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

Alumno: Ing. Maximiliano Marufo da Silva, **Director:** Dr. Andrés Farall

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

Introducción

La tarea de cuantificación consiste en proporcionar predicciones agregadas para conjuntos de datos, en vez de predicciones particulares sobre los datos individuales. Por ejemplo, para el caso de la cuantificación aplicada a la clasificación, se busca predecir la proporción de clases de un conjunto de individuos, en vez de la clase particular de cada individuo. En este caso, se puede utilizar un clasificador para predecir por cada individuo si la clase es positiva (o negativa), y luego estimar la proporción de clases positivas contándolas y dividiéndolos por el total.

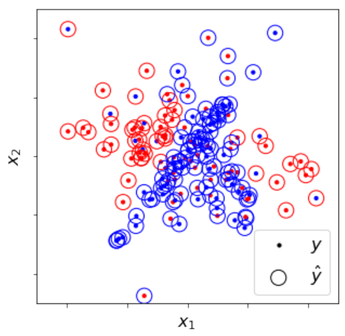

Figura 1: Clasificación

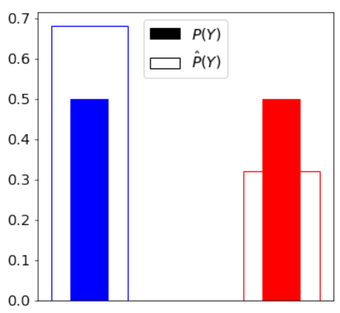

Figura 2: Cuantificación

Sin embargo, esta estrategia (método *baseline* de cuantificación, conocido como *Classify & Count* -*CC*-) es subóptima: si bien un clasificador perfecto es también un cuantificador perfecto, **un buen clasificador puede ser un mal cuantificador**, principalmente si la distribución de los datos utilizados para entrenar es diