UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Maestría en Estadística Matemática

Estimación de proporción de clases en muestras no etiquetadas mediante modelos de cuantificación

Tesis presentada para optar al título de Magíster de
la Universidad de Buenos Aires en Estadística Matemática

Ing. Maximiliano Marufo da Silva

Director de tesis: Dr. Andrés Farall

Buenos Aires, 2025.

Fecha de defensa: 13 de octubre de 2025.

Resumen

La cuantificación consiste en proporcionar predicciones agregadas para conjuntos de datos, en lugar de predicciones individuales para cada dato. En el contexto de la clasificación, esto se traduce en predecir la proporción de cada clase dentro de un conjunto de instancias, en lugar de la clase particular de cada instancia individualmente.

Un ejemplo práctico es la predicción de la proporción de comentarios positivos y negativos sobre un producto, servicio o candidato en redes sociales. Si bien se podría utilizar un clasificador para predecir el sentimiento de cada comentario y, posteriormente, derivar las proporciones de clase, esta estrategia es subóptima y a menudo produce estimaciones sesgadas de la prevalencia, lo que resulta en una baja precisión en la cuantificación. Por consiguiente, se han desarrollado métodos específicos para abordar la cuantificación como una tarea independiente.

Los modelos de cuantificación se entrenan con datos cuya distribución puede diferir de la de los datos de prueba. En el contexto de la cuantificación binaria, para cada instancia $i \in \{1, \ldots, n\}$, consideramos un vector de variables aleatorias $(\boldsymbol{X}_i, Y_i, S_i)$, donde $\boldsymbol{X}_i \in \mathbb{R}^d$ representa las características de la instancia, $Y_i \in \{0, 1\}$ denota su etiqueta de clase, y $S_i \in \{0, 1\}$ indica si la instancia está etiquetada (y, por lo tanto, pertenece al conjunto de entrenamiento). Cuando $S_i = 0$, la etiqueta Y_i no es observable. El objetivo es estimar $p := \mathbb{P}(Y = 1|S = 0)$, es decir, la prevalencia de etiquetas positivas entre las instancias no etiquetadas. No se asume que esta prevalencia sea igual a la de las instancias etiquetadas, $\mathbb{P}(Y = 1|S = 1)$. Además, el estimador de p debe depender únicamente de los datos disponibles: las características de todas las instancias y las etiquetas observadas.

El objetivo de este trabajo es describir el problema de la cuantificación, justificando la necesidad de utilizar modelos optimizados para estos casos, y presentar una revisión del estado del arte en este campo, evaluando mediante simulaciones los principales modelos propuestos.

Palabras Clave: Cuantificación, estimación de proporción de clases, cambio de distribución

Abstract

Estimating class proportions in unlabeled samples using quantification models

Quantification aims to provide aggregate predictions for datasets, rather than individual predictions for each data point. In the context of classification, this translates to predicting the proportion of each class within a set of instances, rather than the specific class of each instance individually.

A practical example is predicting the proportion of positive and negative comments regarding a product, service, or candidate on social media. While a classifier could be used to predict the sentiment of each comment and subsequently derive class proportions, this strategy is suboptimal and often yields biased prevalence estimates, resulting in poor quantification accuracy. Consequently, dedicated methods have been developed to address quantification as an independent task.

Quantification models are trained on data whose distribution may differ from that of the test data. In the context of binary quantification, for each instance $i \in \{1, \dots, n\}$, we consider a vector of random variables $(\boldsymbol{X}_i, Y_i, S_i)$, where $\boldsymbol{X}_i \in \mathbb{R}^d$ represents the instance's features, $Y_i \in \{0, 1\}$ denotes its class label, and $S_i \in \{0, 1\}$ indicates whether the instance is labeled (and therefore belongs to the training set). When $S_i = 0$, the label Y_i is unobserved. The objective is to estimate $p := \mathbb{P}(Y = 1|S = 0)$, i.e., the prevalence of positive labels among unlabeled instances. This prevalence is not assumed to be equal to that of the labeled instances, $\mathbb{P}(Y = 1|S = 1)$. Furthermore, the estimator of p must depend solely on the available data: the features of all instances and the observed labels.

This work aims to describe the quantification problem, justifying the need for optimized models in these cases, and to present a review of the state of the art in this field, evaluating the main proposed models through simulations.

Keywords: Quantification, class proportion estimation, distribution shift

Índice general

Capítulo 1

Problema

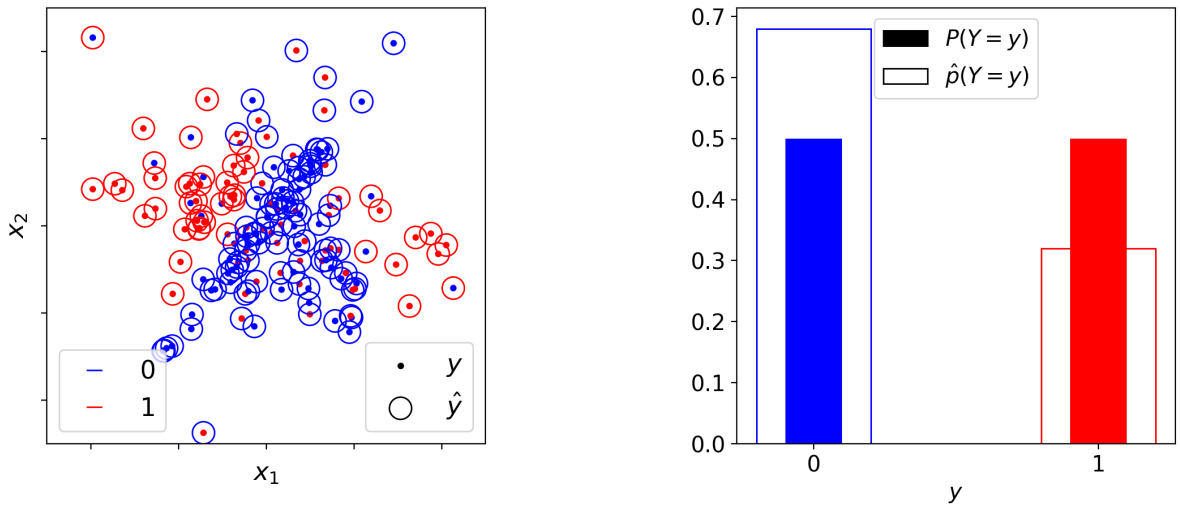
1.1. Introducción

En algunas aplicaciones vinculadas a la clasificación, el objetivo final no es determinar a qué clase (o clases) pertenece cada una de las instancias individuales de un conjunto de datos no etiquetado, sino estimar la proporción (también llamada 'prevalencia', 'frecuencia relativa' o 'probabilidad *prior*') de cada clase en los datos no etiquetados. En los últimos años se ha señalado que, en estos casos, tiene sentido optimizar directamente algoritmos de aprendizaje automático para este objetivo, en lugar de simplemente optimizar clasificadores para etiquetar instancias individuales.

La tarea de ajustar estimadores de prevalencia de clases a través del aprendizaje supervisado se conoce como 'aprender a cuantificar' o, más simplemente, cuantificar o *quantification* (término acuñado por Forman [1], quien planteó el problema por primera vez). Se sabe que cuantificar mediante la clasificación de cada instancia no etiquetada a través de un clasificador estándar y luego contando las instancias que han sido asignadas a cada clase (el primer método de cuantificación que veremos, denominado *Classify & Count -CC-*) generalmente conduce a estimadores de prevalencia de clases sesgados, es decir, obtienen poca exactitud en la cuantificación. Como resultado, se han desarrollado métodos que abordan la cuantificación como una tarea en sí.

Para ver la importancia de diferenciar el problema de cuantificación del de clasificación, veamos dos ejemplos. En el primero, una empresa que ofrece un servicio a sus clientes realiza una encuesta con varias preguntas para determinar el grado de satisfacción de cada persona. El objetivo de le empresa es determinar aquellos clientes que podrían no estar conformes con el servicio y ofrecerles una mejora en las condiciones para retenerlos. En el segundo ejemplo, una consultora analiza *tweets* para estimar el grado de aprobación de candidatos políticos. Aquí, la consultora no está interesado en predecir si un individuo específico está a favor o en contra, sino en cuántos encuestados, del número total de encuestados, aprueban al candidato, es decir, en conocer la prevalencia de la clase positiva.

Mientras en el primer escenario el interés es a nivel individual, en el último, el nivel agregado es lo que importa; en otras palabras, en el primer escenario la clasificación es el objetivo, mientras que en el segundo el verdadero objetivo es la cuantificación. De hecho, en muchas de las aplicaciones las predicciones que interesan no son a nivel individual sino a nivel colectivo; ejemplos de tales campos son la investigación de mercado, la ciencia política, las ciencias sociales, modelado ecológico y epidemiología.



(a) Clasificación **(b)** Cuantificación

Figura 1.1: En la clasificación, la predicción es a nivel individual, mientras que en la cuantificación es a nivel agregado (para mayor información sobre los datos del gráfico consultar el Capítulo 2).

En resumen, y generalizando no solo para clasificación sino también a otros problemas (regresión, ordinalidad, etc.), la tarea de cuantificación consiste en proporcionar predicciones agregadas para conjuntos de datos, en vez de predicciones particulares sobre los datos individuales. Si bien en principio no es necesario realizar predicciones por cada individuo, muchos de los métodos se basan en obtener la cuantificación de esa manera, ya que hacer predicciones individuales suele ser un requisito de por sí de las aplicaciones prácticas, o porque ya existen en ellas modelos que las generen.

La literatura sobre métodos relacionados con cuantificación está un tanto desconectada entre sí. Algunos de los métodos que pueden usarse como cuantificadores han sido ideados para otros fines, principalmente para mejorar la precisión en clasificación cuando cambia el dominio. El desempeño de este último grupo ha sido normalmente estudiado solo en términos de mejora en las tareas de clasificación pero no como cuantificadores. Dado este escenario, y debido a la variedad de campos en los que ha surgido como una necesidad de aplicación, los algoritmos que se pueden aplicar para tareas de cuantificación aparecen en artículos que usan diferentes palabras clave y nombres, como *counting* [2], *prior probability shift* [3, 4], *posterior probability estimation* [5], *class prior estimation* [6–8], *class prior change* [9], *prevalence estimation* [10], *class ratio estimation* [11] o *class distribution estimation* [12–14], por citar solo algunos de ellos.

1.2. Tipos de cuantificación

Aunque el estudio de la cuantificación se ha centrado principalmente en el dominio de clasificación, la cuantificación también aparece en otros tipos de problemas de aprendizaje automático, como la regresión, la clasificación ordinal, el aprendizaje sensible al costo y la cuantificación en redes.

De manera similar a la regresión, aprender a cuantificar admite diferentes problemas de interés aplicativo, basados en cuántas clases distintas existen en el problema, y en cuántas de las clases se

pueden atribuir al mismo tiempo al mismo individuo. Así, los problemas de cuantificación se dividen de esta manera:

1. Etiquetado simple *(Single-Label Quantification -SLQ-)*: cuando cada individuo pertenece exactamente a una de las clases en $C = \{c_1, \ldots, c_{\#C}\}$.

2. Etiquetado múltiple *(Multi-Label Quantification -MLQ-)*: cuando cada individuo puede pertenecer a cualquier número de clases (cero, una o varias) en $C = \{c_1, \ldots, c_{\#C}\}$.

3. Cuantificación binaria *(Binary Quantification -BQ-)*: se puede definir de dos formas (equivalentes) según el tipo de etiquetado definido:

   a) en *SLQ* con $\#C = 2$, (en este caso $C = \{c_1, c_2\}$, y cada individuo pertenece a c_1 o c_2)

   b) en *MLQ* con $\#C = 1$, (en este caso $C = \{c\}$, y cada individuo pertenece o no a c)

4. Cuantificación ordinal *(Ordinal Quantification -OQ-)*: cuando existe un orden $c_1 \prec \cdots \prec c_{\#C}$ en $C = \{c_1, \ldots, c_{\#C}\}$.

5. Cuantificación de regresión *(Regression Quantification -RQ-)*: cuando no hay un conjunto de clases involucradas, sino que cada individuo está etiquetado con una puntuación de valor real y la cuantificación equivale a estimar la fracción de ítems cuya puntuación está en un intervalo dado $[a, b]$ con $a, b \in \mathbb{R}^d$.

1.3. Marco teórico

Si hablamos entonces de cuantificación binaria, se tiene que por cada muestra $i \in \{1, \ldots, n\}$, $(\boldsymbol{X}_i, Y_i, S_i)$ es un vector de variables aleatorias tal que $\boldsymbol{X}_i \in \mathbb{R}^d$ son las características de la muestra, $Y_i \in C$ con $C = \{1, 0\}$ indica la clase a la que pertenece y $S_i \in \{1, 0\}$ indica si pertenece al conjunto de entrenamiento o al de prueba. Es decir, cuando $S_i = 0$, entonces Y_i no es observable. El objetivo es estimar $p := \mathbb{P}(Y = 1|S = 0)$[1], es decir, la prevalencia de etiquetas positivas en conjuntos de prueba. Esta prevalencia no se asume de ser la misma que en el conjunto de entrenamiento, $\mathbb{P}(Y = 1|S = 1)$. Además, el estimador de p debe depender solo de los datos disponibles, es decir, de las características de la muestra (tanto del conjunto de entrenamiento como de prueba) y de las etiquetas del conjunto de entrenamiento. Los supuestos que se asumen [15] son:

- $(\boldsymbol{X}_1, Y_1, S_1) \ldots (\boldsymbol{X}_n, Y_n, S_n)$ son independientes. Es decir, el conocimiento del valor de una muestra no proporciona ninguna información sobre el valor de la otra, y viceversa.

- Por cada $s \in \{0, 1\}$, $(\boldsymbol{X}_1, Y_1)|S_1 = s, \ldots, (\boldsymbol{X}_n, Y_n)|S_n = s$ son idénticamente distribuidas. En otras palabras, que si separamos los conjuntos de entrenamiento y de prueba, en cada uno de ellos los individuos son tomados de la misma distribución de probabilidad.

- Por cada $(y_1, \ldots, y_n) \in \{0, 1\}^n$, $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ es independiente de $(S_1, \ldots, S_n)$ condicionado a $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$. Este es el supuesto más fuerte, ya que implica que la relación entre las características de un individuo y su etiqueta no está intermediada por el conjunto (entrenamiento o prueba) al que pertenece. Es decir, que una vez que conocemos la etiqueta

[1] Para el caso multiclase, $p_1, \ldots, p_{\#C}$ o $p(c)$.

del individuo, sabemos la distribución de sus características, independientemente del conjunto al que pertenece.

Usando la distribución de probabilidad conjunta, podemos factorizar usando las distribuciones condicionales:

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, Y = y, S = s) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y, S = s)\mathbb{P}(Y = y|S = s)\mathbb{P}(S = s) \tag{1.1}$$

Luego, usando el tercer supuesto mencionado, podemos hacer:

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, Y = y, S = s) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)\mathbb{P}(Y = y|S = s)\mathbb{P}(S = s) \tag{1.2}$$

Si bien existen varios métodos propuestos para el aprendizaje de cuantificación [16, 17], este campo de estudio es todavía relativamente desconocido incluso para expertos en aprendizaje automático. La razón principal es la creencia errónea de que es una tarea trivial que se puede resolver usando un método directo, como *CC*. La cuantificación requiere métodos más sofisticados si el objetivo es obtener modelos óptimos, y su principal dificultad radica en la definición del problema, ya que las distribuciones de los datos de entrenamiento y de prueba pueden ser distintas. Por ejemplo, si la diferencia entre $\mathbb{P}(Y = 1|S = 0)$ y $\mathbb{P}(Y = 1|S = 1)$ es grande, los métodos simples como *CC* suelen tener bajo rendimiento.

1.4. Cambios en las distribuciones de los datos

En los últimos años ha habido un interés creciente en las aplicaciones que presentan cambios en las distribuciones de datos (conocido en la bibliografía por su término en inglés *dataset shift*). Estos problemas comparten el hecho de que la distribución de los datos utilizados para entrenar es diferente a la de los datos que se usan para predecir. Al igual que para el área de la cuantificación, aquí también la literatura sobre el tema está dispersa y diferentes autores usan diferentes nombres para referirse a los mismos conceptos, o usan el mismo nombre para diferentes conceptos.

Teniendo en cuenta que en los problemas de clasificación tenemos:

- un conjunto de características o covariables $\boldsymbol{X}$,
- una variable de respuesta Y,
- una distribución de probabilidad conjunta. $\mathbb{P}(Y = y, \boldsymbol{X} = \boldsymbol{x})$.

La probabilidad conjunta $\mathbb{P}(Y = y, \boldsymbol{X} = \boldsymbol{x})$ se puede escribir como $\mathbb{P}(Y = y|\boldsymbol{X} = \boldsymbol{x})\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})$ o como $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)\mathbb{P}(Y = y)$. Por otro lado, cuando usamos los términos de entrenamiento (*train*) y prueba (*test*), nos referimos a las datos disponibles para entrenar al clasificador y los datos presentes en el entorno en el que se implementará el clasificador, respectivamente. Podemos entonces también separar los datos en dos distribuciones distintas, condicionando a la variable S definida en 1.3, siendo $\mathbb{P}_{tr}(Y = y, \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y = y, \boldsymbol{X} = \boldsymbol{x}|S = 1)$ y $\mathbb{P}_{tst}(Y = y, \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y = y, \boldsymbol{X} = \boldsymbol{x}|S = 0)$.

El *dataset shift* aparece cuando las distribuciones conjuntas de entrenamiento y de prueba son diferentes, es decir, cuando $\mathbb{P}_{tr}(Y = y, \boldsymbol{X} = \boldsymbol{x}) \neq \mathbb{P}_{tst}(Y = y, \boldsymbol{X} = \boldsymbol{x})$. Moreno-Torres et al. [3] distingue las distintas variantes del *dataset shift* según qué elementos mencionados anteriormente cambian:

- *Covariate shift*, cuando $\mathbb{P}_{tr}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_{tst}(Y = y|\boldsymbol{X} = \boldsymbol{x})$ y $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}) \neq \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x})$,
- *Prior probability shift*, cuando $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}|Y = y) = \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$ y $\mathbb{P}_{tr}(Y = y) \neq \mathbb{P}_{tst}(Y = y)$,
- *Concept shift*, cuando $\mathbb{P}_{tr}(Y = y|\boldsymbol{X} = \boldsymbol{x}) \neq \mathbb{P}_{tst}(Y = y|\boldsymbol{X} = \boldsymbol{x})$ y $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x})$ o $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}|Y = y) \neq \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$ y $\mathbb{P}_{tr}(Y = y) = \mathbb{P}_{tst}(Y = y)$.

Otros tipos de *dataset shift* surgen cuando $\mathbb{P}_{tr}(Y = y|\boldsymbol{X} = \boldsymbol{x}) \neq \mathbb{P}_{tst}(Y = y|\boldsymbol{X} = \boldsymbol{x})$ y $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}) \neq \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x})$ y cuando $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}|Y = y) \neq \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$ y $\mathbb{P}_{tr}(Y = y) \neq \mathbb{P}_{tst}(Y = y)$. Sin embargo, estos tipos de cambios no se consideran generalmente en la literatura ya que aparecen mucho más raramente, o porque son difíciles o imposibles de resolver.

El problema de cuantificación se trata de un caso donde $\mathbb{P}_{tst}(Y = y)$ es desconocido. Además, la mayoría de los métodos de cuantificación propuestos asumen que $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}|Y = y) = \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$, por lo que están dentro de los casos de *prior probability shift*. Esto implica que se asumen los tres supuestos mencionados en 1.3.

Por otro lado, en la mayoría de los casos el objetivo final de la implementación es estimar algún parámetro de $\mathbb{P}_{tst}(Y = y)$. Por ejemplo, como ya mencionamos anteriormente, en la cuantificación binaria, se desea estimar $p := \mathbb{P}(Y = 1|S = 0)$, o lo que es lo mismo, $p_{tst} := \mathbb{P}_{tst}(Y = 1)$. Es decir, en la cuantificación la tarea indirectamente suele ser aprender a aproximar una distribución desconocida (observando solo características de una muestra) mediante una distribución conocida. En consecuencia, prácticamente todas las medidas de evaluación para la cuantificación son divergencias, es decir, medidas de cómo una distribución pronosticada difiere de la distribución real.

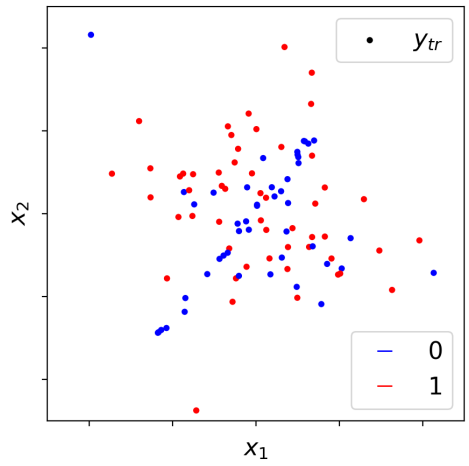
1.5. El problema de clasificar y contar

En ausencia de métodos para estimar los valores de prevalencia de clase de forma directa, el primer método que suele pensarse para hacerlo es *Classify & Count* o *CC*, es decir, clasificar cada individuo no etiquetado y estimar los valores de prevalencia de clase contando los individuos que fueron asignados a cada clase. Sin embargo, esta estrategia es subóptima: si bien un clasificador perfecto produce también un cuantificador perfecto, un buen clasificador no necesariamente produce un mejor cuantificador que el que puede producir un mal clasificador. Para ver esto, se puede ver la definición de F_1, una función de evaluación estándar para la clasificación binaria, que se define como:

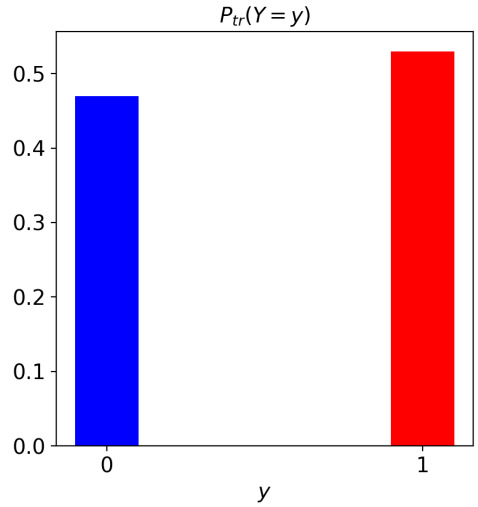
$$F_1 = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{1.3}$$

donde tp, fp, fn indican el número de verdaderos positivos, falsos positivos y falsos negativos, respectivamente. Un buen clasificador, según la medida F_1, puede producir un mal cuantificador ya que F_1 considera buenos aquellos clasificadores que mantienen la suma $fp + fn$ al mínimo; sin embargo, el objetivo de un algoritmo de cuantificación debe ser mantener al mínimo $|fp - fn|$. Es decir, que un mal clasificador con fp y fn altos pero similares podría llegar a producir un mejor cuantificador que un buen clasificador.

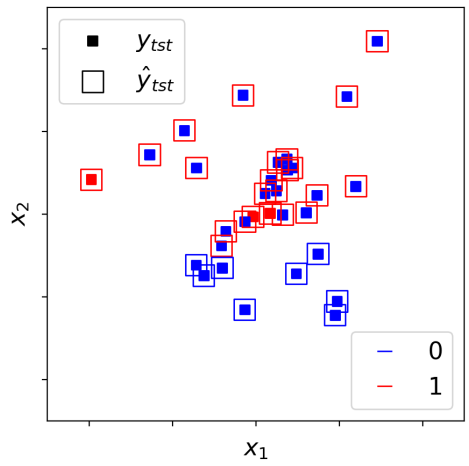
El análisis teórico de esta cuestión se basa en el supuesto de *prior probability shift* 1.4. Bajo tal supuesto, la estimación $\hat{p}$ obtenida por el enfoque *CC* depende solo de las características del



(a) Muestra de entrenamiento.

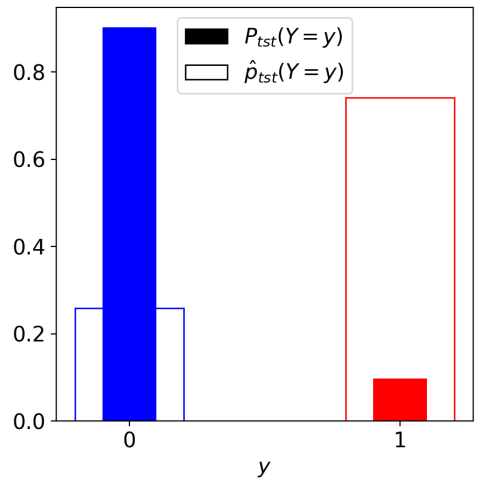


(b) Prevalencia de clases en muestra de entrenamiento.



(c) Muestra de prueba[a] y su clasificación.

[a] con $\mathbb{P}_{tst}(\boldsymbol{X}=\boldsymbol{x}|Y=y)=\mathbb{P}_{tr}(\boldsymbol{X}=\boldsymbol{x}|Y=y)$.



(d) Prevalencia de clases verdadera y estimada (por el método *CC* usando un clasificador Naive Bayes) en muestra de prueba.

Figura 1.2: En la muestra de entrenamiento (1.2a y 1.2b), se observa una determinada distribución de datos y su correspondiente prevalencia de clases. En la muestra de prueba (1.2c), aunque la relación entre las características y las clases se mantiene, el hecho de que la prevalencia de clases haya cambiado (*prior probability shift*) significativamente hace que el método *CC* pueda estimar incorrectamente (si el clasificador para predecir las clases de los individuos es imperfecto) la prevalencia de clases en el conjunto de prueba (1.2d).

clasificador, definido (para el caso binario) por su tasa de verdaderos positivos (tpr), su tasa de falsos positivos (fpr) y de la prevalencia real (p):

$$\hat{p}(p) = p \cdot tpr + (1-p) \cdot fpr \tag{1.4}$$

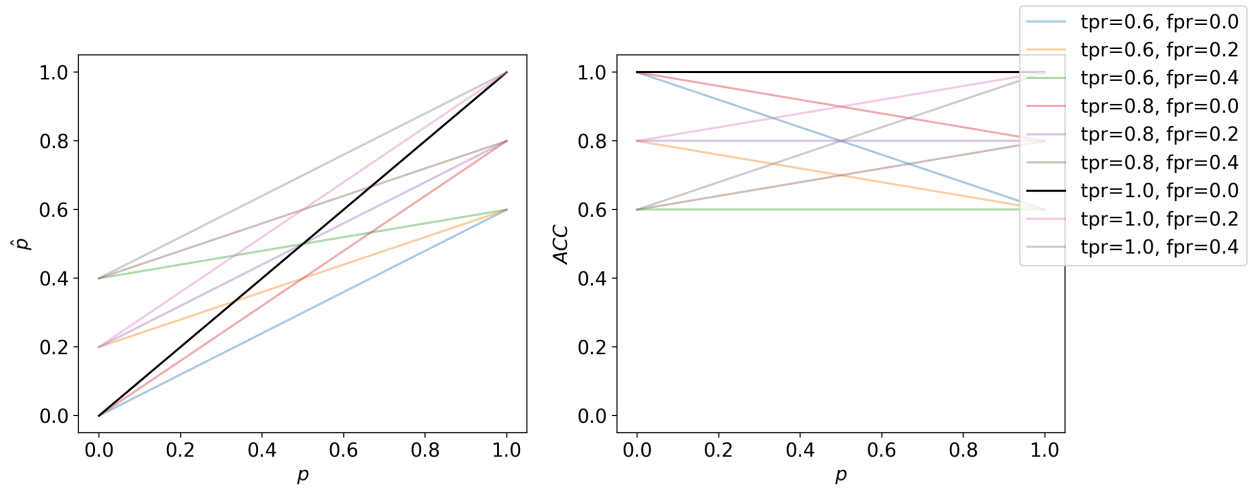


Figura 1.3: La línea negra representa el cuantificador y clasificador perfecto, respectivamente. Las otras líneas muestran las estimaciones teóricas de $\hat{p}$ resultantes de aplicar la ecuación 1.4, y el *accuracy* correspondiente al clasificador, según se varían los valores de tpr y fpr.

El mal desempeño de CC fue demostrado mediante el siguiente teorema por Forman [18]:

Teorema 1.1 (Teorema de Forman)**.** [18, p.169] Para un clasificador imperfecto, el método CC subestimará la proporción verdadera de ejemplos positivos p en un conjunto de prueba para $p > p^*$, y sobreestimará para $p < p^*$, donde p^* es la proporción particular en la cual el método CC estima de forma correcta. Es decir, el método CC estima exactamente p^* para un conjunto de prueba con p^* muestras positivas.

La demostración (ver todos los detalles en [18, p.170]) supone que el cuantificador CC produce una predicción perfecta para una prevalencia concreta, llamada p^*, y estudia su comportamiento cuando la prevalencia cambia ligeramente. Suponiendo que $\hat{p}(p^*) = p^*$, cuando la prevalencia cambia en una cantidad $\Delta \neq 0, p^* + \Delta$, la estimación del método CC en tal caso será:

$$\begin{aligned} \hat{p}(p^* + \Delta) &= (p^* + \Delta) \cdot tpr + (1 - (p^* + \Delta)) \cdot fpr \\ &= \hat{p}(p^*) + (tpr - fpr) \cdot \Delta \\ &= p^* + (tpr - fpr) \cdot \Delta \end{aligned} \tag{1.5}$$

La predicción del método CC será perfecta, $\hat{p}(p^* + \Delta) = (p^* + \Delta)$, solo cuando el clasificador también es perfecto ($tpr = 1$, $fpr = 0$ y por tanto $tpr - fpr = 1$). Pero en el caso habitual, en el cual el clasificador es imperfecto ($0 \leq tpr - fpr < 1$), cuando la prevalencia aumenta ($\Delta > 0$), CC la subestima ($\hat{p} < p^* + \Delta$), y cuando la prevalencia disminuye ($\Delta < 0$), CC la sobreestima ($\hat{p} > p^* + \Delta$).

Un buen clasificador puede estar sesgado, es decir, puede mantener sus falsos positivos al mínimo solo a expensas de una cantidad sustancialmente mayor de falsos negativos (o viceversa); si este es el caso, el clasificador es un mal cuantificador. Este fenómeno no es infrecuente, especialmente en presencia de datos desbalanceados. En tales casos, los algoritmos que minimizan las funciones de pérdida de clasificación (*Hamming*, *hinge*, etc.) suelen generar clasificadores con tendencia a elegir la clase mayoritaria, lo que implica un número mucho mayor de falsos positivos que de falsos negativos

para la clase mayoritaria, lo que significa a su vez que tal algoritmo tenderá a subestimar las clases minoritarias.

Los argumentos anteriores indican que no se debe considerar la cuantificación como un mero subproducto de la clasificación, y debe estudiarse y resolverse como una tarea en sí misma. Hay al menos otros dos argumentos que apoyan esta idea. Uno, es que las funciones que se utilizan para evaluar la clasificación no pueden utilizarse para evaluar la cuantificación, ya que estas miden, en general, cuántos individuos han sido mal clasificados, y no cuánto difiere la prevalencia de clase estimada del valor real. Esto significa que los algoritmos que minimizan estas funciones están optimizados para la clasificación, y no para la cuantificación. Un segundo argumento presentado por Forman [18] es que los métodos diseñados específicamente para cuantificar requieren menos datos de entrenamiento para alcanzar la misma precisión de cuantificación que los métodos estándar basados en *CC*. Si bien esta observación es de naturaleza empírica, también existen argumentos teóricos que sustentan este hecho [19].

1.6. Cuantificadores para la mejora de la clasificación

Debido a los problemas mencionados anteriormente de los clasificadores frente a cambios en las distribuciones de los datos y frente a datos desbalanceados, los algoritmos de cuantificación están cada vez más frecuentemente siendo usados en tareas que requieren predicciones individuales. Los mismos se emplean como suplemento de clasificadores para suplir sus defectos frente a estos problemas, ya que en algunos casos no solo predicen los valores agregados, sino que también mejoran las predicciones a nivel individual.

Por ejemplo, el *prior probability shift* puede hacer que los clasificadores performen de manera subóptima. Considerando como generalización el clasificador óptimo de Bayes, dado por:

$$h(\boldsymbol{x}) = \arg\max_{y \in C} p_{Y|\boldsymbol{X}=\boldsymbol{x}}(y), \tag{1.6}$$

y si aplicamos el Teorema de Bayes:

$$h(\boldsymbol{x}) = \arg\max_{y \in C} \frac{p_{\boldsymbol{X}|Y=y}(\boldsymbol{x}) p_Y(y)}{p_{\boldsymbol{X}}(\boldsymbol{x})}, \tag{1.7}$$

la decisión del clasificador depende de $p_Y(y)$, que es estimado generalmente con la muestra de entrenamiento, usando entonces $\hat{p}_Y(y=1) = p_{tr}$. Es decir, que en caso de $\mathbb{P}_{tr}(Y=y) \neq \mathbb{P}_{tst}(Y=y)$, la decisión final del clasificador puede verse afectada negativamente. Para mejorar el rendimiento del clasificador frente a estos casos, se debería usar $\hat{p}_Y(y=1) = p_{tst}$, pero como p_{tst} es generalmente desconocido, se puede usar un método de cuantificación para estimarlo [5, 8, 14, 20].

Los métodos de cuantificación pueden usarse no solo para mejorar el rendimiento general de un clasificador, sino también para mejorar su equidad o *fairness* [21, 22], es decir, su posibilidad de predecir resultados independientes de un cierto conjunto de variables que consideramos sensibles y no relacionadas con él (e.j.: género, etnia, orientación sexual, etc.). Por ejemplo, suponiendo que una variable Z debe considerarse sensible, se puede estimar $\mathbb{P}_{tr}(Y=y|Z=z)$. Luego, si los datos de entrenamiento están sesgados, por ejemplo, con $\mathbb{P}_{tr}(Y=1|Z=1) \gg \mathbb{P}_{tr}(Y=1|Z=0)$, pero se sabe que en las muestras a inferir esto no es así, se puede optimizar el modelo de clasificación imponiendo alguna penalidad basada en la estimación de $\mathbb{P}_{tst}(Y=y|Z=z)$, siendo esta última obtenido por un cuantificador.

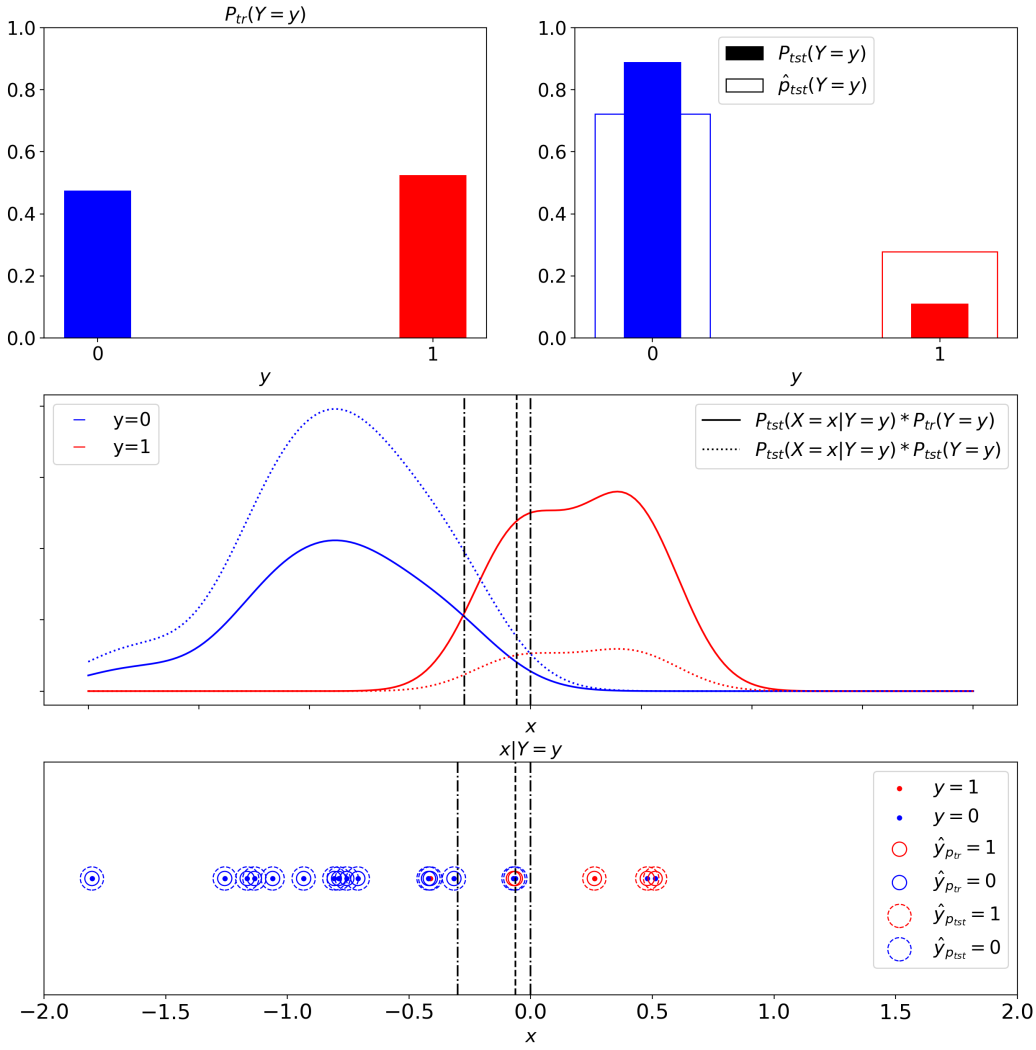


Figura 1.4: Ejemplo de *prior probability shift* con $d = 1$ (es decir, $\boldsymbol{X}_i \in \mathbb{R}$) e implementando la regla óptima de Bayes para un modelo Naive Bayes Gaussiano. Se puede observar como cambia la clase predicha para el punto $\boldsymbol{x}$ marcado con línea punteada vertical, si se usa $\hat{p}_Y(y) = \mathbb{P}_{tst}(Y = y)$ en vez de $\hat{p}_Y(y) = \mathbb{P}_{tr}(Y = y)$, debido al corrimiento de la frontera de decisión.

Capítulo 2

Métodos de estimación

Durante los últimos años, se han propuesto varios métodos de cuantificación desde diferentes perspectivas y con diferentes objetivos. En términos generales, se pueden distinguir dos grandes clases de métodos en la literatura. La primera clase es la de métodos agregativos, es decir, métodos que requieren la clasificación de todos los individuos como un paso intermedio. Dentro de los métodos agregativos, se pueden identificar dos subclases. La primera subclase incluye métodos basados en clasificadores de propósito general; en estos métodos la clasificación de los elementos individuales realizados como un paso intermedio puede lograrse mediante cualquier clasificador. La segunda subclase se compone, en cambio, de métodos que para clasificar los individuos, se basan en métodos de aprendizaje diseñados con la cuantificación en mente. La segunda clase es la de métodos no agregativos, es decir, métodos que resuelven la tarea de cuantificación "holísticamente", es decir, sin clasificar a los individuos. La idea de esta tesis no es la de mostrar todos los métodos propuestos hasta la actualidad, sino la de mencionar los métodos más populares.

Caso de ejemplo

Como ejemplo de muestra para comparar los distintos métodos de estimación, se usará el mismo set de datos artificiales que el de las figuras 1.1 y 1.2. Para crear el set de datos se utilizó el algoritmo prepuesto por Guyon et al. [23] mediante la función *make_classification* de *scikit-learn*, usando los siguientes parámetros de entrada:

```
population_size = 150 X, y = make_classification( n_samples=
   population_size,
n_features=2, n_informative=2, n_redundant=0, n_repeated=0,
n_clusters_per_class=2, n_classes=2, weights=None, class_sep=0.1,
random_state=42, )
```

Esta función, al usar `weights=None`, crea un set de datos balanceados en cuanto a las clases de los individuos (figura 1.1). Sin embargo, para poder evaluar los distintos métodos, lo que haremos ahora es seleccionar 100 individuos y usarlos como datos de entrenamiento, y hacer un sub-muestreo de los 50 restantes de forma tal de aproximarnos a un $p_{tst} = 0.1$:

```
train_size = 100 prev_test = 0.1 X_train, y_train = X[:train_size],
y[:train_size] X_test, y_test = X[train_size:], y[train_size:]
```

```
idx_negatives_test = np.argwhere(y_test==0).flatten() idx_positives_test =
np.random.choice( np.argwhere(y_test==1).flatten(),
size=round((prev_test)*len(idx_negatives_test)/(1-prev_test)), replace=
    False )
idx_test = np.concatenate([idx_positives_test, idx_negatives_test]) X_test
    =
X_test[idx_test] y_test = y_test[idx_test]
```

El código de arriba termina generando un set de entrenamiento de tamaño $n_{tr} = 100$ con $p_{tr} = 0.53$ (figuras 1.2a y 1.2b) y uno de prueba de tamaño $n_{tst} = 31$ con $p_{tst} \approx 0.097$ (figuras 1.2c y 1.2d).

El modelo de clasificación utilizado para generar las predicciones mostradas en ambas figuras 1.1 (entrenando y prediciendo con todos los datos) y 1.2 (entrenando con los datos de entrenamiento y prediciendo los de prueba) es el clasificador *Naive Bayes* generado mediante la clase *GaussianNB* de *scikit-learn*, usando sus parámetros por defecto.

El código en Python para generar los resultados de los ejemplos de cada método se encuentra en el repositorio https://github.com/maxi-marufo/tesis.

2.1. Métodos agregativos

2.1.1. Con clasificadores generales

Dentro de los métodos agregativos, algunos de ellos requieren como entrada las etiquetas de clases predichas (es decir, el tipo de salida que devuelven los clasificadores denominados duros), otros requieren un *score* de decisión (como podría ser la distancia al hiperplano de separación en el clasificador SVM), y otros que requieren como entrada las probabilidades *a posteriori* de pertenencia a cada clase (es decir, el tipo de salida que devuelven los clasificadores denominados blandos)[1]. En estos últimos, además, las probabilidades *a posteriori* deben estar calibradas (para mayor información sobre calibración consultar el Apéndice A). Para estos casos, en los ejemplos a continuación que requieren clasificadores blandos se separó de las muestras de entrenamiento un 15 % de datos para el proceso de calibración (y estratificando para que la proporción de etiquetas se mantenga igual).

Clasificar y Contar (CC)

El método más sencillo y directo para construir un cuantificador para clasificación (tanto binaria como multiclase) es aplicar el enfoque *Classify & Count* [1]. *CC* juega un papel importante en la investigación de cuantificación ya que siempre se utiliza como el *baseline* que cualquier método de cuantificación razonable debe mejorar. Este método consiste simplemente en: (i) ajustar un clasificador duro, y luego (ii), utilizando dicho clasificador, clasificar las instancias de la muestra de prueba, contando la proporción de cada clase. Generalizando el estimador de *CC* para el caso

[1] Los clasificadores blandos y de *score* se pueden convertir en duros usando umbrales de clasificación

multiclase, el mismo queda entonces definido por:

$$\hat{p}_{tst}^{CC}(c) = \frac{\#\{\boldsymbol{x} \in \boldsymbol{X}_{tst} | h_{tr}(\boldsymbol{x}) = c\}}{\#\boldsymbol{X}_{tst}} \tag{2.1}$$

donde se usó h_{tr} para la función de decisión del clasificador duro ajustado con la muestra de entrenamiento.

Es evidente que podemos obtener un cuantificador perfecto si el clasificador es también perfecto. El problema es que obtener un clasificador perfecto es casi imposible en aplicaciones reales, y luego el cuantificador hereda el sesgo del clasificador. Este aspecto se analiza en varios artículos tanto desde una perspectiva teórica como práctica, como lo hizo Forman [18], y como también ya lo hemos mencionado en 1.5.

Ejemplo: Para el caso de ejemplo, y manteniendo el clasificador allí usado, debemos contar la cantidad de predicciones positivas (rojas) en la figura 1.2, y dividirlas por el tamaño de la muestra de prueba. Es decir, $\hat{p}_{tst}^{CC}(c = 1) = \frac{23}{31} \approx 0.742$.

Clasificar, Contar y Ajustar (ACC)

Conocido en inglés como *Adjusted Classify & Count*, *Adjusted Count* [18] o también como *Confusion Matrix Method* [20], este método se basa en corregir las estimaciones de CC teniendo en cuenta la tendencia del clasificador a cometer errores de cierto tipo. Un modelo ACC está compuesto por dos elementos: un clasificador duro (como en CC) y de las estimaciones de tpr y fpr. Dichas estimaciones pueden obtenerse usando validación cruzada o *cross-validation*, ya sea mediante la técnica de *k-folds* o un *held-out*. Luego, en la fase de predicción, el modelo obtiene una primera estimación $\hat{p}$ de la misma forma que en CC que luego, para el caso binario, es ajustado aplicando la siguiente fórmula[2]:

$$\hat{p}_{tst}^{ACC}(c = 1) = \frac{\hat{p}_{tst}^{CC}(c = 1) - \hat{fpr}}{\hat{tpr} - \hat{fpr}} \tag{2.2}$$

Esta expresión se obtiene despejando la verdadera prevalencia p de la ecuación 1.4 y reemplazando fpr y tpr por sus estimadores.

El método ACC es teóricamente perfecto, independientemente del valor de *accuracy* obtenido con el clasificador, cuando se cumple el supuesto de *prior probability shift* 1.4 y cuando las estimaciones de tpr y fpr son perfectas. Desafortunadamente, es raro que se cumplan ambas condiciones en aplicaciones del mundo real: $\mathbb{P}(\boldsymbol{X}|Y)$ puede tener variaciones entre los datos de entrenamiento y los de predicción, y es difícil obtener estimaciones perfectas para tpr y fpr en algunos dominios, ya que suele haber pequeñas muestras disponibles y/o están muy desequilibradas. Pero incluso en estos casos, el rendimiento del método ACC suele ser mejor que el de CC.

Partiendo de la ecuación 2.1 y utilizando el teorema de probabilidad total, podemos extender la ecuación 2.2 para el caso multiclase:

$$\begin{aligned} \hat{p}_{tst}^{CC}(c = c_k) &= \hat{\mathbb{P}}_{tst}(h_{tr}(\boldsymbol{x}) = c_k) \\ &= \sum_{j=1}^{\#C} \hat{\mathbb{P}}(h_{tr}(\boldsymbol{x}) = c_k | y = c_j) \hat{p}_{tst}^{ACC}(c = c_j) \end{aligned} \tag{2.3}$$

[2] A veces, esta expresión conduce a un valor inválido de $\hat{p}_{tst}^{ACC}$ que debe recortarse en el rango $[0, 1]$ en un último paso.

donde $\hat{p}_{tst}^{CC}(c = c_k)$ es la fracción de datos de prueba que el clasificador h asigna a c_k (y por ende, es conocido), y $\hat{\mathbb{P}}(h_{tr}(\boldsymbol{x}) = c_k | y = c_j)$ es la estimación de probabilidad de que el clasificador h asigne la clase c_k a $\boldsymbol{x}$ cuando este pertenece a la clase c_j. Estas probabilidades, al igual que *tpr* y *fpr* en el caso binario, deben estimarse mediante validación cruzada [1, 10, 18]. Luego, $\hat{p}_{tst}^{ACC}(c = c_j)$, nuestras incógnitas (una por cada c_j), pueden calcularse mediante un sistema de ecuaciones lineales con $\#C$ ecuaciones y $\#C$ incógnitas.

Ejemplo: Aquí debemos estimar el *tpr* y *fpr*. Para ello, se separó de la muestra de entrenamiento un 15 % de datos. Con el 85 % de la muestra se entrenó el clasificador y se obtuvo un $\hat{p}_{tst}^{CC}(c = 1) \approx 0.71$, y con el 15 % separado se obtuvo $\hat{tpr} \approx 0.625$ y $\hat{fpr} \approx 0.714$, y por lo tanto, $\hat{p}_{tst}^{ACC}(c = 1) \approx 0.0516$.

Clasificar y Contar Probabilístico (PCC)

Este método, conocido en inglés como *Probabilistic Classify and Count* [24, 25], es una variante de *CC* que utiliza un clasificador blando en vez de uno duro. Es decir, que la salida del clasificador blando ajustado con la muestra de entrenamiento, $s(\boldsymbol{x}, y)$, será una estimación de la probabilidad *a posteriori* $p_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ por cada individuo $\boldsymbol{x} \in \boldsymbol{X}_{tst}$ y cada $y \in C$. El método consiste en estimar las $p_{tst}(c = c_j)$ mediante el valor esperado de la proporción de items que se predijeron como pertenecientes a cada clase c_j:

$$\begin{aligned} \hat{p}_{tst}^{PCC}(c = c_j) &= \hat{\mathbb{E}}[p_{Y|\boldsymbol{X}=\boldsymbol{x}}(y = c_j)] \\ &= \frac{1}{m} \sum_{i=1}^{m} \hat{p}_{Y|\boldsymbol{X}=\boldsymbol{x}_i}(y = c_j) \\ &= \frac{1}{m} \sum_{i=1}^{m} s(\boldsymbol{x}_i, y = c_j) \end{aligned} \tag{2.4}$$

con $m = \#\boldsymbol{X}_{tst}$. La intuición detrás de *PCC* es que las probabilidades *a posteriori* contienen mayor información que las decisiones de un clasificador duro y, por lo tanto, deberían ser usadas en su lugar. Sin embargo, Tasche [26, Corolario 6, p.157 y p.163] demuestra que el comportamiento de *PCC* será similar al de *CC*, en cuanto a que ambos subestiman o sobreestiman la prevalencia verdadera cuando la distribución de clases cambia entre los datos de entrenamiento y de prueba.

Ejemplo: Como este método utiliza un clasificador blando, se separó primero un 15 % de los datos de entrenamiento. Con el 85 % se entrenó el clasificador, y luego con el 15 % se realizó la calibración. Luego, debemos sumar las salidas del clasificador calibrado para la clase positiva. Para el ejemplo, se obtuvieron las siguientes salidas:

$s(\boldsymbol{x}_i, y = 1)$	0.54	0.54	0.49	0.50	0.45	0.56	0.57	0.60	0.52	0.50	0.58	0.54 ...	0.49

siendo $\hat{p}_{tst}^{PCC}(c = 1) \approx 0.527$.

Clasificar, Contar y Ajustar Probabilístico (PACC)

Presentado como *Probabilistic Adjusted Classify and Count* o también como *Probabilistic Adjusted Count*, este método combina las ideas de ACC y de PCC [24, 25].

$$\begin{aligned}
\hat{p}_{tst}^{PCC}(c=c_k) &= \hat{\mathbb{E}}[\mathbb{P}_{tst}(h_{tr}(\boldsymbol{x})=c_k)] \\
&= \hat{\mathbb{E}}\left[\sum_{j=1}^{\#C}\mathbb{P}(h_{tr}(\boldsymbol{x})=c_k|y=c_j)p_{tst}^{PACC}(c=c_j)\right] \\
&= \sum_{j=1}^{\#C}\hat{\mathbb{E}}\left[\mathbb{P}(h_{tr}(\boldsymbol{x})=c_k|y=c_j)p_{tst}^{PACC}(c=c_j)\right] \\
&= \sum_{j=1}^{\#C}\hat{\mathbb{E}}\left[\mathbb{P}(h_{tr}(\boldsymbol{x})=c_k|y=c_j)\right]\hat{p}_{tst}^{PACC}(c=c_j) \\
&= \sum_{j=1}^{\#C}\left[\frac{1}{\#U_j}\sum_{\boldsymbol{x}\in U_j}\hat{\mathbb{P}}(h_{tr}(\boldsymbol{x})=c_k)\right]\hat{p}_{tst}^{PACC}(c=c_j)
\end{aligned} \tag{2.5}$$

donde $U_j = \{(\boldsymbol{x}, y) \in (\boldsymbol{X}_{tst}, Y_{tst})|y = c_j\}$. Luego, $\hat{p}_{tst}^{PCC}(c = c_k)$ se calcula mediante PCC y, como en ACC, las $[\frac{1}{\#U_j}\sum_{\boldsymbol{x}\in U_j}\hat{\mathbb{P}}(h_{tr}(\boldsymbol{x}) = c_k)]$ deben estimarse mediante validación cruzada, quedando nuevamente un sistema de ecuaciones lineales de $\#C$ ecuaciones y $\#C$ incógnitas.

Para el caso particular binario, y relacionando con la ecuación 2.2, tenemos:

$$\hat{p}_{tst}^{PACC}(c=1) = \frac{\hat{p}_{tst}^{PCC}(c=1) - \hat{fp}_{pa}}{\hat{tp}_{pa} - \hat{fp}_{pa}} \tag{2.6}$$

donde tp_{pa} y fp_{pa} (pa: *probability average*) son los dos parámetros propios del cuantificador a estimar mediante validación cruzada, siendo tp_{pa} el promedio de las probabilidades *a posteriori* para la clase positiva estimadas por el clasificador correspondientes a los individuos cuya etiqueta es positiva, y del mismo modo fp_{pa} pero para individuos con etiqueta negativa. En este método hay que tener en cuenta ambas consideraciones sobre las estimaciones de $\hat{p}$ dentro del rango $[0, 1]$ y sobre la calibración -ver A-.

Ejemplo: Del mismo modo que para el ejemplo de PCC, se separó de la muestra de entrenamiento un 15 % de datos para realizar la calibración del clasificador blando. Pero también se separó otro 15 % para realizar el ajuste del propio método de cuantificación. Con el 70 % de datos se entrenó el clasificador que luego fue calibrado usando el primer 15 % separado, obteniendo un $\hat{p}_{tst}^{PCC}(c = 1) \approx 0.51$. Luego, con el segundo 15 % de datos, se procedió a estimar tp_{pa} y fp_{pa}. Teniendo en cuenta entonces ahora tanto las salidas del clasificador calibrado como las etiquetas de la muestra, tenemos:

$s(\boldsymbol{x}_i, y = 0)$	$s(\boldsymbol{x}_i, y = 1)$	c
0.23	0.77	1
0.78	0.22	0
0.40	0.60	1
0.43	0.57	1
0.32	0.68	1
...		
0.58	0.42	0

siendo entonces $\hat{tp_{pa}} \approx 0.551$ y $\hat{fp_{pa}} \approx 0.548$, por lo que $\hat{p}_{tst}^{PACC}(c = 1) \approx 1.00$ (teniendo que haber truncado).

Selección de umbrales (TH)

Cuando los datos de entrenamiento presentan un desbalance significativo (generalmente los casos positivos son los escasos), la precisión de ACC se ve considerablemente afectada [27]. En estas situaciones, el clasificador tiende a favorecer la predicción de la clase mayoritaria (negativa), lo que disminuye la cantidad de fp pero a expensas de un bajo tpr. Esto se traduce en un denominador reducido en la ecuación 2.2, lo que hace que el método sea más sensible a las estimaciones de tpr y fpr.

Esta serie de métodos se fundamenta en la elección de un umbral que reduzca la varianza en las estimaciones de tpr y fpr, o de $tpr - fpr$. La premisa es identificar un umbral que aumente el número de tp, aunque generalmente esto conlleve un incremento fpr. Además, siempre que $tpr \gg fpr$, el denominador en 2.2 aumenta, lo que resulta en métodos más robustos ante pequeños errores en las estimaciones de tpr y fpr. Siguiendo esta lógica, Forman [18, 27] propone una serie de métodos basados en clasificadores que entreguen *scores* (no necesariamente probabilísticos ni calibrados) con distintas estrategias de selección de umbrales[3]:

- MAX: selecciona el umbral que maximiza $tpr - fpr$. Esto resulta en el mayor denominador posible en la ecuación 2.2 para el clasificador entrenado, lo que suaviza las correcciones.
- X: busca obtener $fpr = 1 - tpr$ para evitar los extremos de las curvas de fpr y de $1 - tpr$.
- T50: elige el umbral con $tpr = 0.5$, asumiendo que los positivos conforman la clase minoritaria. El objetivo es nuevamente evitar los extremos de la curva tpr.
- *Median Sweep* (MS): adopta un enfoque conjunto, calculando la prevalencia para todos los umbrales que modifiquen los posibles valores de fpr y tpr, y devolviendo la mediana de estas prevalencias como la predicción final.

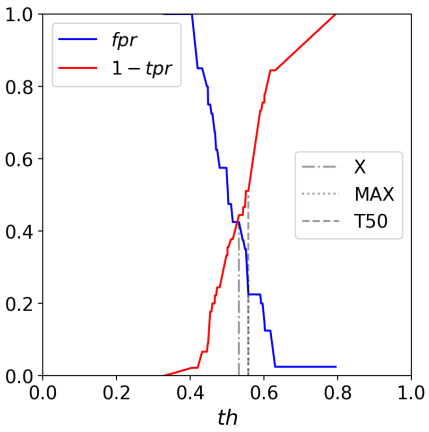


Figura 2.1: Curvas de ejemplo de fpr y $1 - tpr$ utilizadas visualizar la selección de umbrales según distintos criterios (MAX, X y T50). Estos umbrales son luego usados para definir la clase a predecir para cada individuo.

[3]Los métodos aquí se describen son exclusivamente de cuantificación binaria (las versiones multiclase no han sido abordadas en la literatura y no son sencillas de implementar).

Ejemplo: En la figura 2.2 se visualiza la selección de umbral según los criterios MAX, X y T50. Con estos umbrales, se computa luego la etapa de clasificación y, utilizando los correspondientes fpr y tpr, se utiliza la ecuación 2.2:

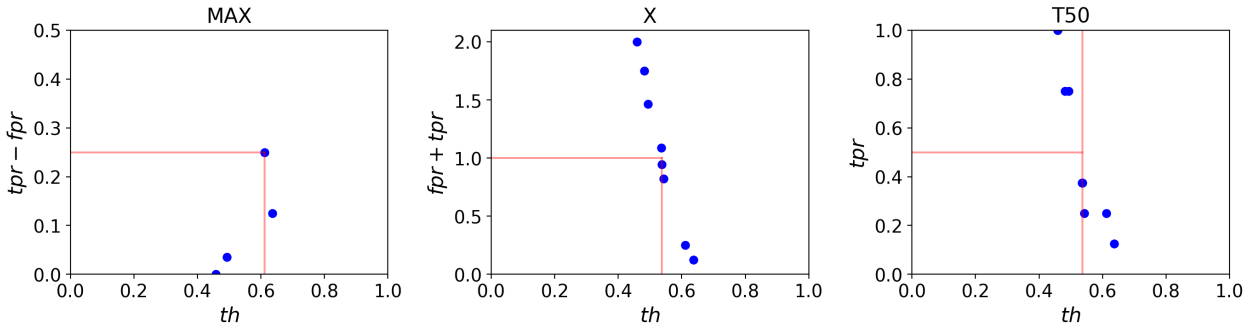


Figura 2.2: Selección de umbrales según distintos criterios (MAX, X y T50) para el ejemplo comparativo.

Para el criterio MS (figura 2.3), en cambio, por cada umbral que cambie el fpr o el tpr se calcula una prevalencia (se descartan los casos indeterminados por 2.2), y luego la mediana de todas ellas será la predicción final del método.

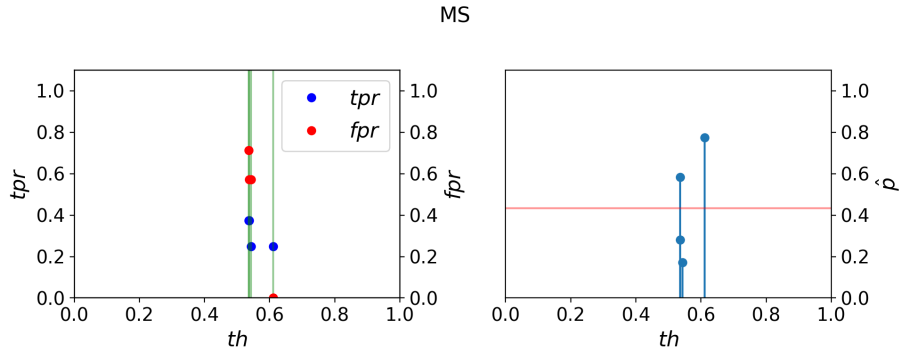


Figura 2.3: Para el criterio *Median Sweep*, primero se computan todos los umbrales que modifiquen ya sea fpr o tpr (izquierda), y luego se computa la mediana del conjunto de prevalencias obtenidas con esos umbrales (derecha).

Los resultados obtenidos fueron:

- $\hat{p}_{tst}^{MAX}(c=1) \approx 0.774$
- $\hat{p}_{tst}^{X}(c=1) \approx 0.282$
- $\hat{p}_{tst}^{T50}(c=1) \approx 0.282$
- $\hat{p}_{tst}^{MS}(c=1) \approx 0.433$

Esperanza-Maximización (EMQ)

Aunque este método se propuso originalmente para mejorar las probabilidades *a posteriori* de modelos de clasificación bajo *dataset shift* (ver 1.4), el mismo también sirve para mejorar la estimación de prevalencias. También conocido como *SLD* por las iniciales de sus autores, este método fue

propuesto por Saerens et al. [20] y aplica el algoritmo de Esperanza-Maximización (EM) [28], un conocido algoritmo iterativo para encontrar estimaciones de máxima verosimilitud de parámetros (los valores de prevalencia de clase) para modelos que dependen de variables no observadas (las etiquetas de clase). Esencialmente, *EMQ* actualiza incrementalmente las probabilidades *a posteriori* utilizando los valores de prevalencia de clases calculados en el último paso de la iteración, y actualiza los valores de prevalencia de clases utilizando las probabilidades *a posteriori* calculadas en el último paso de la iteración, de forma mutuamente recursiva, y tomando como punto de partida un valor determinado para la prevalencia de clases (generalmente el valor correspondiente a la muestra de entrenamiento o una estimación *a priori* dada por algún conocimiento de la muestra de prueba, aunque puede ser cualquier otro valor), y repitiendo las iteraciones hasta alcanzar la convergencia.

Saerens et al. [20, Apéndice, p.23 a p.25] demuestra, mediante el Teorema de Bayes y el Teorema de probabilidad total, que el algoritmo de EM aplicado a este problema resulta en los siguientes pasos (el paso 0 se aplica una sola vez, luego se iteran el E y M):

0 - Inicialización de $\hat{p}_Y^{(0)}(y = c_k)$, generalmente haciendo $\hat{p}_Y^{(0)}(y = c_k) = \hat{p}_{tr}(c = c_k)$

E - Esperanza:
$$\hat{p}_{Y|\boldsymbol{X}=\boldsymbol{x}_i}^{(s)}(y = c_k) = \frac{\frac{\hat{p}_Y^{(s)}(y = c_k)}{\hat{p}_{tr}(c = c_k)} s(\boldsymbol{x}_i, y = c_k)}{\sum_{j=1}^{\#C} \frac{\hat{p}_Y^{(s)}(y = c_j)}{\hat{p}_{tr}(c = c_j)} s(\boldsymbol{x}_i, y = c_k)}$$

M - Maximización:
$$\hat{p}_Y^{(s+1)}(y = c_k) = \frac{1}{m} \sum_{i=1}^{m} \hat{p}_{Y|\boldsymbol{X}=\boldsymbol{x}_i}^{(s)}(y = c_k)$$

Finalmente, cuando se alcanza la convergencia, se obtiene: $\hat{p}_{tst}^{EMQ}(c = c_k) = \hat{p}_Y(y = c_k)$. Aunque ya mencionamos que el modelo supone que las probabilidades *a posteriori* de modelos de clasificación ya están calibradas, se ha estudiado también que el método *EMQ* mejora las predicciones de cuantificación si el clasificador utilizado está calibrado [29, 30].

Ejemplo: Comenzamos con la inicialización, dando en nuestro caso como resultado $\hat{p}_Y^{(0)}(y = 1) = \hat{p}_{tr}(c = 1) = 0.5$ y $\hat{p}_Y^{(0)}(y = 0) = \hat{p}_{tr}(c = 0) = 0.5$. En la primera iteración, para el paso E queda $\hat{p}_{Y|\boldsymbol{X}=\boldsymbol{x}_i}^{(s=0)}(y = c_k) = s(\boldsymbol{x}_i, y = c_k)$ es decir, las mismas salidas del clasificador calibrado. Luego, para el paso M, y al igual que en *PCC*, debemos promediar cada salida individual del clasificador calibrado por cada una de las clases existentes, quedando en nuestro caso $\hat{p}_Y^{(s=1)}(y = 1) = p_{tst}^{PCC}(c = 1) \approx 0.527$. Ahora, en el paso E de la segunda iteración, se usará este último valor junto con los valores de prevalencia de la muestra de entrenamiento para ajustar las salidas del clasificador calibrado. Por ejemplo, para ajustar la salida del primer individuo correspondiente a la clase positiva, sería:

$$\hat{p}_{Y|\boldsymbol{X}=\boldsymbol{x}_i}^{(s=1)}(y = 1) \approx \frac{\frac{0.527}{0.5} 0.539}{\frac{0.527}{0.5} 0.539 + \frac{0.473}{0.5} 0.461} \tag{2.7}$$

Si continuamos repitiendo los pasos E y M de forma sucesiva, y definiendo un criterio de corte para la convergencia (ya sea por máxima cantidad de iteraciones o por un umbral de diferencia entre $\hat{p}_Y^{(s)}(y = c_k)$ y $\hat{p}_Y^{(s+1)}(y = c_k)$), se obtuvo $p_{tst}^{EMQ}(c = 1) \approx 0.164$.

Usando la distancia de Hellinger en y (HDy)

González-Castro et al. [12] proponen dos métodos fundamentados en la comparación de distribuciones. Aunque difieren en la manera de representar estas distribuciones, ambos comparten un elemento esencial: emplean la distancia de Hellinger como medida para cuantificar la disparidad entre ellas. El primer método, conocido como *HDy*, es un método agregativo, ya que emplea las salidas del clasificador para describir las distribuciones tanto de la muestra de entrenamiento como la de prueba. El método se basa en el cálculo de:

$$p_{tst}^{HDy}(c=1) = \underset{0\leq\alpha\leq 1}{\arg\min}\, \mathrm{HD}(\alpha f_{tr}(s(\boldsymbol{x}, y=1|c=1)) + (1-\alpha) f_{tr}(s(\boldsymbol{x}, y=1|c=0)), f_{tst}(s(\boldsymbol{x}, y=1))) \tag{2.8}$$

donde:

$$\mathrm{HD}(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i}-\sqrt{q_i})^2} \text{ con } P=(p_1,\ldots,p_k), Q=(q_1,\ldots,q_k) \tag{2.9}$$

y $f_{tr}(\circ)$ y $f_{tst}(\circ)$ son las funciones de densidad de probabilidad de las salidas del clasificador para la muestra de entrenamiento y de evaluación, respectivamente. Estas densidades son aproximadas empíricamente mediante histogramas, siendo k el número de *bins* utilizados. Dado que el número de *bins* k podría tener un impacto significativo en la estimación, normalmente se utiliza como estimador la mediana de la distribución de los α encontrados para un rango de k.

El segundo método propuesto por González-Castro et al. [12] pertenece a los métodos no agregativos y será desarrollado en la correspondiente sección 2.2.

Ejemplo: En la figura 2.4 vemos cómo son las estimaciones de tres distribuciones estimadas para el ejemplo, la gráfica de la función de costo usada, y cómo el mínimo encontrado se usa para combinar las distribuciones de entrenamiento y compararlas con la de prueba. En este caso, en vez de utilizar un rango de *bins*, se utilizó un solo valor, siendo $k = 200$. Se observa que el valor mínimo de HDy se da en $p_{tst}^{HDy}(c=1) = 0.87$.

2.1.2. Con clasificadores específicos

Los métodos presentados anteriormente implican el uso de un clasificador, a menudo seguido por una fase de ajuste para contrarrestar cualquier tendencia del clasificador a subestimar o sobreestimar las proporciones de clases. Los algoritmos discutidos en esta sección están específicamente diseñados con este propósito en mente: durante el entrenamiento, tienen en cuenta que el modelo será utilizado para cuantificar.

Minimización de pérdida explícita (ELM)

Esta familia de métodos se aplican, en principio, a la cuantificación binaria, pero son fácilmente extensibles a la cuantificación multiclase. La idea propuesta por Esuli y Sebastiani [31] es seleccionar una medida de rendimiento de cuantificación y entrenar un algoritmo de optimización para construir el modelo óptimo según esa medida. Las diferencias entre ellos se deben a la medida de rendimiento seleccionada y al algoritmo de optimización utilizado.

Esuli y Sebastiani [31–33] proponen utilizar SVM_{perf} [34] para optimizar la divergencia KL -ver 3.2-, mientras que Barranquero et al. [35] también emplean SVM_{perf} pero con una pérdida

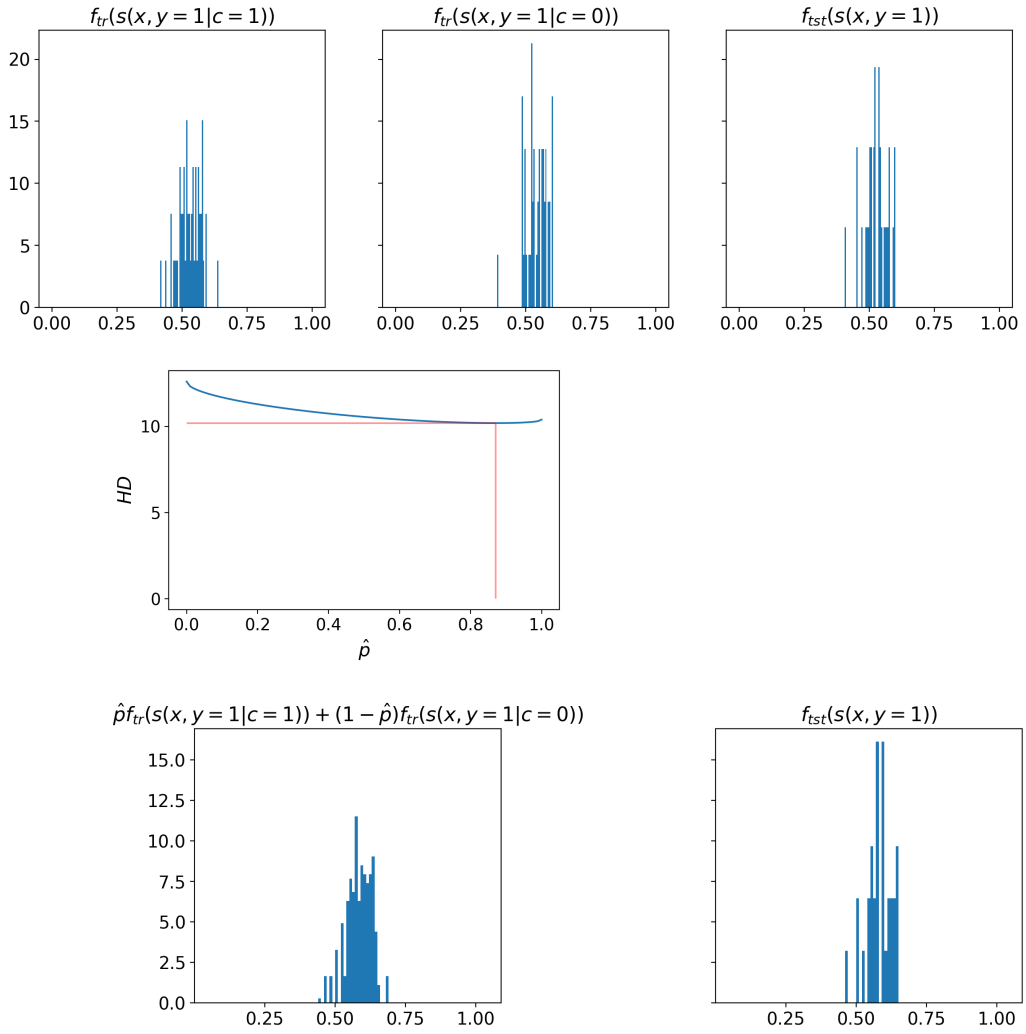


Figura 2.4: Histogramas de las distribuciones que conforman la ecuación 2.8, junto con la minimización de la función HD, y de las dos distribuciones con la distancia de Hellinger minimizada.

diferente, argumentando que la cuantificación pura no considera la precisión del clasificador subyacente (pudiendo generar un modelo que, aunque cuantifique bien, clasifique mal). Para abordar esto, introducen la medida Q, que combina una evaluación de cuantificación con una evaluación de clasificación, permitiendo un equilibrio entre ellas. Más recientemente Moreo y Sebastiani [36] reincorporaran la idea de utilizar SVM_{perf}, pero sugieren usar las medidas de evaluación de error absoluto (AE) y error absoluto negativo (RAE) -ver 3.2-.

Existen dos inconvenientes asociados con SVM_{perf}: podría resultar en un modelo subóptimo para la medida elegida, y no escala para grandes cantidades de datos de entrenamiento. Para abordar estas limitaciones, Kar et al. [37] proponen algoritmos de optimización estocástica. Además, plantean distintas medidas multivariadas para evaluar el rendimiento de cuantificación. Siguiendo esta línea, Sanyal et al. [38] introducen una serie de algoritmos que permiten el entrenamiento directo de redes neuronales profundas y la generación de clasificadores no lineales. Estos métodos están diseñados para optimizar funciones de pérdida de cuantificación como la divergencia KL.

Ejemplo: A diferencia de los casos anteriores, aquí no usamos el mismo clasificador con el que veníamos trabajando. En cambio, se ajusta un nuevo clasificador pero con una función de pérdida

más acorde al problema de cuantificación. A modo de ejemplo, usaremos el método propuesto por Esuli et al. [31], es decir, usando la divergencia KL como pérdida y SVM_{perf} como algoritmo de optimización. De esta forma, obtuvimos un $p_{tst}^{SVM_{perf},KLD}(c=1) \approx 0.613$, lo cual efectivamente implica una mejora del KLD con respecto a usar el clasificador anterior con *CC*, ya que $KLD_{tst}^{CC} \approx 0.887$ y $KLD_{tst}^{SVM_{perf},KLD} \approx 0.556$.

2.2. Métodos no agregativos

Hasta ahora, hemos utilizado métodos que agregan predicciones individuales de un clasificador para poder cuantificar. Sin embargo, también es posible estimar valores de prevalencia de clase sin generar decisiones binarias o probabilidades *a posteriori* para cada ítem. Esta alternativa se fundamenta en el principio de Vapnik, que sugiere resolver problemas directamente con la información disponible en lugar de abordar un problema más general. En cuantificación, esto significa que podemos estimar prevalencias de clase directamente sin clasificar cada individuo.

Usando la distancia de Hellinger en x (HDx)

Este método está obviamente relacionado con *HDy* (2.1.1), con la diferencia de considerar distribuciones de probabilidad multidimensionales $f(\boldsymbol{x})$ en lugar de distribuciones unidimensionales $f(s(\boldsymbol{x}))$. En vez de utilizar las salidas del clasificador, se estiman, con los datos de entrenamiento, las funciones densidad de las características de los individuos condicionadas a sus etiquetas.

Debido a la multidimensionalidad de $\boldsymbol{x} \in \mathbb{R}^d$, González-Castro et al. [12] proponen minimizar el promedio de las divergencias de Hellinger por cada $\boldsymbol{x}^j$:

$$p_{tst}^{HDx}(c=1) = \underset{0\leq\alpha\leq1}{\arg\min} \frac{1}{d}\sum_{j=1}^{d} \mathrm{HD}(\alpha f_{tr}(\boldsymbol{x^j}|c=1) + (1-\alpha) f_{tr}(\boldsymbol{x^j}|c=0), f_{tst}(\boldsymbol{x^j})) \tag{2.10}$$

Ejemplo: En la figura 2.5 repetimos el tipo de gráficos mostrados para el caso de *HDy* (figura 2.4), siendo en este caso $k = 10$. Se observa que el valor mínimo de HDx se da en $p_{tst}^{HDx}(c=1) = 0.7$.

Usando modelos generativos

Keith and O'Connor [39] presentan un enfoque utilizando modelos generativos, es decir, aquellos que modelan $p_{\boldsymbol{X},Y}$, a diferencia de los discriminativos, que modelan $p_{Y|\boldsymbol{X}=\boldsymbol{x}}$. Este método realiza una inferencia directa de la prevalencia desconocida, además de abordar el cálculo de intervalos de confianza[4].

Su propuesta, inspirada también en Saerens et al. [20], se basa en modelar la distribución conjunta de las características de los individuos y sus etiquetas. Para ello, se computa la verosimilitud marginal sobre θ (prevalencia de clases en la muestra de prueba) para obtener la distribución *a posteriori* de θ:

[4]Este enfoque es pionero en la cuantificación, ya que introduce un modelo directo para el cálculo de los intervalos de confianza, los cuales eran estimados mediante *bootstraping* en trabajos anteriores.

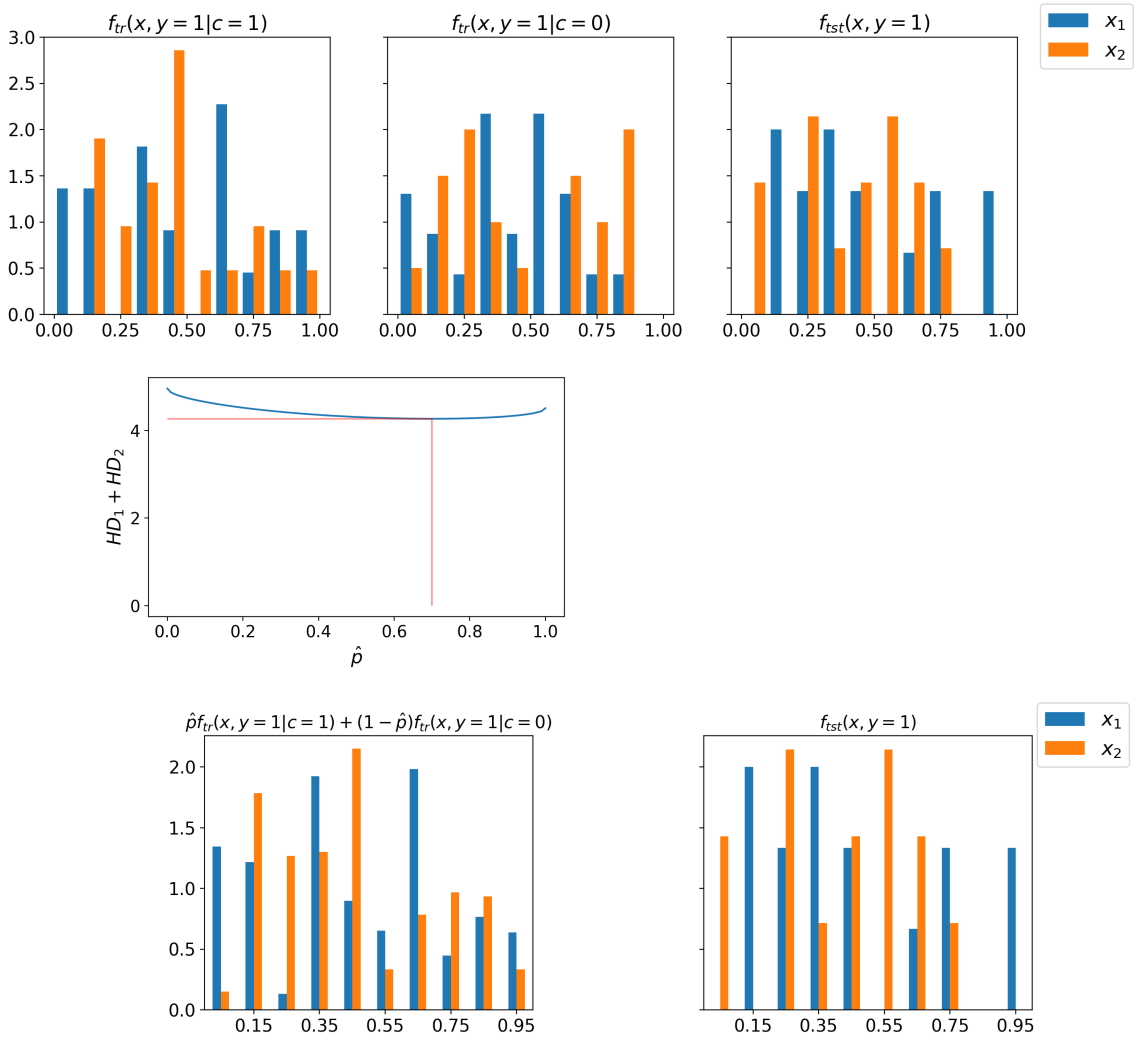


Figura 2.5: Histogramas de las distribuciones que conforman la ecuación 2.10, junto con la minimización de la función HD, y de las dos distribuciones con la distancia de Hellinger minimizada.

$$MLL(\theta) = \sum_{i=1}^{m} \log \sum_{c \in C} p_{\boldsymbol{X},Y|\theta}(\boldsymbol{x}_i, c) \tag{2.11}$$

siendo para el caso binario:

$$MLL(\theta) = \sum_{i=1}^{m} \log(p_{\boldsymbol{X},Y|\theta}(\boldsymbol{x}_i, 1) + p_{\boldsymbol{X},Y|\theta}(\boldsymbol{x}_i, 0)) \tag{2.12}$$

y usando la Ley de Probabilidad Condicionada con $p_{Y|\theta}(1) = \theta$ y $p_{Y|\theta}(0) = 1 - \theta$:

$$MLL(\theta) = \sum_{i=1}^{m} \log(\theta p_{\boldsymbol{X}|Y=1,\theta}(\boldsymbol{x}_i) + (1 - \theta) p_{\boldsymbol{X}|Y=0,\theta}(\boldsymbol{x}_i)) \tag{2.13}$$

y como se asume $X_i \perp\!\!\!\perp \theta \mid Y_i$, entonces:

$$MLL(\theta) = \sum_{i=1}^{m} \log(\theta p_{\boldsymbol{X}|Y=1}(\boldsymbol{x}_i) + (1 - \theta) p_{\boldsymbol{X}|Y=0}(\boldsymbol{x}_i)) \tag{2.14}$$

donde se aplica el supuesto de *prior probability shift*, es decir, $\mathbb{P}_{tr}(\boldsymbol{X} = \boldsymbol{x}|Y = y) = \mathbb{P}_{tst}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$, utilizando las distribuciones modeladas con los datos de entrenamiento para aplicarlas a los datos de prueba.

Luego, para obtener la predicción se obtiene el máximo de la distribución. Es decir, que al igual que el método *EMQ*, se busca maximizar la verosimilitud, pero en este caso no necesariamente utilizando el algoritmo *EM*. Esta función es unimodal en $\theta \in [0, 1]$. Como es cóncava y hay un solo parámetro, se pueden emplear muchas técnicas para encontrar la moda, incluyendo *EM*, Newton-Rapshon o computacionalmente mediante una grilla de valores.

Keith and O'Connor [39] proponen particularmente dos modelos generativos enfocados en problemas de procesamiento de lenguaje (*MNB y Loglin*), al que llaman explícitos. Pero aún más interesante es el tercer método que proponen, al que llaman *LR-Implicit*, el cual se basa en estimar de forma implícita las $p_{\boldsymbol{X}|Y=y}(\boldsymbol{x})$ mediante las $p_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ obtenidas con modelos discriminativos, utilizando el Teorema de Bayes:

$$p_{discY|\boldsymbol{X}=\boldsymbol{x}}(y) = \frac{p_{imp_{\boldsymbol{X}|Y=y}}(\boldsymbol{x}) p_{trY}(y)}{p_{\boldsymbol{X}}(\boldsymbol{x})}, \tag{2.15}$$

siendo $p_{discY|\boldsymbol{X}=\boldsymbol{x}}(1) = h_{tr}(\boldsymbol{x})$, entonces podemos utilizar en 2.14:

$$\begin{aligned} MLL(\theta) &= \sum_{i=1}^{m} \log\left(\theta \frac{h_{tr}(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}_i)}{p_{trY}(1)} + (1-\theta)\frac{(1-h_{tr}(\boldsymbol{x}))p_{\boldsymbol{X}}(\boldsymbol{x}_i)}{1-p_{trY}(1)}\right) \\ &= \sum_{i=1}^{m} \log\left(p_{\boldsymbol{X}}(\boldsymbol{x}_i)\left(\theta \frac{h_{tr}(\boldsymbol{x})}{p_{trY}(1)} + (1-\theta)\frac{(1-h_{tr}(\boldsymbol{x}))}{(1-p_{trY}(1))}\right)\right) \\ &= \sum_{i=1}^{m} \log(p_{\boldsymbol{X}}(\boldsymbol{x}_i)) + \sum_{i=1}^{m} \log\left(\theta \frac{h_{tr}(\boldsymbol{x})}{p_{trY}(1)} + (1-\theta)\frac{(1-h_{tr}(\boldsymbol{x}))}{(1-p_{trY}(1))}\right) \end{aligned} \tag{2.16}$$

donde para obtener el θ que maximiza la función, basta con analizar el segundo término de la ecuación.

Un modelo generativo que utiliza un clasificador discriminativo como intermediario para estimar $p_{\boldsymbol{X}|Y=y}(\boldsymbol{x})$ a partir de $p_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ (es decir, el método *LR-Implicit*) pertenece en realidad a los métodos agregativos (mencionados en la Sección 2.1). No obstante, dado que el marco generativo presentado por Keith and O'Connor [39] solo requiere un modelo condicionado por las etiquetas de clase, como ocurre con las versiones explícitas (usando los modelos *MNB y Loglin* como ejemplo), enmarcamos este método más general dentro de los métodos no agregativos.

Ejemplo: En este caso utilizaremos el método *LR-Implicit* aplicado al clasificador con el que venimos trabajando. Utilizaremos el método de grilla para buscar el máximo de la curva de $MLL(\theta)$ en la figura 2.6:

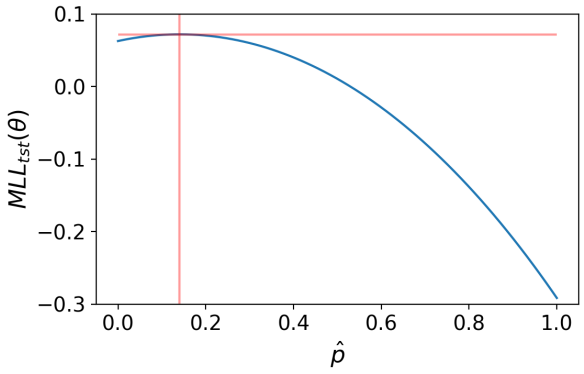


Figura 2.6: Curva de la función de máxima verosimilitud marginal (*MLL*) en función de θ para el método *LR-Implicit*. El máximo de la curva indica la prevalencia estimada para la clase positiva en la muestra de prueba.

Se observa que el máximo se da en $p_{tst}^{LR-Implicit}(c = 1) = 0.14$.

Capítulo 3

Métodos de evaluación

La evaluación de métodos de cuantificación es más compleja que en otros problemas. En aprendizaje supervisado, típicamente se mide el rendimiento estimando la probabilidad de predecir correctamente ejemplos individuales no observados. Sin embargo, en cuantificación, el rendimiento se evalúa para conjuntos de datos. Esto implica que necesitamos una colección de muestras para evaluar el rendimiento de un método. Dado un método $\overline{h}$, una función de pérdida $L(\cdot,\cdot)$, y un conjunto de muestras de evaluación $T_1,\ldots,T_s$, el rendimiento de $\overline{h}$ es:

$$Rendimiento(\overline{h}, L, T_1, \ldots, T_s) = \frac{1}{s}\sum_{j=1}^{s} L(\overline{h}, T_j) \tag{3.1}$$

Calcular la pérdida de un método de cuantificación sobre una muestra de prueba ($L(\overline{h}, T_j)$) no equivale a promediar o sumar pérdidas obtenidas a nivel individual (es decir, pérdidas para clasificación o regresión), debido a la interdependencia de estas últimas a nivel muestra. Por ejemplo, para el caso binario, si en una muestra de evaluación hay una mayor incidencia de falsos positivos que de falsos negativos, la presencia de un falso negativo adicional puede en realidad mejorar el error de cuantificación general, gracias al efecto de compensación mutua entre fp y fn mencionado en 1.5.

Además, el problema de evaluación en cuantificación se relaciona con el cambio en la distribución de datos entre la fase de entrenamiento y la de implementación del método. Se requiere una colección de muestras de prueba variada y que represente diversas distribuciones para evaluar correctamente el rendimiento del método y evitar sesgos. Por esta razón, la mayoría de los experimentos reportados en la literatura emplean conjuntos de datos tomados de otros problemas y se crean conjuntos de prueba con cambios en las distribuciones creados artificialmente. Este enfoque tiene la ventaja de que la cantidad del *dataset shift* se puede controlar para estudiar el rendimiento de los métodos en diferentes situaciones.

Las funciones de pérdida $L(\cdot,\cdot)$ serán elegidas de acuerdo al tipo de problema y al objetivo particular de la aplicación. Como ya se mencionó, el rendimiento de $\overline{h}$ será el promedio del resultado de la función de pérdida por cada muestra de evaluación, de acuerdo a la ecuación 3.1. Se han propuesto en la literatura distintas medidas de evaluación para problemas de *Single-Label Quantification (SLQ)*. Estas también se pueden usar para *Binary Quantification (BQ)*, ya que es un caso espacial del anterior, y para *Multi-Label Quantification*, ya que se pueden usar para cada $y \in C$. Esencialmente todas las medidas de evaluación que se han propuesto son divergencias, es

decir, medidas de cómo una distribución difiere de otra. No se desarrollarán en esta tesis medidas para *Ordinal Quantification* ni para *Regression Quantification*, ya que no son útiles para nuestro objeto de estudio.

3.1. Propiedades

Sebastiani [40] define una serie de propiedades interesantes para medidas de evaluación en *SLQ*. Un importante resultado de este artículo es que ninguna medida de evaluación existente para *SLQ* satisface todas las propiedades identificadas como deseables; aún así, se ha demostrado que algunas medidas de evaluación son "menos inadecuadas" que otras. Aquí mencionamos brevemente las cuatro propiedades principales que habría que considerar en cada medida M a emplear (el resto son propiedades que suelen ser satisfechas por todas las medidas).

- **Máximo (MAX)**: si $\beta > 0, \beta \in \mathbb{R}$ tal que por cada $c \in C$ y por cada p, (i) existe $\hat{p}$ tal que $M(p, \hat{p}) = \beta$, y (ii) para ninguna $\hat{p}$ se cumple que $M(p, \hat{p}) > \beta$. Si se cumple MAX, la imagen de M es independiente del problema, y esto permite juzgar si un valor dado significa un error de cuantificación alto o bajo. Si M no cumple cumple MAX, cada muestra de evaluación tendrá un peso distinto en el resultado final.

- **Imparcial (IMP)**: si M penaliza igualmente la subestimación de p por una cantidad a (es decir, con $\hat{p} = p - a$) o su sobreestimación por la misma cantidad a (es decir, con $\hat{p} = p + a$). Si se cumple IMP, la subestimación y la sobreestimación se consideran igualmente indeseables. Esto es generalmente lo deseable, a menos que exista una razón específica para no hacerlo.

- **Relativo (REL)**: si M penaliza más gravemente un error de magnitud absoluta a (es decir, cuando $\hat{p} = p \pm a$) si p es menor. Por ejemplo, predecir $\hat{p} = 0.011$ cuando $p = 0.001$ es un error mucho más serio que predecir $\hat{p} = 0.11$ cuando $p = 0.1$.

- **Absoluto (ABS)**: si M penaliza un error de magnitud independientemente del valor de p. Mientras algunas aplicaciones requieren REL, otras requieren ABS. Si bien REL y ABS son mutuamente excluyentes, no son redundantes ya que no abarcan todos los posibles casos (ninguna cubre el caso cuando M considera un error de magnitud absoluta a menos grave cuando p es menor).

3.2. Medidas de Evaluación

El rendimiento se calcula haciendo un promedio entre un conjunto de muestras de evaluación 3.1, por lo que la definición matemática de cada medida de evaluación (o función de pérdida) se reduce a obtener un "puntaje" para una sola muestra. Como ya mencionamos, todas las medidas de evaluación usadas en cuantificación son divergencias. Una divergencia D es una medida de cómo una distribución predicha $\hat{p}$ diverge (es decir, difiere) de la distribución real p, y se define de tal manera que (1) $D(p, \hat{p}) = 0$ si y solo si $p = \hat{p}$, y (2) $D(p, \hat{p}) > 0$ para todo $\hat{p} \neq p$. Para cuantificación binaria, el problema se reduce a la predicción de la proporción de la clase positiva, p. Así, para un modelo $\overline{h}$ y un conjunto de prueba dado T, solo necesitamos comparar la prevalencia predicha, $\hat{p}$, con la real, p.

Error (solo para Cuantificación Binaria)

El error técnicamente no es una medida de evaluación para la cuantificación, ya que no se aplica a toda una distribución p sino solo a una clase específica $c \in C$, y se define como:

$$\mathrm{e}(c) = \hat{p}(c) - p(c) \tag{3.2}$$

Incluso usado en cuantificación binaria, se debe especificar a cuál de las clases hace referencia (en este caso, como bien dijimos, suele hacer referencia a la clase positiva). Si se usa como una medida de evaluación para la cuantificación, un problema con el error es que promediar los valores para diferentes clases produce resultados poco intuitivos, ya que el error positivo de una clase y el error negativo de otra clase se anulan entre sí. El mismo problema ocurre cuando se trata de la misma clase pero se promedia entre diferentes muestras. Como resultado, esta medida se puede utilizar para determinar si un método tiene una tendencia a subestimar o sobrestimar la prevalencia de una clase específica (esto es, promediar el error para una clase particular entre distintas muestras, y de esta manera estimar el sesgo o *bias*) en BQ, pero no como una medida de evaluación general para usar.

Las siguientes medidas de evaluación se pueden emplear en cuantificación multiclase ya que, suman siempre valores positivos para todas las clases.

Error Absoluto

El error absoluto o *absolute error* es una de las medidas más empleadas ya que, al ser simplemente la diferencia entre ambas magnitudes, es simple y fácilmente interpretable.

$$\mathrm{AE}(p, \hat{p}) = \frac{1}{\#C} \sum_{j=1}^{\#C} |\hat{p}(c = c_j) - p(c = c_j)| \tag{3.3}$$

Como en este caso las diferencias positivas y negativas son igualmente indeseables, promediar el AE entre varias clases, o varias muestras, no es problemático. Como se muestra en [40], AE cumple IMP y ABS pero no cumple MAX ni REL. Su rango va de 0 (mejor) a:

$$z_{\mathrm{AE}} = \frac{2[1 - \min\limits_{j \in \{1,\ldots,\#C\}} p(c = c_j)]}{\#C} \tag{3.4}$$

(peor)[1] por lo que su rango depende de la distribución de p y de $\#C$.

Error Absoluto Normalizado

El error absoluto normalizado *normalised absolute error*, definido como:

$$\mathrm{NAE}(p, \hat{p}) = \frac{\mathrm{AE}(p, \hat{p})}{z_{\mathrm{AE}}} = \frac{\sum\limits_{j=1}^{\#C} |\hat{p}(c = c_j) - p(c = c_j)|}{2[1 - \min\limits_{j \in \{1,\ldots,\#C\}} p(c = c_j)]}, \tag{3.5}$$

es una versión de AE que oscila entre 0 (mejor) y 1 (peor), por lo que cumple MAX. NAE no cumple ABS (a pesar de su nombre) ni tampoco REL, pero sí IMP.

[1] Ocurre cuando toda la probabilidad predicha se asigna a la clase menos probable, es decir, cuando $\hat{p}(c = c_j) = 1$ para $c_j = \arg\min_j p(c = c_j)$ y $\hat{p}(c = c_k) = 0$ para $k \neq j$.

Error Cuadrático

El error cuadrático o *squared error*, definido como:

$$\mathrm{SE}(p, \hat{p}) = \frac{1}{\#C} \sum_{j=1}^{\#C} (\hat{p}(c = c_j) - p(c = c_j))^2, \tag{3.6}$$

comparte los mismos pros y contras de AE, pero penalizando más cuanto mayor es la diferencia entre el valor real y el predicho, por lo que se usa cuando se quiere castigar los valores atípicos u *outliers*. SE cumple IMP y ABS pero no cumple MAX ni REL.

Error Absoluto Relativo

El error absoluto relativo o *relative absolute error* es una adaptación del AE que impone REL al hacer que AE sea relativo a p.

$$\mathrm{RAE}(p, \hat{p}) = \frac{1}{\#C} \sum_{j=1}^{\#C} \frac{|\hat{p}(c = c_j) - p(c = c_j)|}{p(c = c_j)} \tag{3.7}$$

RAE cumple IMP y REL pero no cumple MAX ni ABS. Su rango va de 0 (mejor) a:

$$z_{\mathrm{RAE}} = \frac{\#C - 1 + \dfrac{1 - \min\limits_{j \in \{1,\ldots,\#C\}} p(c = c_j)}{\min\limits_{j \in \{1,\ldots,\#C\}} p(c = c_j)}}{\#C} \tag{3.8}$$

(peor)[2], por lo que su rango depende de la distribución de p y de $\#C$.

Error Absoluto Relativo Normalizado

El error absoluto relativo normalizado *normalised relative absolute error*, definido como:

$$\mathrm{NRAE}(p, \hat{p}) = \frac{\mathrm{RAE}(p, \hat{p})}{z_{\mathrm{RAE}}} = \frac{\sum\limits_{j=1}^{\#C} \frac{|\hat{p}(c=c_j) - p(c=c_j)|}{p(c=c_j)}}{\#C - 1 + \dfrac{1 - \min\limits_{j \in \{1,\ldots,\#C\}} p(c = c_j)}{\min\limits_{j \in \{1,\ldots,\#C\}} p(c = c_j)}}, \tag{3.9}$$

es una versión de RAE que oscila entre 0 (mejor) y 1 (peor), por lo que cumple MAX. NRAE no cumple REL ni ABS, pero sí IMP.

Tanto RAE como NRAE no están definidas cuando sus denominadores sean nulos. Para resolver este problema, se puede suavizar tanto $p(c = c_j)$ como $\hat{p}(c = c_j)$ mediante suavizado aditivo:

$$\underline{p}(c = c_j) = \frac{\epsilon + p(c = c_j)}{\epsilon \#C + \sum\limits_{j=1}^{\#C} p(c = c_j)}, \tag{3.10}$$

donde $\underline{p}(c = c_j)$ es la versión suavizada de $p(c = c_j)$ y el denominador es solo un un factor de normalización (lo mismo para $\underline{\hat{p}}(c = c_j)$).

[2]Ocurre cuando toda la probabilidad predicha se asigna a la clase menos probable, es decir, cuando $\hat{p}(c = c_j) = 1$ para $c_j = \arg\min_j p(c = c_j)$ y $\hat{p}(c = c_k) = 0$ para $k \neq j$.

Divergencia de Kullback-Leibler

Para distribuciones de probabilidad discretas P y Q definidas en el mismo espacio muestral $\mathcal{X}$, su divergencia KL se define como:

$$\mathrm{DKL}(P,Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right). \tag{3.11}$$

En cuantificación, se quiere comparar la prevalencia real p y la prevalencia predicha $\hat{p}$, y el espacio muestral corresponde a las posibles clases, con lo cuál será:

$$\mathrm{DKL}(p,\hat{p}) = \sum_{j=1}^{\#C} p(c=c_j)\log\left(\frac{p(c=c_j)}{\hat{p}(c=c_j)}\right), \tag{3.12}$$

que va de 0 (mejor) a $+\infty$ (peor) -por lo tanto, no cumple con MAX. Esta medida no es simétrica, es menos interpretable que otras medidas de rendimiento, y no está definida cuando $\hat{p}$ es 0 o 1.

Divergencia de Kullback-Leibler Normalizada

Para suplir los problemas de DKL, se puede utilizar la función logística, quedando:

$$\mathrm{NDKL}(p,\hat{p}) = 2\cdot\frac{e^{\mathrm{DKL}(p,\hat{p})}}{1+e^{\mathrm{DKL}(p,\hat{p})}} - 1, \tag{3.13}$$

que va de 0 (mejor) a 1 (peor) -por lo tanto, cumple con MAX-. Sin embargo, como se muestra en [40], ni DKL ni NDKL cumplen con IMP, REL y ABS, lo que hace que su uso como medidas de evaluación para cuantificación sea cuestionable, además de ser difíciles de interpretar.

3.3. Elección de la medida de evaluación

Es evidente que ninguna de las medidas propuestas hasta ahora satisface todas las propiedades. DKL y NDKL son los menos satisfactorios y parecen fuera de discusión. Respecto a las demás, el problema es que MAX parece ser incompatible con REL/ABS, y viceversa.

Sebastiani [40] sostiene que cumplir con REL o ABS parece más importante que cumplir con MAX. La elección de REL o ABS es importante para que en la evaluación se refleje si un error de cuantificación de una magnitud absoluta dada es más grave cuando la verdadera prevalencia de la clase afectada es menor o no. En cambio, si no se satisface MAX, algunas muestras tendrán mayor peso que otras en el resultado final; si bien esto no es deseable, no afecta la comparación experimental entre diferentes sistemas de cuantificación, ya que cada uno de ellos se ve afectado por esta disparidad de la misma manera.

Por lo tanto, si aceptamos la idea de "sacrificar" MAX para conservar REL o ABS, esto sugiere que AE, RAE y SE son las mejores medidas a elegir. Se debe preferir AE cuando un error de estimación de una magnitud absoluta dada debe considerarse más grave cuando la verdadera prevalencia de la clase afectada es menor. RAE debe ser elegido cuando un error de estimación de una magnitud absoluta dada tiene el mismo impacto independientemente de la verdadera prevalencia de la clase afectada. Si se quiere penalizar mayormente errores atípicos, considerando mucho más graves a los errores cuanto mayor es la diferencia entre el valor real y el predicho, entonces SE es la medida más conveniente.

Tabla 3.1: Resumen de medidas de evaluación para cuantificación y sus propiedades principales.

Medida	MAX	IMP	REL	ABS	Uso recomendado
AE	–	✓	–	✓	Cuando se desea penalizar errores absolutos de igual manera, independientemente de la prevalencia.
NAE	✓	✓	–	–	Cuando se requiere comparar resultados entre distintos tipos de muestras.
SE	–	✓	–	✓	Cuando se desea penalizar fuertemente los errores grandes.
RAE	–	✓	✓	–	Cuando los errores son más graves si la prevalencia real es baja.
NRAE	✓	✓	–	–	Cuando se requiere comparar resultados entre distintas tipos de muestras.
DKL	–	–	–	–	Poco recomendable; difícil de interpretar y no cumple propiedades deseables.
NDKL	✓	–	–	–	Poco recomendable; sólo si se requiere rango [0,1] y se acepta baja interpretabilidad.

3.4. Protocolos

Mientras que en la clasificación, un conjunto de datos de tamaño k proporciona k puntos de evaluación, para la cuantificación, el mismo conjunto solo proporciona 1 punto. Evaluar algoritmos de cuantificación es, por lo tanto, un reto, debido a que la disponibilidad de datos etiquetados con fines de prueba es más restringido. Hay principalmente dos protocolos experimentales que se han tomado para tratar con este problema: el Protocolo de Prevalencia Natural (*NPP*) y el Protocolo de Prevalencia Artificial (*APP*).

- *NPP*: Consiste en, una vez entrenado un cuantificador, tomar un conjunto de prueba (no observado en el entrenamiento) lo suficientemente grande, dividirlo en un número de muestras de manera uniformemente aleatoria, y llevar a cabo la evaluación individualmente en cada muestra.

- *APP*: Consiste en, previo al entrenamiento, tomar un conjunto de datos, dividirlo en un conjunto de entrenamiento y en un conjunto de evaluación de manera aleatoria, y realizar experimentos repetidos en los que la prevalencia del conjunto de entrenamiento o la prevalencia del conjunto de prueba de una clase se varía artificialmente a través del submuestreo.

Ambos protocolos tienen diferentes pros y contras. Una ventaja de *APP* es que permite crear muchas puntos de prueba de la misma muestra. Además, *APP* permite simular distintos *Prior probability shift*, mientras que con *NPP* se estaría evaluando sólo con las distribuciones originales de los datos de entrenamiento y prueba. Sin embargo, una desventaja de *APP* es que puede no saberse

cuán realistas son estas diferentes situaciones en la aplicación real, por lo que se podría estar destinando recursos a una evaluación errónea o pobre. Una solución intermedia podría ser utilizar un protocolo que utilice conocimientos previos sobre la distribución de prevalencias "probables" que se podría esperar encontrar en el dominio específico en cuestión.

3.5. Selección de modelos

El rendimiento de muchos algoritmos de aprendizaje automático es altamente sensible a la configuración de sus hiperparámetros. Estos hiperparámetros, a diferencia de los parámetros, no se aprenden durante el entrenamiento, debiendo ser ajustados previamente para cada problema específico. Aunque muchos algoritmos ofrecen valores predeterminados, la optimización de estos hiperparámetros es esencial para maximizar el rendimiento en aplicaciones concretas. Los métodos de cuantificación no son una excepción en este sentido [16].

El proceso de selección de modelos consiste en evaluar diferentes combinaciones de hiperparámetros hasta quedarse con la combinación que obtenga la mejor evaluación. Para garantizar una evaluación rigurosa, es recomendable utilizar validación cruzada. En el caso de la cuantificación, la optimización de hiperparámetros debería imitar el protocolo de evaluación al evaluar cada una de las configuraciones candidatas. En otras palabras, dado que el objetivo de la selección de modelos es encontrar la configuración de hiperparámetros que funcione mejor de acuerdo con un protocolo experimental dado y una medida de evaluación dada, resulta adecuado adoptar los mismos protocolos de evaluación (3.4) y medidas de evaluación (3.2) que se usan habitualmente en la evaluación de sistemas de cuantificación [41, 42].

Algunos de los métodos de cuantificación que presentan hiperparámetros *per-se* y que hemos mencionado en el Capítulo 2 son el *HDy* y *HDx* (con respecto a la cantidad de *bins*) y *ACC* y *PACC* para el caso multiclase (con respecto al método para resolver el sistema de ecuaciones lineales), entre otros. Adicionalmente, todos los métodos agregativos (2.1) heredan los hiperparámetros de los clasificadores que emplean. En este sentido, Moreo y Sebastiani [36] afirman que el método *CC* y sus variantes, con una adecuada optimización, pueden ofrecer un rendimiento competitivo, aunque siguen siendo inferiores a los métodos de cuantificación más sofisticados, además de requerir recursos y tiempo de cómputo para la búsqueda de hiperparámetros.

Capítulo 4

Experimentos

En esta sección se evaluarán todos los métodos mencionados en el Capítulo 2 mediante nuevos casos de ejemplo simulados. Dicha evaluación se basará en las medidas de evaluación elegidas en 3.3 (RAE, AE y SE) además del error. Además, la simulación se basará en el protocolo *APP* definido en 3.4. No se realizará el proceso de selección de modelos mencionado en 3.5 ya que el objetivo de estos experimentos no es el de buscar los mejores cuantificadores para los casos de ejemplo, sino hacer una comparación de los mismos frente a iguales condiciones (hiperparámetros por defecto, mismo clasificador, etc., ya que esto permite una comparación justa y evita introducir sesgos debidos a la optimización específica de cada método). Finalmente, se elaboran conclusiones en base a los resultados y se comparan con los resultados de otros trabajos [36, 41, 43–45]. El código en Python para realizar las simulaciones se encuentra en el repositorio https://github.com/maxi-marufo/tesis.

4.1. Simulación

4.1.1. Poblaciones

La simulación consiste, en primer lugar, en generar dos poblaciones sintéticas de datos. Al igual que en 2, para sintetizar los datos se utiliza el algoritmo propuesto por Guyon et al. [23] mediante la función `make_classification` de `scikit-learn`. En este caso, se crean dos poblaciones, F (fácil) y D (difícil), cuyos argumentos de creación son (los parámetros no mencionados usan sus valores por defecto):

	Población	
	F	D
n_samples	51000	51000
n_features	2	100
n_informative	2	5
n_redundant	0	15
n_repeated	0	15
flip_y	0	0
class_sep	0.6	0.2

Tabla 4.1: Poblaciones.

La elección de estos valores tiene como objetivo generar una población fácil de clasificar y con características de baja dimensionalidad, y otra más difícil y con alta dimensionalidad, para probar bajo estas dos condiciones cada método. Además, al usar los parámetros *n_classes* y *weights* por defecto, estaremos generando datasets binarios y perfectamente balanceados.

4.1.2. Datasets

Tenemos entonces dos poblaciones sintéticas, cada una de 51000 individuos. Luego, procedemos a sub-dividir ambas poblaciones de forma estratificada (es decir, manteniendo la proporción de clases balanceadas) para crear los datasets de entrenamiento y de prueba en cada población. Dichos datasets son de 50000 y 1000 individuos respectivamente.

Siguiendo el protocolo *APP* ya mencionado, la simulación consiste en muestrear de forma iterativa el dataset de entrenamiento de cada población con distintos tamaños de muestra (500 y 5000), distintas prevalencias[1] (0.01, 0.25, 0.5, 0.75 y 0.99) y de forma repetida (5 veces) -es decir, un total de 50 muestras distintas-. Luego, por cada muestra de entrenamiento, a su vez, se muestrea de forma iterativa el dataset de evaluación. En este caso también se toman distintos tamaños de muestra (10 y 100), distintas prevalencias (0, 0.2, 0.5, 0.8 y 1.0 para las muestras de tamaño $n = 10$, y 0.01, 0.25, 0.5, 0.75 y 0.99 para las muestras de tamaño $n = 100$) y de forma repetida (5 veces) -50 muestras de prueba por cada muestra de entrenamiento-.

<table>
<tr><td></td><td colspan="4">Dataset</td></tr>
<tr><td></td><td colspan="2">Train</td><td colspan="2">Test</td></tr>
<tr><td>n_samples</td><td>500</td><td>5000</td><td>10</td><td>100</td></tr>
<tr><td rowspan="5">prev</td><td colspan="2">0.01</td><td>0</td><td>0.01</td></tr>
<tr><td colspan="2">0.25</td><td>0.2</td><td>0.25</td></tr>
<tr><td colspan="2">0.5</td><td>0.5</td><td>0.5</td></tr>
<tr><td colspan="2">0.75</td><td>0.8</td><td>0.75</td></tr>
<tr><td colspan="2">0.99</td><td>1.0</td><td>0.99</td></tr>
<tr><td>n_repetitions</td><td colspan="2">5</td><td colspan="2">5</td></tr>
</table>

Tabla 4.2: Datasets.

4.1.3. Cuantificación

Por cada muestra de entrenamiento, se procede a ajustar cada uno de los métodos a evaluar. Empezando por los métodos agregativos que utilizan clasificadores generales (2.1.1), tenemos los que requieren como entrada las etiquetas de las clases predichas (es decir, que usan clasificadores duros). En nuestro caso, estos métodos son el *CC* y el *ACC*. Es decir, que primero debemos ajustar un modelo de clasificación. En esta simulación hemos utilizado dos modelos distintos, un modelo de regresión logística y el clasificador de XGBoost, con la idea de probar un modelo simple y uno complejo, con el objetivo de determinar si la complejidad del modelo mejora o no la cuantificación. En este caso no es necesario calibrarlos ya que usaremos solamente las etiquetas predichas. Para sus hiperparámetros usaremos los valores por defecto de las librerías `scikit-learn` y `xgboost` respectivamente.

[1]Para evitar repeticiones, cuando mencionemos prevalencia en estos casos se hará referencia siempre a la clase positiva.

Clasificación	
Modelo	Complejidad
LogisticRegression	Baja
XGBoost	Alta

Tabla 4.3: Modelos de clasificación.

En cambio, para los métodos agregativos que utilizan clasificadores generales pero que requieren como entrada las probabilidades *a posteriori* de pertenencia a cada clase (es decir, clasificadores blandos), debemos no solo ajustar los modelos de clasificación, sino que también conviene calibrarlos (ya que es un supuesto de estos métodos). En nuestro caso, estos métodos son el *PCC*, *PACC*, *EMQ* y *HDy*. Con respecto a los algoritmos de calibración, utilizaremos los cuatro métodos para modelos binarios propuestos por Guo et al. [46] y mencionados en el Apéndice A, además de probar también los clasificadores sin calibrar. Como los métodos de calibración requieren utilizar validación cruzada para ajustar sus parámetros, utilizaremos el método de *held-out*, es decir que destinaremos una porción (20 %) del dataset de entrenamiento para ello.

Calibración	
Método	Held-out
No Calibration	0 %
Histogram Binning	20 %
Isotonic Regression	20 %
BBQ	20 %
Platt Scaling	20 %

Tabla 4.4: Modelos de calibración.

Para los métodos agregativos que utilizan clasificadores específicos (2.1.2) -en nuestro caso, únicamente el método *ELM*-, debemos también ajustar un modelo de clasificación previo a la cuantificación, pero en este caso no utilizaremos los nombrados en 4.3, sino uno optimizado para ser utilizado en cuantificación. En este caso usaremos, al igual que en el Capítulo 3, el clasificador SVM_{perf}, pero esta vez optimizado no solo para KLD, sino también AE y RAE.

Para los modelos no agregativos (2.2), no se usan modelos de clasificación, y por lo tanto tampoco se calibran; simplemente se aplica directamente el método de cuantificación. En nuestro caso, para este grupo usaremos solo el método *HDx*, ya que no usaremos modelos de clasificación generativos explícitos. Sin embargo, sí usaremos el método *LR-Implicit*, el cual, como bien dijimos, se debe considerar en realidad un método agregativo, y así lo haremos en las simulaciones.

Cabe mencionar que para los métodos de cuantificación que requieren de validación cruzada para estimar sus parámetros internos también utilizaremos el método de *held-out*, destinando otra porción (20 %) del dataset de entrenamiento para ello.

Cuantificación	
Método	Held-out
CC	0 %
ACC	20 %
PCC	0 %
PACC	20 %
TH_MAX	20 %
TH_X	20 %
TH_T50	20 %
TH_MS	20 %
EMQ	0 %
HDy	20 %
HDx	0 %
ELM-SVMperfKLD	0 %
ELM-SVMperfAE	0 %
ELM-SVMperfRAE	0 %
LR-Implicit	0 %

Tabla 4.5: Modelos de cuantificación.

Finalmente, una vez ajustado el método de cuantificación para una muestra de entrenamiento, se procede a tomar una muestra de prueba (recordando que probaremos con distintos tamaños de muestra, prevalencia y repitiendo el muestreo) y estimar su prevalencia.

Para la implementación de los métodos de cuantificación en la simulación se utiliza la librería `Quapy` [43] (`https://github.com/HLT-ISTI/QuaPy`), excepto para el método *LR-Implicit* [39], para el cual se emplea el código provisto por los autores (`https://github.com/slanglab/freq-e`). Para la implementación de los modelos de calibración utilizamos la librería `net:cal` [47] (`https://github.com/EFS-OpenSource/calibration-framework`).

4.1.4. Evaluación

Como ya se mencionó, se realiza la evaluación basándose en las medidas elegidas en 3.3, es decir, se utilizan RAE, AE y SE. La evaluación se hizo según distintos criterios, agrupando los resultados basándonos en estos criterios y calculando su estimación puntual (promedio) e intervalo de confianza del 95 (usando la distribución t de Student) %. Además, en los casos que corresponda, usaremos también la medida de clasificación F1 y la medida de calibración ECE para elaborar conclusiones.

4.2. Conclusiones

Empezamos por analizar el rendimiento de los cuantificadores en general. En la tabla B.1 se muestran los resultados de las estimaciones de las medidas RAE, AE y SE para cada cuantificador usado en la simulación. Los métodos basados en selección de umbrales (*TH*), *PACC*, *LR-Implicit* y *EMQ* son los que en general mejor desempeño tienen, mientras que *CC*, *PCC* y los basados en minimización de pérdida explícita (*ELM*) son los peores. Estos resultados son congruentes con los de Schumacher et al. [45] (en donde *TH_MS* es el método con mejor AE), los de Moreo and Sebastiani [36] (que afirma que *PACC* es el mejor método dentro de las variantes de *CC*), los de Moreo et al. [43] (donde *EMQ* resulta el método con mejor AE), los de Moreo and Sebastiani

[41] (donde también *EMQ* y *PACC* resultan los mejores métodos evaluados con AE) y los de Tasche [44] (que concluye que los métodos *CC* y *PCC* son limitados frente a *ACC* y *PACC*).

Si queremos ver si los métodos sobre o sub estiman las prevalencias, debemos entonces analizar los errores (y principalmente, el sesgo o *bias*). Podemos observar en la figura 4.1 que *HDy* está levemente sesgado a subestimar, mientras que el resto de los métodos no presentan, en general, un sesgo marcado. Sin embargo, si analizamos el sesgo en función de la verdadera prevalencia en la muestra de prueba (figura 4.2), todos los métodos si tienen un sesgo en favor de la clase minoritaria (excepto justamente para algunos valores de *HDy*).

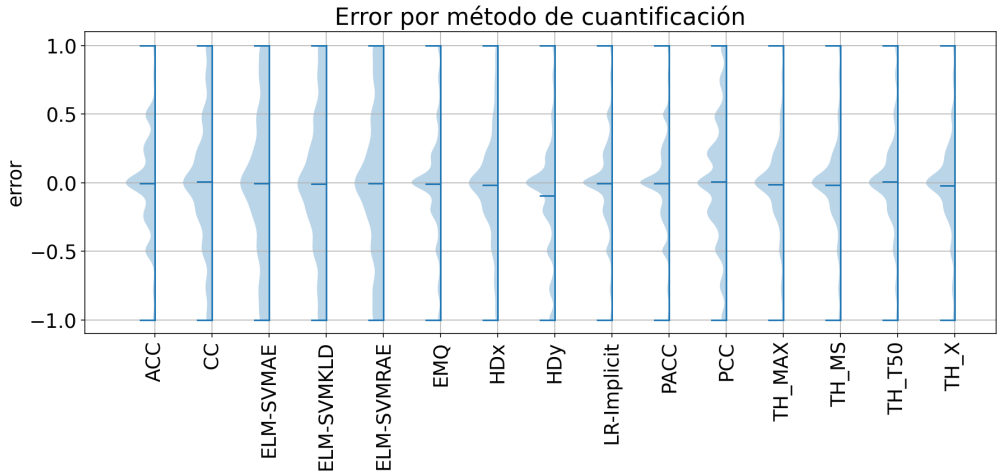


Figura 4.1: Error por método de cuantificación. HDy tiende a subestimar levemente la prevalencia, mientras que el resto no presenta sesgos marcados, pero sí diferencias en sus desvíos.

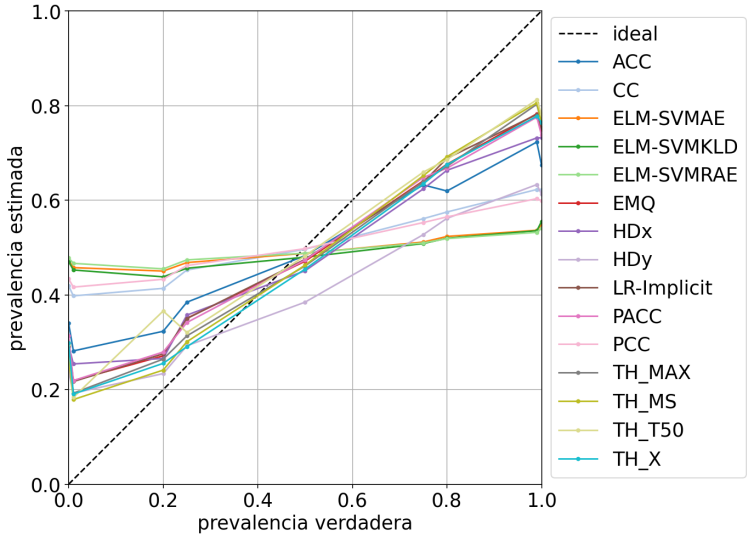


Figura 4.2: Prevalencia estimada por método de cuantificación según la verdadera prevalencia de prueba. La mayoría tiende a sobrestimar la clase menos representada en la muestra de prueba.

Algunas otras conclusiones que podemos sacar de los resultados son:

- En los métodos de cuantificación agregativos con clasificadores generales (2.1.1), a mejor desempeño en la clasificación (según su F1), mejor desempeño en la cuantificación (tabla B.2).

- A menor cantidad de características y mayor separación entre clases en las poblaciones, mejor rendimiento de cuantificación (tabla B.3).
- A mayor tamaño en muestra de entrenamiento, mejor rendimiento de cuantificación (tabla B.4) (coincide con los resultados de Schumacher et al. [45]).

Vemos también que cuanto más balanceada sea la muestra de entrenamiento, mejores resultados de cuantificación se obtienen (tabla B.5 y figuras 4.3 y 4.4). Este es un dato bastante interesante, ya que implicaría que podríamos aplicar en cuantificación las mismas técnicas de balance de clases que se usan en problemas de clasificación para tratar de conseguir mejores resultados que si usáramos datos desbalanceados. También vemos como la prevalencia de la muestra de entrenamiento influye en el sesgo del cuantificador, tendiendo en general a sobrestimar la clase mayoritaria. En la figura 4.5 podemos ver estos resultados desagregados por método de cuantificación.

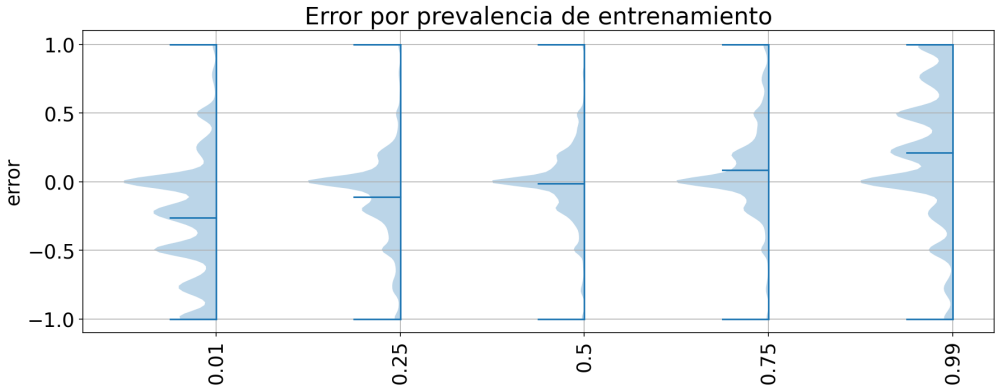


Figura 4.3: Error por prevalencia de entrenamiento. Los cuantificadores tienden a sobrestimar, en la muestra de evaluación, la prevalencia de clase que es mayoritaria en la muestra de entrenamiento.

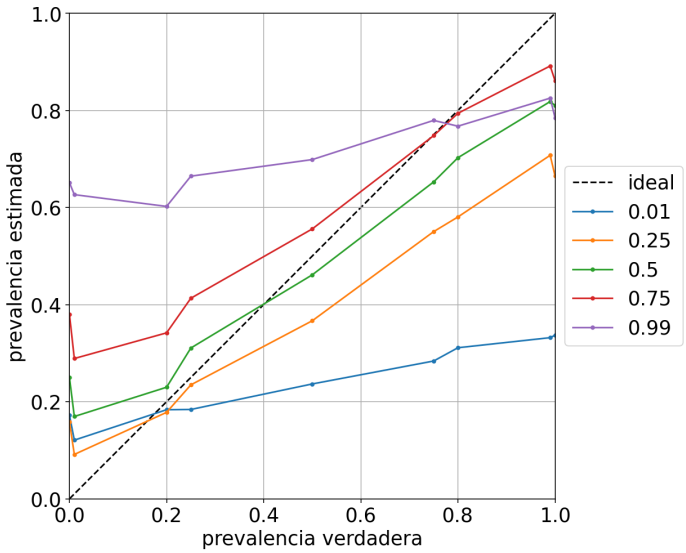


Figura 4.4: Prevalencia estimada por prevalencia de entrenamiento según la verdadera prevalencia de prueba. Cuando la muestra de prueba está muy desbalanceada, se pasa a subestimar su clase mayoritaria.

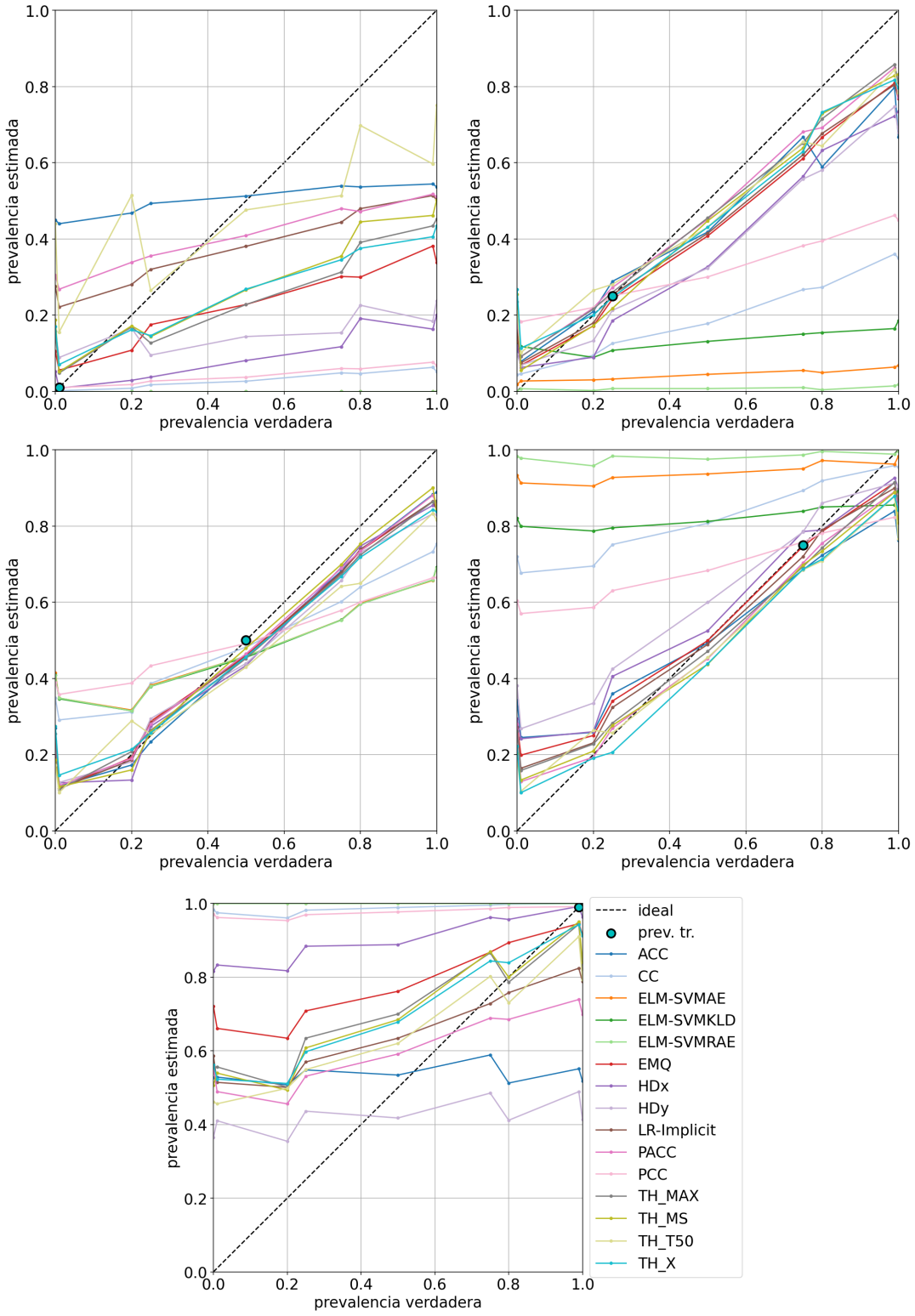


Figura 4.5: Prevalencia estimada por método de cuantificación según la verdadera prevalencia de prueba, para distintas prevalencias de entrenamiento. Se evidencia que el sesgo varía según el método de cuantificación, la prevalencia de entrenamiento y la verdadera prevalencia de prueba.

Con respecto al balance en la muestra de prueba (tabla B.6), debemos considerar los distintos valores de prevalencia para los distintos posibles tamaños de muestra. De todas formas, llama la atención que las medidas de evaluación son mejores cuando el desbalance de clases lleva a la clase

positiva a ser la minoritaria. Esto se debe a que los métodos de cuantificación binaria basados en la estimación del *fpr* y *tpr* (*ACC*, *PACC*, y *TH*) no son simétricos en cuanto a las clases positivas y negativas, sino que son sensibles a si la clase mayoritaria es una u otra.

Si tenemos en cuenta la diferencia absoluta entre las prevalencias de la muestra de entrenamiento y la de prueba (tabla B.7 y figura 4.6), vemos que también en general, a menor diferencia, mejor rendimiento. Sin embargo, llama la atención cuan dinámico es este comportamiento según el método de cuantificación (tabla B.8). En general, los métodos que mejor funcionan cuando el *dataset shift* es bajo son los que peor funcionan cuando es alto. A pesar de que en Moreo and Sebastiani [41] y Moreo et al. [43] el método *EMQ* es el de mejor rendimiento para los casos de mayor *dataset shift*, en nuestras simulaciones fue *TH_T50* el método con mejor evaluación ante estos casos.

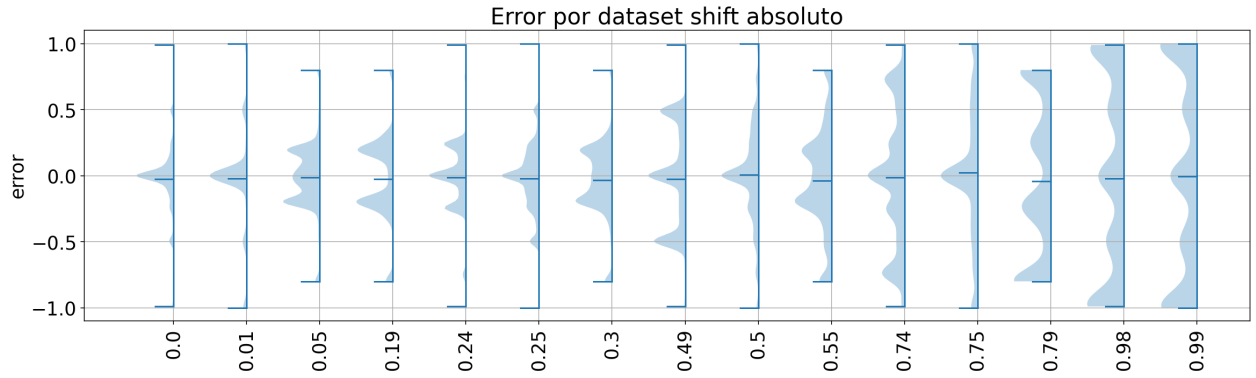


Figura 4.6: Error por *dataset shift* absoluto de calibración. A menor diferencia de prevalencias, mejor rendimiento

Analizando ahora la influencia de la calibración en la cuantificación, la calibración por regresión isotónica lleva una muy leve ventaja en el rendimiento de la cuantificación en general (considerando solo métodos de cuantificación agregativos con clasificadores generales y blandos) (tabla B.9 y figura 4.7), aunque llamativamente no presenta ventajas en cuánto a la medidas de clasificación (F1) ni de calibración (ECE).

Si tenemos en cuenta tanto el método de cuantificación como el de calibración (en los casos que aplique) (tabla B.10 y figura 4.8), se observa que *LR-Implicit*, *EMQ* y *PACC* en los casos de calibración isotónica y de no calibración se vuelven más competitivos frente a los métodos *TH*.

Sin embargo, también notamos que esta mejora en la cuantificación mediante la calibración surge si están algo desbalanceados los datos de entrenamiento, no habiendo mejoras significativas cuando están balanceados, y no siendo tan significativa si están extremadamente desbalanceados. (tabla B.11).

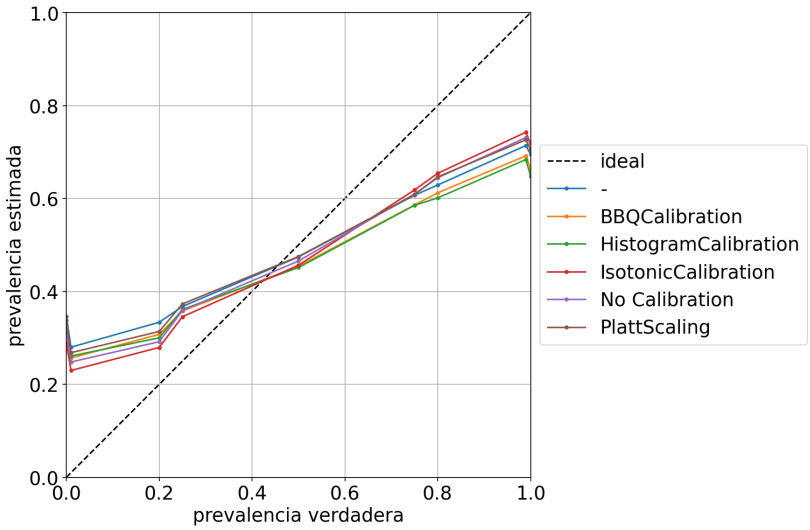


Figura 4.7: Prevalencia estimada por método de calibración según la verdadera prevalencia de prueba. Se observa que la calibración isotónica mejora ligeramente la estimación de prevalencias, siendo más notoria cuando la muestra de prueba está más desbalanceada.

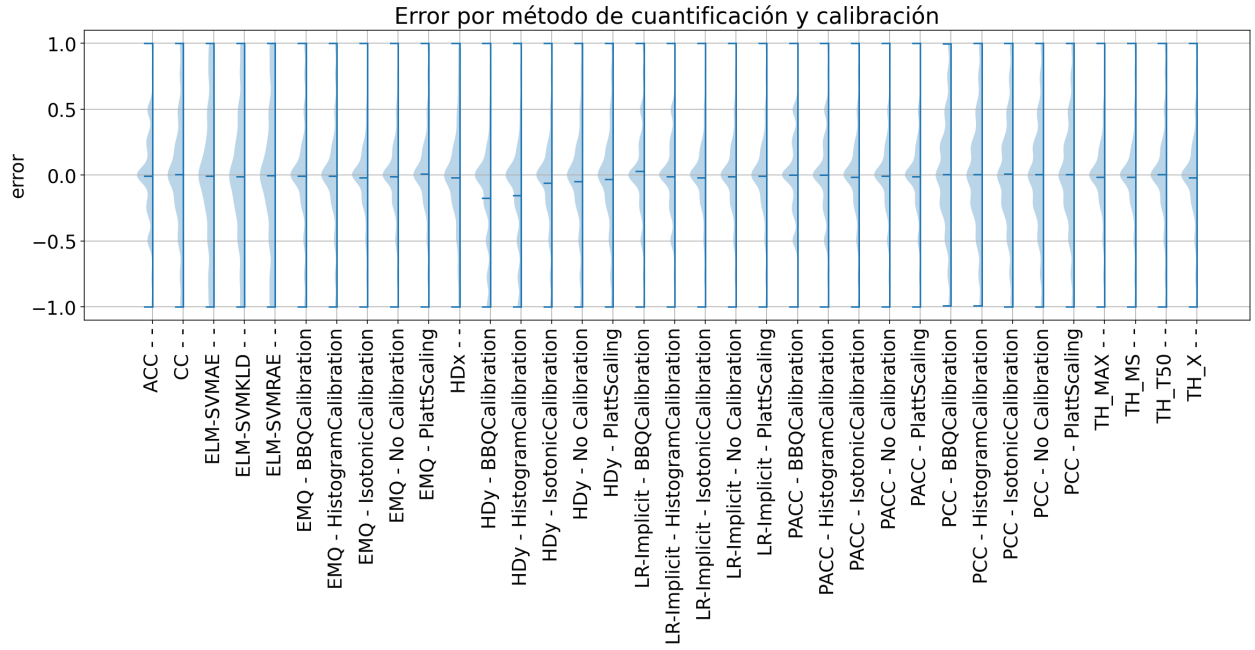


Figura 4.8: Error por método de calibración y cuantificación. Los métodos *LR-Implicit*, *EMQ* y *PACC*, especialmente con calibración isotónica o sin calibración, son en general los más competitivos.

Apéndice A

Calibración

En problemas de clasificación, el subproblema de la predicción de estimaciones de probabilidad representativas de las probabilidades verdaderas es conocido como calibración. En los sistemas del mundo real, los clasificadores no sólo deben ser precisos, sino que también deben indicar cuando es probable que sean incorrectos. Es decir, deben estimar su nivel de incertidumbre o confiabilidad. En otras palabras, las probabilidades asociadas con la etiquetas de clase predichas deben reflejar su verosimilitud real. Para ejemplificar este concepto traducimos una cita de Nate Silver en su libro *The Signal and the Noise*:

> "- Una de las pruebas más importantes del pronóstico del tiempo se llama calibración. De todas las veces que dijiste que había un 40 % de probabilidad de lluvia, ¿con qué frecuencia llovió realmente? Si, a largo plazo, realmente llovió alrededor del 40 % de las veces, eso significa que sus pronósticos estaban bien calibrados. Si terminó lloviendo solo el 20 % de las veces, o el 60 %, no fue así. "

Uno de los casos en donde se debe contemplar este problema es en la toma de decisiones (es decir, en casi toda aplicación real). Por ejemplo, en sistemas utilizados para la salud, un diagnóstico con un bajo nivel de confianza puede significar realizar otro tipo de chequeo. A su vez, las estimaciones de probabilidad, pueden ser utilizadas para ser incorporadas en otro modelo probabilístico. Por ejemplo, se podrían combinar distintas salidas de distintos modelos de forma ponderada para obtener una predicción más robusta frente los casos en donde cada modelo individual falla. Sin embargo, si únicamente es de interés la etiqueta generada por el clasificador, la calibración no aporta valor.

Si bien la calibración parece una propiedad sencilla y quizás trivial, los modelos mal calibrados son bastante comunes. Por ejemplo, algunos modelos están mal calibrados desde su origen, como Naive Bayes, Random Forest y algunas redes neuronales modernas. Incluso, hay modelos que no devuelven probabilidades *a posteriori*, sino genéricas puntuaciones de confianza, como es el caso de SVM. En estos dos últimos casos es posible mapear las salidas de los clasificadores a probabilidades *a posteriori* calibradas a través de algunos método de calibración [46, 48–50].

Dentro de los distintos modelos de clasificación, la regresión logística es un caso especial ya que está bien calibrado por diseño, dado que su función objetivo minimiza la función de pérdida logarítmica o *log-loss* [51]. Sin embargo, si no se cuenta con un conjunto de entrenamiento lo suficientemente grande, es posible que el modelo no tenga suficiente información para calibrar las probabilidades y sufrir la maldición de la dimensionalidad. Las cosas se complican aún más cuando

se incorporan técnicas como regularización o *boosting*. Dadas todas las posibles causas de una mala calibración, no se debe asumir que el modelo ajustado estará calibrado.

Por último, una buena calibración no implica que el modelo sea bueno clasificando. Por ejemplo, podemos tener un conjunto de datos balanceado (50 % clase positiva, 50 % clase negativa) con el cual ajustamos un modelo que termina clasificando aleatoriamente las observaciones en cada grupo con un 50 % de probabilidad. Este modelo tendrá solo un 50 % de *accuracy* pero estará calibrado de forma perfecta. Lo mismo puede ocurrir al contrario, por ejemplo, un modelo con muy buena capacidad de clasificación pero que siempre sobrestime las probabilidades predichas.

A.1. Definición

Podemos definir la calibración perfecta como:

$$\mathbb{P}(\hat{Y} = y|\hat{P} = p) = p, p \in [0, 1], \tag{A.1}$$

donde $\hat{Y}$ es la variable aleatoria asociada a la clase predicha y $\hat{P}$ la de la probabilidad *a posteriori* entregada por el modelo de clasificación. La probabilidad $\mathbb{P}$ no puede ser computada de forma práctica, ya que $\hat{P}$ es una variable aleatoria continua. Esto motiva la necesidad de aproximaciones empíricas.

A.2. Métodos de calibración

Se pueden clasificar en binarios o multiclase, y en no paramétricos y paramétricos. Para cualquier caso, se deben conocer las y_i clases reales del set de calibración. Luego, para los métodos no paramétricos, debemos tener acceso a las probabilidades $\hat{p}_i(x_i)$ estimadas por el modelo. Y para los métodos paramétricos, debemos tener acceso a las salidas no probabilísticas del modelo, $z_i(x_i)$, o bien calcularlas haciendo $z_i(x_i) = logit(\hat{p}_i(x_i))$ (siendo σ la función sigmoide o logística -inversa de *logit*-, y σ_{SM} la función softmax, su generalización para un vector $\boldsymbol{x}_i$).

A.2.1. Modelos binarios

Métodos no paramétricos

Histograma: Las predicciones $\hat{p}_i(x_i)$ se dividen en *bins* $B_1, \dots B_M$ mutuamente excluyentes. Cada *bin* se asigna a una probabilidad calibrada θ_m (si $\hat{p}_i(x_i)$ se asigna al *bin* B_m, entonces $\hat{q}_i = \theta_m$). Al predecir, si la predicción $\hat{p}_t$ cae en el *bin* B_m, entonces la probabilidad calibrada $\hat{q}_t$ es θ_m. Los límites de los *bins* pueden elegirse para que tengan la misma longitud o bien tengan la misma cantidad de muestras.

Regresión Isotónica: Se estima la función por partes mediante constantes $\hat{q}_i = f(\hat{p}_i)$ que minimiza $\sum_{i=1}^{n}(f(\hat{p}_i(x_i)) - y_i)^2$. Esta técnica es una generalización de la anterior, en la que los límites de los *bins* y las predicciones se optimizan simultáneamente.

Agrupamiento bayesiano en cuantiles (BBQ): Es una extensión del histograma, donde múltiples opciones de *bins* son consideradas y luego combinadas. Todas las opciones consideradas son de igual frecuencia, pero varían en la cantidad de *bins* que emplean. Luego se combinan mediante un score Bayesiano (que no analizaremos por su complejidad).

Métodos paramétricos

Platt Scaling: Las salidas no probabilísticas $z_i(x_i) = logit(\hat{p}_i(x_i))$ del clasificador son usadas como covariables para entrenar un modelo de regresión logística, el cual es entrenado con el set destinado a calibración para que retorne las probabilidades reales. Es decir, la idea es estimar los parámetros $a, b \in IR$ tal que $\hat{q}_i = \sigma(a.z_i + b)$. Estos parámetros se buscan optimizando para la *Negative Log-Likelihood*.

A.2.2. Modelos Multiclase

Métodos no paramétricos

Una forma de extender los métodos binarios no paramétricos a multiclase es tratando el problema como K problemas de uno-vs-todos. Esto dará K modelos de calibración, cada uno para una clase en particular. Al predecir, obtenemos un vector de probabilidades no normalizado $[\hat{q}_i^{(1)}, \ldots, \hat{q}_i^{(K)}]$ donde $\hat{q}_i^{(k)}$ es la probabilidad calibrada para la clase k. Luego, esto vector es normalizado dividiendo por $\sum_{k=1}^{K} \hat{q}_i^{(k)}$. Esta extensión se puede aplicar a cualquiera de los métodos no paramétricos binarios.

Métodos paramétricos

Vector scaling: Es una extensión del método de Platt Scaling, en donde se aplica una transformación lineal $W \cdot z_i + b$ a los *logits*:

$$\hat{q}_i = \max_k \sigma_{SM}(W \cdot z_i + b)^K,$$

los parámetros W y b se buscan optimizando para la *Negative Log-Likelihood* con $W \in \mathbb{R}^{K \times K}$ y $b \in \mathbb{R}^K$. Si se restringe a que W sea diagonal, estamos dentro de la variante Vector Scaling.

Temparature Scaling: Es la extensión más simple de Platt Scaling, ya que usa un sólo parámetro $T > 0$ para todas las clases:

$$\hat{q}_i = \max_k \sigma_{SM}(z_i/T)^K,$$

T es llamado temperatura, y suaviza a la función $softmax$ cuando $T > 1$. A medida que $T \to \infty$, $\hat{q}_i \to 1/K$, representando máxima incertidumbre. Con $T = 1$, se recupera la $\hat{p}_i$. A medida que $T \to 0$, $\hat{q}_i \to 1$. El valor de T se busca optimizando con respecto a la *Negative Log-Likelihood*. Como el parámetro T no altera el máximo de la función $softmax$, la predicción de las clases post-calibración $\hat{y}_i'$ se mantendrán igual a las de antes de calibrar. Es decir, que este método de calibración no cambiará el *accuracy* del modelo.

A.3. Métodos de evaluación

A.3.1. Diagramas de confiabilidad

También llamadas curvas de calibración, se construyen a partir de los valores verdaderos de las clases y las probabilidades predichas para la clase principal. Para generar una curva de calibración debemos seguir los siguientes pasos:

1. Ordenar las probabilidades predichas por el modelo para la clase principal de menor a mayor.

2. Dividir dichas probabilidades en M bines de tamaño fijo $(1/M)$.

3. Calcular la fracción de verdaderos positivos en cada bin.

4. Calcular el promedio de las probabilidades en cada bin.

5. Graficar la fracción de verdaderos positivos en el eje y el promedio de las probabilidades en el eje x.

siendo B_m el set de índices de las muestras cuyas probabilidades predichas caen en el intervalo $I_m = (\frac{m-1}{M}, \frac{m}{M}]$:

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i), \tag{A.2}$$

donde $\hat{y}_i$ y y_i son las clases predichas y reales respectivamente. Se puede demostrar que $acc(B_m)$ es un estimador insesgado y consistente de $P(\hat{Y} = Y|\hat{P} \in I_m)$. Por otro lado, tenemos:

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{A.3}$$

donde $\hat{p}_i$ es la probabilidad predicha para la muestra i.

Las medidas $acc(B_m)$ y $conf(B_m)$ aproximan el lado izquierdo y derecho de A.1 respectivamente, para el *bin* B_m. Por lo tanto, un modelo perfectamente calibrado tendrá:

$$acc(B_m) = conf(B_m), m \in 1, \ldots, M \tag{A.4}$$

Cuanto mejor calibrado esté el modelo, la curva de calibración obtenida se aproximará en mayor medida a la diagonal. La curva de calibración quedará por encima de la diagonal si el modelo tiende a subestimar las probabilidades y por debajo si las sobreestima.

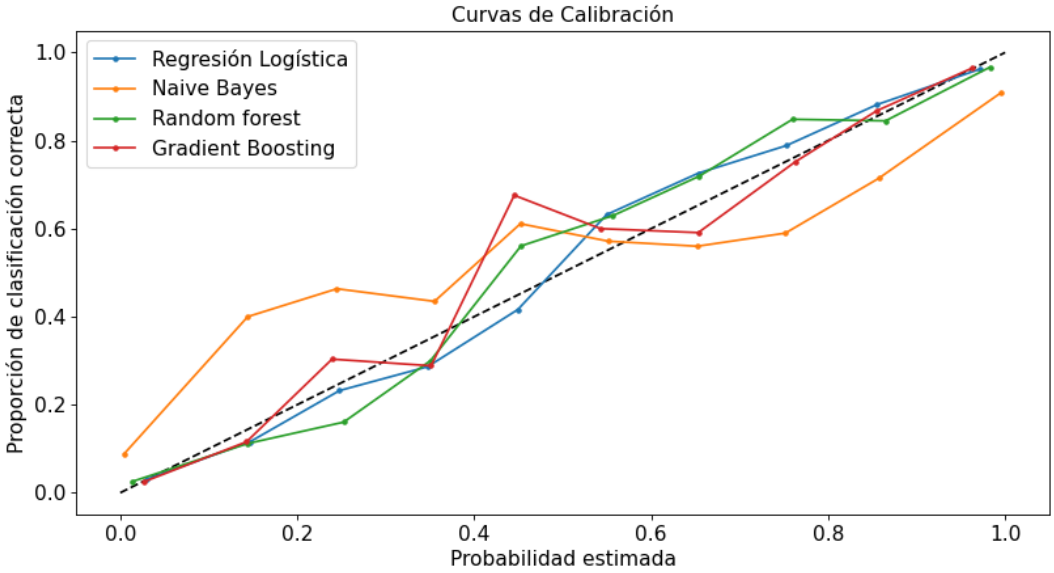


Figura A.1: Ejemplo de curvas de calibración, que muestran la relación entre las probabilidades estimadas por los modelos y la proporción real de ejemplos positivos en cada intervalo de probabilidad. Una curva perfectamente calibrada sigue la diagonal (línea punteada), lo que indica que las probabilidades predichas corresponden fielmente a la frecuencia observada. Los datos utilizados corresponden un conjunto de datos generados sintéticamente, distinto al de los experimentos 4.

A.3.2. Medidas de Evaluación

Aunque las curvas de calibración aportan información detallada, es interesante disponer de un único valor que permita cuantificar la calidad de calibración del modelo. Vamos a definir primero dos medidas de evaluación que también se basan en *bins*, y luego otras que no.

Expected Calibration Error (ECE)

Una posible medida para evaluar la descalibración podría basarse en tomar ambos términos A.1, restarlos, y calcular el valor esperado de la diferencia:

$$E_{\hat{P}}[|P(\hat{Y} = Y|\hat{P} = p) - p|],$$

que podemos estimar usando la siguiente fórmula:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|,$$

donde n es el número de muestras. El ECE es el promedio de las diferencias de los *bins* en las curvas de calibración.

Maximum Calibration Error (MCE)

Usando el mismo concepto, pero queriendo minimizar la peor desviación, podemos buscar:

$$\max_{p\in[0,1]} |P(\hat{Y} = Y|\hat{P} = p) - p|,$$

el cual también se aproxima mediante *bins*:

$$MCE = \max_{p\in[0,1]} |acc(B_m) - conf(B_m)|,$$

El MCE es la máxima diferencia de los *bins* en las curvas de calibración.

Cross-Entropy

Teniendo en cuenta que la descalibración puede deberse a un sobreajuste de la *Cross-Entropy*, podemos evaluar dicha medida en el set destinado a calibración.

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \log \hat{p}_{ij}(x_i),$$

donde N es la cantidad de muestras y C la cantidad de clases. Si este valor es relativamente menor (o mayor) al del entrenamiento, es un síntoma de una posible des-calibración. Un problema de esta medida es que da infinito si alguna $\hat{p}_{ij} = 0$.

Brier score

Otra posible medida para evaluar la bondad de ajuste de las probabilidades es el *Brier score*:

$$BS = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(\hat{p}_{ij}(x_i) - y_{ij})^2,$$

que con respecto al *Cross-Entropy* tiene la ventaja de que es siempre finito.

Apéndice B

Tablas de Resultados

	RAE			AE			SE		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Cuantificador									
TH_MS	2.10	2.00	2.21	0.22	0.22	0.23	0.12	0.12	0.12
TH_T50	2.17	2.07	2.28	0.25	0.25	0.26	0.14	0.13	0.14
TH_MAX	2.19	2.08	2.29	0.23	0.23	0.24	0.12	0.11	0.12
TH_X	2.31	2.19	2.42	0.24	0.24	0.25	0.13	0.13	0.14
LR-Implicit	2.36	2.31	2.41	0.24	0.24	0.24	0.12	0.12	0.12
EMQ	2.38	2.33	2.43	0.25	0.25	0.26	0.14	0.14	0.14
PACC	2.40	2.35	2.45	0.23	0.23	0.23	0.11	0.11	0.11
HDx	2.72	2.54	2.89	0.27	0.26	0.28	0.15	0.14	0.15
ACC	2.89	2.77	3.00	0.25	0.25	0.26	0.12	0.12	0.13
HDy	2.96	2.90	3.02	0.30	0.30	0.30	0.19	0.18	0.19
CC	3.91	3.76	4.06	0.36	0.36	0.37	0.23	0.22	0.23
PCC	3.99	3.92	4.05	0.34	0.34	0.34	0.20	0.19	0.20
ELM-SVMKLD	4.57	4.33	4.80	0.41	0.40	0.42	0.27	0.26	0.28
ELM-SVMAE	4.64	4.39	4.88	0.43	0.42	0.44	0.30	0.29	0.30
ELM-SVMRAE	4.72	4.47	4.97	0.44	0.43	0.45	0.31	0.30	0.32

Tabla B.1: Por método de cuantificación. Los métodos basados en selección de umbrales (*TH*), *PACC*, *LR-Implicit* y *EMQ* son los que en general mejor desempeño tienen, mientras que *CC*, *PCC* y los basados en minimización de pérdida explícita (*ELM*) son los peores.

	RAE			AE			SE			F1		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Clasificador												
XGBClassifier	2.31	2.28	2.34	0.23	0.23	0.23	0.12	0.12	0.12	0.52	0.52	0.52
LogisticRegression	3.24	3.21	3.27	0.31	0.31	0.31	0.18	0.18	0.18	0.38	0.38	0.39

Tabla B.2: Por modelo de clasificación (considerando solo métodos de cuantificación agregativos con clasificadores generales). A mejor desempeño en la clasificación (según su F1), mejor desempeño en la cuantificación.

	RAE			AE			SE		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Population									
F	2.32	2.30	2.35	0.24	0.24	0.24	0.12	0.12	0.12
D	3.39	3.36	3.43	0.32	0.32	0.32	0.19	0.19	0.19

Tabla B.3: Por población. A menor cantidad de características y mayor separación entre clases en las poblaciones, mejor rendimiento de cuantificación.

	RAE			AE			SE		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
n train									
5000	2.43	2.40	2.46	0.25	0.25	0.25	0.13	0.13	0.13
500	3.29	3.25	3.32	0.31	0.31	0.31	0.18	0.18	0.18

Tabla B.4: Por tamaño de muestra de entrenamiento. A mayor tamaño en muestra de entrenamiento, mejor rendimiento de cuantificación.

	RAE			AE			SE		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Train prev +									
0.50	1.86	1.83	1.90	0.18	0.18	0.18	0.07	0.07	0.07
0.25	2.10	2.07	2.14	0.22	0.22	0.22	0.11	0.10	0.11
0.75	2.16	2.12	2.20	0.22	0.22	0.22	0.10	0.10	0.10
0.01	4.04	3.98	4.10	0.38	0.38	0.38	0.25	0.24	0.25
0.99	4.12	4.06	4.18	0.39	0.39	0.39	0.25	0.25	0.26

Tabla B.5: Por prevalencia de clase positiva en muestra de entrenamiento. Cuanto más balanceada sea la muestra de entrenamiento, mejores resultados de cuantificación se obtienen. Este es un dato bastante interesante, ya que implicaría que podríamos aplicar en cuantificación las mismas técnicas de balance de clases que se usan en problemas de clasificación para tratar de conseguir mejores resultados que si usáramos datos desbalanceados.

	RAE			AE			SE		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Test prev +									
0.50	0.49	0.49	0.49	0.26	0.26	0.26	0.10	0.10	0.10
0.25	0.62	0.62	0.63	0.24	0.24	0.24	0.10	0.10	0.11
0.75	0.70	0.69	0.71	0.27	0.26	0.27	0.13	0.12	0.13
0.20	0.72	0.71	0.72	0.28	0.27	0.28	0.13	0.13	0.13
0.80	0.81	0.80	0.81	0.31	0.31	0.31	0.16	0.16	0.17
1.00	3.23	3.19	3.27	0.31	0.30	0.31	0.23	0.23	0.24
0.00	3.38	3.34	3.42	0.32	0.32	0.33	0.23	0.23	0.24
0.01	8.63	8.51	8.75	0.25	0.25	0.26	0.17	0.17	0.17
0.99	9.52	9.39	9.64	0.28	0.28	0.29	0.20	0.20	0.20

Tabla B.6: Por prevalencia de clase positiva en muestra de prueba. Llama la atención que las medidas de evaluación son mejores cuando el desbalance de clases lleva a la clase positiva a ser la minoritaria. Esto se debe a que los métodos de cuantificación binaria basados en la estimación del *fpr* y *tpr* (ACC, $PACC$, y TH) no son simétricos en cuanto a las clases positivas y negativas, sino que son sensibles a si la clase mayoritaria es una u otra.

	RAE			AE			SE		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Abs Dataset Shift									
[0]	1.85	1.79	1.90	0.14	0.14	0.14	0.06	0.06	0.06
(0-0.25]	1.10	1.09	1.11	0.19	0.19	0.19	0.08	0.08	0.08
(0.25-0.5]	1.77	1.75	1.80	0.26	0.26	0.26	0.12	0.12	0.12
(0.5-0.75]	3.88	3.82	3.93	0.36	0.36	0.36	0.22	0.22	0.22
(0.75-1.0]	9.96	9.84	10.07	0.60	0.60	0.61	0.48	0.48	0.49

Tabla B.7: Por *dataset shift* absoluto. A menor diferencia, mejor rendimiento.

		RAE			AE			SE		
Abs Dataset Shift	Cuantificador	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
[0]	**PCC**	0.07	0.07	0.07	0.02	0.02	0.02	0.00	0.00	0.00
	CC	0.27	0.26	0.28	0.07	0.07	0.08	0.01	0.01	0.01
	ELM-SVMKLD	0.31	0.29	0.32	0.09	0.08	0.09	0.02	0.01	0.02
	HDx	0.29	0.27	0.32	0.09	0.08	0.10	0.02	0.02	0.02
	ELM-SVMAE	0.35	0.33	0.36	0.10	0.09	0.11	0.02	0.02	0.02
	EMQ	0.79	0.70	0.87	0.11	0.11	0.11	0.04	0.03	0.04
	ELM-SVMRAE	0.37	0.36	0.39	0.11	0.10	0.12	0.02	0.02	0.03
	TH_MS	0.79	0.68	0.89	0.12	0.11	0.12	0.03	0.03	0.04
	TH_MAX	0.83	0.69	0.98	0.12	0.12	0.13	0.04	0.03	0.04
	TH_X	0.99	0.81	1.17	0.14	0.13	0.15	0.05	0.05	0.06
	LR-Implicit	2.45	2.29	2.61	0.17	0.16	0.17	0.07	0.07	0.07
	TH_T50	1.67	1.37	1.98	0.17	0.16	0.18	0.07	0.06	0.08
	PACC	3.17	3.01	3.33	0.19	0.19	0.20	0.08	0.07	0.08
	HDy	3.63	3.39	3.86	0.21	0.20	0.22	0.12	0.12	0.13
	ACC	5.20	4.76	5.65	0.27	0.26	0.29	0.12	0.11	0.13
(0-0.25]	**HDx**	0.72	0.66	0.78	0.15	0.14	0.15	0.04	0.03	0.04
	PCC	1.13	1.11	1.16	0.15	0.15	0.16	0.03	0.03	0.03
	CC	0.53	0.51	0.55	0.16	0.15	0.16	0.04	0.04	0.05
	EMQ	0.86	0.83	0.90	0.16	0.16	0.17	0.06	0.06	0.07
	TH_MS	1.05	0.97	1.13	0.17	0.17	0.18	0.07	0.06	0.07
	TH_MAX	1.03	0.96	1.11	0.17	0.17	0.18	0.07	0.06	0.07
	TH_X	1.16	1.05	1.26	0.18	0.18	0.19	0.08	0.07	0.08
	ELM-SVMKLD	0.99	0.90	1.09	0.19	0.18	0.20	0.06	0.06	0.06
	ELM-SVMAE	0.55	0.53	0.58	0.19	0.18	0.20	0.06	0.06	0.07
	LR-Implicit	1.11	1.07	1.16	0.19	0.19	0.20	0.08	0.08	0.09
	ELM-SVMRAE	0.49	0.47	0.51	0.20	0.19	0.21	0.07	0.07	0.08
	PACC	1.26	1.21	1.30	0.20	0.20	0.21	0.08	0.08	0.09
	HDy	1.26	1.21	1.30	0.24	0.24	0.24	0.13	0.13	0.13
	TH_T50	1.45	1.34	1.56	0.24	0.23	0.25	0.13	0.12	0.14
	ACC	1.58	1.47	1.69	0.25	0.24	0.25	0.11	0.11	0.12
(0.25-0.5]	**ACC**	1.24	1.08	1.39	0.15	0.14	0.16	0.07	0.06	0.08
	PACC	1.28	1.22	1.35	0.18	0.17	0.18	0.08	0.08	0.08
	LR-Implicit	1.34	1.27	1.40	0.21	0.20	0.21	0.09	0.09	0.10
	TH_MS	1.26	1.12	1.39	0.21	0.20	0.22	0.09	0.09	0.10
	TH_MAX	1.38	1.25	1.51	0.22	0.21	0.23	0.09	0.09	0.10
	TH_X	1.62	1.45	1.80	0.24	0.23	0.25	0.11	0.11	0.12
	TH_T50	1.48	1.32	1.64	0.24	0.23	0.25	0.12	0.11	0.13
	EMQ	1.39	1.33	1.45	0.25	0.24	0.25	0.11	0.11	0.11
	HDx	1.46	1.30	1.62	0.25	0.24	0.26	0.10	0.09	0.11
	HDy	1.60	1.53	1.67	0.28	0.27	0.28	0.13	0.13	0.14
	CC	2.62	2.48	2.76	0.36	0.35	0.37	0.16	0.16	0.17
	PCC	3.08	3.00	3.15	0.37	0.37	0.37	0.15	0.15	0.15
	ELM-SVMKLD	3.15	2.91	3.38	0.41	0.40	0.42	0.20	0.19	0.20
	ELM-SVMAE	3.22	2.98	3.45	0.43	0.42	0.44	0.22	0.21	0.23
	ELM-SVMRAE	3.23	2.99	3.46	0.43	0.42	0.44	0.23	0.22	0.24
(0.5-0.75]	**PACC**	2.21	2.11	2.31	0.24	0.23	0.25	0.12	0.12	0.13
	TH_T50	2.07	1.84	2.29	0.25	0.23	0.26	0.14	0.13	0.15
	TH_MS	2.20	1.94	2.46	0.25	0.23	0.26	0.15	0.14	0.16
	TH_X	2.18	1.95	2.42	0.25	0.24	0.27	0.14	0.13	0.15
	TH_MAX	2.26	2.05	2.48	0.25	0.24	0.27	0.14	0.13	0.15
	LR-Implicit	2.55	2.45	2.66	0.26	0.26	0.27	0.14	0.13	0.14
	ACC	3.14	2.83	3.44	0.28	0.27	0.30	0.16	0.15	0.17
	EMQ	2.77	2.66	2.88	0.30	0.30	0.31	0.17	0.17	0.18
	HDy	3.61	3.47	3.74	0.36	0.35	0.36	0.22	0.22	0.23
	HDx	3.46	3.09	3.84	0.36	0.34	0.38	0.20	0.19	0.22
	PCC	6.83	6.67	6.99	0.56	0.55	0.56	0.34	0.33	0.34
	CC	8.08	7.62	8.54	0.64	0.63	0.65	0.46	0.45	0.48
	ELM-SVMKLD	9.86	9.10	10.62	0.75	0.73	0.76	0.60	0.58	0.61
	ELM-SVMAE	11.14	10.32	11.97	0.83	0.82	0.84	0.71	0.69	0.72
	ELM-SVMRAE	11.79	10.91	12.67	0.87	0.86	0.88	0.77	0.76	0.78
(0.75-1.0]	**TH_T50**	6.36	5.78	6.94	0.38	0.36	0.40	0.28	0.26	0.30
	ACC	7.46	7.04	7.89	0.43	0.42	0.45	0.23	0.22	0.24
	PACC	7.54	7.32	7.76	0.46	0.45	0.46	0.29	0.28	0.30
	LR-Implicit	7.80	7.56	8.04	0.48	0.47	0.49	0.33	0.32	0.34
	TH_MS	8.12	7.50	8.75	0.49	0.47	0.51	0.38	0.36	0.40
	TH_MAX	8.52	7.94	9.10	0.51	0.49	0.53	0.37	0.35	0.39
	TH_X	8.59	7.98	9.21	0.52	0.50	0.54	0.40	0.38	0.42
	HDy	9.26	8.96	9.56	0.56	0.55	0.57	0.47	0.46	0.48
	EMQ	10.00	9.72	10.29	0.62	0.61	0.63	0.51	0.50	0.52
	HDx	12.65	11.70	13.61	0.75	0.73	0.77	0.62	0.59	0.64
	PCC	14.49	14.17	14.81	0.88	0.87	0.88	0.79	0.78	0.79
	CC	14.69	13.96	15.42	0.89	0.88	0.90	0.81	0.80	0.82
	ELM-SVMAE	15.35	14.28	16.41	0.93	0.92	0.94	0.87	0.86	0.89
	ELM-SVMKLD	15.35	14.28	16.41	0.93	0.92	0.94	0.87	0.86	0.89
	ELM-SVMRAE	15.35	14.28	16.41	0.93	0.92	0.94	0.87	0.86	0.89

Tabla B.8: Por *dataset shift* absoluto, y método de cuantificación. En general, los métodos que mejor funcionan cuando el *dataset shift* es bajo son los que peor funcionan cuando es alto.

	RAE			AE			SE			F1			ECE		
	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Calibración															
IsotonicCalibration	2.57	2.52	2.62	0.25	0.25	0.26	0.14	0.13	0.14	0.45	0.45	0.45	32.79	32.55	33.03
No Calibration	2.73	2.68	2.78	0.27	0.27	0.28	0.15	0.15	0.15	0.46	0.45	0.46	32.55	32.31	32.80
PlattScaling	2.85	2.80	2.90	0.28	0.28	0.28	0.16	0.15	0.16	0.45	0.45	0.45	31.71	31.47	31.96
BBQCalibration	2.95	2.89	3.01	0.28	0.28	0.28	0.16	0.16	0.16	0.45	0.45	0.45	33.20	32.94	33.45
HistogramCalibration	2.99	2.93	3.04	0.28	0.28	0.28	0.16	0.15	0.16	0.45	0.45	0.45	33.85	33.61	34.10

Tabla B.9: Por método de calibración (considerando solo métodos de cuantificación agregativos con clasificadores generales y blandos). La calibración por regresión isotónica lleva una muy leve ventaja en el rendimiento de la cuantificación en general, aunque llamativamente no presenta ventajas en cuánto a la medidas de clasificación (F1) ni de calibración (ECE).

		RAE			AE			SE			F1			ECE		
		media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Cuantificador	Calibración															
LR-Implicit	**IsotonicCalibration**	2.10	2.00	2.20	0.22	0.21	0.22	0.11	0.10	0.11	0.45	0.44	0.46	32.71	32.17	33.25
TH_MS	**-**	2.10	2.00	2.21	0.22	0.22	0.23	0.12	0.12	0.12	0.45	0.44	0.46	NaN	NaN	NaN
EMQ	**IsotonicCalibration**	2.10	2.01	2.20	0.22	0.21	0.22	0.11	0.10	0.11	0.45	0.44	0.46	32.29	31.75	32.82
	No Calibration	2.13	2.03	2.23	0.24	0.23	0.24	0.12	0.12	0.13	0.46	0.45	0.47	32.38	31.83	32.92
TH_T50	**-**	2.17	2.07	2.28	0.25	0.25	0.26	0.14	0.13	0.14	0.45	0.44	0.46	NaN	NaN	NaN
PACC	**IsotonicCalibration**	2.18	2.08	2.29	0.23	0.23	0.24	0.12	0.11	0.12	0.45	0.44	0.45	33.22	32.69	33.76
TH_MAX	**-**	2.19	2.08	2.29	0.23	0.23	0.24	0.12	0.11	0.12	0.45	0.44	0.46	NaN	NaN	NaN
PACC	**PlattScaling**	2.19	2.09	2.29	0.23	0.22	0.23	0.12	0.11	0.12	0.45	0.44	0.45	31.80	31.26	32.34
LR-Implicit	**No Calibration**	2.19	2.09	2.29	0.24	0.24	0.25	0.13	0.12	0.13	0.46	0.45	0.47	32.38	31.83	32.92
PACC	**No Calibration**	2.23	2.12	2.33	0.23	0.22	0.23	0.12	0.11	0.12	0.45	0.45	0.46	32.75	32.21	33.30
TH_X	**-**	2.31	2.19	2.42	0.24	0.24	0.25	0.13	0.13	0.14	0.45	0.44	0.46	NaN	NaN	NaN
LR-Implicit	**PlattScaling**	2.42	2.31	2.54	0.25	0.25	0.26	0.14	0.13	0.14	0.45	0.45	0.46	31.69	31.15	32.24
	BBQCalibration	2.44	2.34	2.55	0.24	0.24	0.25	0.13	0.12	0.13	0.45	0.44	0.46	33.32	32.75	33.89
EMQ	**PlattScaling**	2.46	2.35	2.58	0.26	0.25	0.26	0.14	0.14	0.15	0.45	0.45	0.46	31.56	31.01	32.11
HDy	**IsotonicCalibration**	2.55	2.43	2.67	0.27	0.26	0.28	0.16	0.15	0.16	0.44	0.44	0.45	33.40	32.86	33.94
EMQ	**BBQCalibration**	2.57	2.45	2.70	0.28	0.27	0.29	0.17	0.17	0.18	0.45	0.44	0.46	33.28	32.72	33.85
	HistogramCalibration	2.64	2.52	2.76	0.28	0.27	0.28	0.17	0.16	0.17	0.45	0.44	0.46	33.59	33.04	34.14
LR-Implicit	**HistogramCalibration**	2.65	2.54	2.75	0.23	0.23	0.24	0.11	0.11	0.11	0.45	0.44	0.46	33.68	33.13	34.23
PACC	**HistogramCalibration**	2.65	2.55	2.76	0.24	0.23	0.24	0.11	0.11	0.12	0.45	0.44	0.46	34.15	33.60	34.69
HDx	**-**	2.72	2.54	2.89	0.27	0.26	0.28	0.15	0.14	0.15	NaN	NaN	NaN	NaN	NaN	NaN
PACC	**BBQCalibration**	2.74	2.64	2.84	0.23	0.22	0.23	0.10	0.09	0.10	0.45	0.44	0.45	33.33	32.76	33.90
HDy	**BBQCalibration**	2.88	2.74	3.02	0.30	0.30	0.31	0.20	0.19	0.20	0.45	0.44	0.45	33.09	32.52	33.66
ACC	**-**	2.89	2.77	3.00	0.25	0.25	0.26	0.12	0.12	0.13	0.45	0.44	0.46	NaN	NaN	NaN
HDy	**HistogramCalibration**	2.95	2.81	3.09	0.30	0.30	0.31	0.19	0.19	0.20	0.45	0.44	0.45	34.22	33.67	34.77
	No Calibration	3.14	3.01	3.27	0.31	0.31	0.32	0.19	0.18	0.19	0.45	0.44	0.46	32.88	32.34	33.43
	PlattScaling	3.27	3.13	3.40	0.32	0.31	0.33	0.19	0.19	0.20	0.45	0.44	0.46	31.96	31.42	32.51
PCC	**PlattScaling**	3.91	3.76	4.05	0.34	0.33	0.34	0.19	0.19	0.20	0.45	0.44	0.46	31.56	31.01	32.11
CC	**-**	3.91	3.76	4.06	0.36	0.36	0.37	0.23	0.22	0.23	0.46	0.45	0.47	NaN	NaN	NaN
PCC	**IsotonicCalibration**	3.91	3.77	4.05	0.33	0.33	0.34	0.19	0.18	0.19	0.45	0.44	0.46	32.31	31.78	32.85
	No Calibration	3.96	3.81	4.10	0.34	0.33	0.35	0.20	0.19	0.20	0.46	0.45	0.47	32.38	31.83	32.92
	HistogramCalibration	4.05	3.90	4.19	0.34	0.34	0.35	0.20	0.19	0.20	0.45	0.44	0.46	33.63	33.08	34.17
	BBQCalibration	4.11	3.97	4.26	0.35	0.34	0.35	0.20	0.20	0.21	0.45	0.44	0.46	32.96	32.40	33.53
ELM-SVMKLD	**-**	4.57	4.33	4.80	0.41	0.40	0.42	0.27	0.26	0.28	0.39	0.37	0.40	NaN	NaN	NaN
ELM-SVMAE	**-**	4.64	4.39	4.88	0.43	0.42	0.44	0.30	0.29	0.30	0.37	0.36	0.38	NaN	NaN	NaN
ELM-SVMRAE	**-**	4.72	4.47	4.97	0.44	0.43	0.45	0.31	0.30	0.32	0.36	0.35	0.37	NaN	NaN	NaN

Tabla B.10: Por método de cuantificación y de calibración (considerando solo métodos de cuantificación agregativos con clasificadores generales y blandos). *LR-Implicit*, *EMQ* y *PACC*, en los casos de calibración isotónica y de no calibración, se vuelven más competitivos frente a los métodos *TH* (que como vimos en B.1, ganan en la tabla general).

		RAE			AE			SE			F1			ECE		
		media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.	media	lim. inf.	lim. sup.
Train prev +	Calibración															
0.01	**IsotonicCalibration**	3.57	3.43	3.72	0.35	0.35	0.36	0.22	0.21	0.22	0.07	0.07	0.08	45.80	45.14	46.46
	No Calibration	3.90	3.75	4.06	0.39	0.38	0.40	0.25	0.25	0.26	0.08	0.08	0.09	46.42	45.75	47.09
	PlattScaling	4.02	3.86	4.17	0.39	0.38	0.39	0.25	0.25	0.26	0.08	0.07	0.08	45.71	45.05	46.37
	BBQCalibration	4.38	4.22	4.54	0.39	0.38	0.40	0.25	0.24	0.26	0.08	0.07	0.08	47.17	46.49	47.85
	HistogramCalibration	4.40	4.24	4.56	0.39	0.39	0.40	0.25	0.24	0.26	0.07	0.07	0.08	46.65	45.98	47.32
0.25	**IsotonicCalibration**	1.83	1.75	1.91	0.19	0.19	0.20	0.08	0.08	0.09	0.35	0.35	0.36	25.94	25.59	26.29
	BBQCalibration	1.91	1.82	2.00	0.21	0.20	0.21	0.10	0.09	0.10	0.35	0.34	0.36	26.92	26.54	27.30
	HistogramCalibration	2.01	1.92	2.10	0.20	0.20	0.21	0.09	0.09	0.10	0.35	0.35	0.36	27.19	26.83	27.55
	No Calibration	2.07	1.98	2.16	0.21	0.21	0.22	0.10	0.09	0.10	0.37	0.36	0.38	26.28	25.93	26.64
	PlattScaling	2.11	2.02	2.20	0.22	0.21	0.22	0.10	0.10	0.10	0.35	0.35	0.36	25.09	24.73	25.44
0.50	**No Calibration**	1.79	1.71	1.87	0.17	0.17	0.17	0.07	0.06	0.07	0.58	0.57	0.59	15.39	15.22	15.56
	IsotonicCalibration	1.79	1.71	1.87	0.18	0.18	0.18	0.07	0.07	0.08	0.58	0.57	0.58	18.64	18.42	18.86
	BBQCalibration	1.82	1.74	1.90	0.18	0.18	0.19	0.08	0.07	0.08	0.58	0.57	0.58	15.04	14.83	15.25
	PlattScaling	1.87	1.79	1.95	0.18	0.17	0.18	0.07	0.07	0.08	0.58	0.57	0.58	14.55	14.37	14.73
	HistogramCalibration	1.87	1.79	1.95	0.18	0.18	0.19	0.07	0.07	0.08	0.58	0.57	0.58	19.38	19.16	19.60
0.75	**IsotonicCalibration**	1.90	1.82	1.99	0.19	0.18	0.19	0.08	0.08	0.08	0.61	0.61	0.62	26.07	25.71	26.43
	HistogramCalibration	2.06	1.97	2.14	0.20	0.20	0.21	0.09	0.09	0.09	0.61	0.61	0.62	27.88	27.51	28.26
	BBQCalibration	2.08	1.99	2.17	0.21	0.20	0.21	0.10	0.09	0.10	0.61	0.61	0.62	28.03	27.63	28.43
	No Calibration	2.08	1.99	2.17	0.21	0.20	0.21	0.09	0.09	0.09	0.62	0.61	0.62	26.81	26.45	27.17
	PlattScaling	2.15	2.06	2.24	0.21	0.21	0.22	0.10	0.09	0.10	0.61	0.61	0.62	25.71	25.35	26.07
0.99	**IsotonicCalibration**	3.75	3.60	3.90	0.36	0.35	0.37	0.23	0.22	0.23	0.59	0.58	0.59	47.48	46.81	48.16
	No Calibration	3.81	3.66	3.96	0.38	0.37	0.39	0.24	0.24	0.25	0.59	0.58	0.59	47.86	47.18	48.54
	PlattScaling	4.10	3.94	4.26	0.40	0.39	0.40	0.26	0.25	0.27	0.59	0.58	0.59	47.51	46.84	48.18
	BBQCalibration	4.57	4.40	4.74	0.42	0.41	0.42	0.28	0.27	0.29	0.58	0.58	0.59	48.82	48.13	49.51
	HistogramCalibration	4.60	4.43	4.76	0.41	0.41	0.42	0.27	0.27	0.28	0.59	0.58	0.59	48.16	47.48	48.84

Tabla B.11: Por prevalencia de clase positiva en muestra de entrenamiento y de calibración (considerando solo métodos de cuantificación agregativos con clasificadores generales y blandos). La mejora en la cuantificación mediante la calibración surge si están algo desbalanceados los datos de entrenamiento, no habiendo mejoras significativas cuando están balanceados, y no siendo tan significativa si están extremadamente desbalanceados.

Referencias

[1] George Forman. Counting positives accurately despite inaccurate classification. In *European conference on machine learning*, pages 564–575. Springer, 2005.

[2] David D Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 246–254, 1995.

[3] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[4] Amos Storkey et al. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30(3-28):6, 2009.

[5] Rocío Alaíz-Rodríguez, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro. Class and subclass probability re-estimation to adapt a classifier in the presence of concept drift. *Neurocomputing*, 74(16):2614–2623, 2011.

[6] Marthinus Christoffel Du Plessis and Masashi Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE TRANSACTIONS on Information and Systems*, 97(5):1358–1362, 2014.

[7] Yee Seng Chan and Hwee Tou Ng. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 89–96, 2006.

[8] Zhihao Zhang and Jie Zhou. Transfer estimation of evolving class priors in data stream classification. *Pattern Recognition*, 43(9):3151–3161, 2010.

[9] Marthinus Christoffel Du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

[10] Jose Barranquero, Pablo González, Jorge Díez, and Juan José Del Coz. On the study of nearest neighbor algorithms for prevalence estimation in binary problems. *Pattern Recognition*, 46(2): 472–482, 2013.

[11] Hideki Asoh, Kazushi Ikeda, and Chihiro Ono. A fast and simple method for profiling a population of twitter users. In *The Third International Workshop on Mining Ubiquitous and Social Environments*, page 19. Citeseer, 2012.

[12] Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. Class distribution estimation based on the Hellinger distance. *Information Sciences*, 218:146–164, 2013.

[13] Nachai Limsetto and Kitsana Waiyamai. Handling concept drift via ensemble and class distribution estimation technique. In *Advanced Data Mining and Applications: 7th International Conference, ADMA 2011, Beijing, China, December 17-19, 2011, Proceedings, Part II 7*, pages 13–26. Springer, 2011.

[14] Jack Chongjie Xue and Gary M Weiss. Quantification and semi-supervised classification methods for handling changes in class distribution. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 897–906, 2009.

[15] Afonso Fernandes Vaz, Rafael Izbicki, and Rafael Bassi Stern. Quantification under prior probability shift: The ratio estimator and its extensions. *The Journal of Machine Learning Research*, 20(1):2921–2953, 2019.

[16] Andrea Esuli, Alessandro Fabris, Alejandro Moreo, and Fabrizio Sebastiani. *Learning to Quantify*. Springer Nature, 2023.

[17] Pablo González, Alberto Castaño, Nitesh V Chawla, and Juan José Del Coz. A review on quantification learning. *ACM Computing Surveys (CSUR)*, 50(5):1–40, 2017.

[18] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17:164–206, 2008.

[19] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[20] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

[21] Arpita Biswas and Suvam Mukherjee. Ensuring fairness under prior probability shifts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 414–424, 2021.

[22] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. Measuring fairness under unawareness of sensitive attributes: A quantification-based approach. *Journal of Artificial Intelligence Research*, 76:1117–1180, 2023.

[23] Isabelle Guyon et al. Design of experiments for the NIPS 2003 variable selection benchmark. In *NIPS 2003 workshop on feature extraction and feature selection*, volume 253, page 40, 2003.

[24] Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. Quantification via probability estimators. In *2010 IEEE International Conference on Data Mining*, pages 737–742. IEEE, 2010.

[25] Lei Tang, Huiji Gao, and Huan Liu. Network quantification despite biased labels. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 147–154, 2010.

[26] Dirk Tasche. Exact fit of simple finite mixture models. *Journal of Risk and Financial Management*, 7(4):150–164, 2014.

[27] George Forman. Quantifying trends accurately despite classifier error and class imbalance. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 157–166, 2006.

[28] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[29] Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–34, 2020.

[30] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020.

[31] Andrea Esuli, Fabrizio Sebastiani, and Ahmed Abbasi. Sentiment Quantification. *IEEE Intell. Syst.*, 25(4):72–75, 2010.

[32] Andrea Esuli and Fabrizio Sebastiani. Explicit Loss Minimization in Quantification Applications (Preliminary Draft). In *DART@ AI* IA*, pages 1–11. Citeseer, 2014.

[33] Andrea Esuli and Fabrizio Sebastiani. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(4):1–27, 2015.

[34] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384, 2005.

[35] Jose Barranquero, Jorge Díez, and Juan José del Coz. Quantification-oriented learning based on reliable classifiers. *Pattern Recognition*, 48(2):591–604, 2015.

[36] Alejandro Moreo and Fabrizio Sebastiani. Re-assessing the "classify and count" quantification method. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 75–91. Springer, 2021.

[37] Purushottam Kar, Shuai Li, Harikrishna Narasimhan, Sanjay Chawla, and Fabrizio Sebastiani. Online optimization methods for the quantification problem. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1625–1634, 2016.

[38] Amartya Sanyal, Pawan Kumar, Purushottam Kar, Sanjay Chawla, and Fabrizio Sebastiani. Optimizing non-decomposable measures with deep networks. *Machine Learning*, 107:1597–1620, 2018.

[39] Katherine Keith and Brendan O'Connor. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4575–4585, 2018.

[40] Fabrizio Sebastiani. Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal*, 23(3):255–288, 2020.

[41] Alejandro Moreo and Fabrizio Sebastiani. Tweet sentiment quantification: An experimental re-evaluation. *PLoS One*, 17(9):e0263449, 2022.

[42] Waqar Hassan, André Gustavo Maletzke, and Gustavo Enrique de Almeida Prado Alves Batista. Pitfalls in quantification assessment. In *Proceedings*, 2021.

[43] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. QuaPy: A Python-based framework for quantification. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4534–4543, 2021.

[44] Dirk Tasche. Does quantification without adjustments work? *arXiv preprint arXiv:1602.08780*, 2016.

[45] Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. A comparative evaluation of quantification methods. *arXiv preprint arXiv:2103.03223*, 2021.

[46] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[47] Fabian Kuppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 326–327, 2020.

[48] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[49] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

[50] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[51] Geoffrey Stewart Morrison. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197, 2013.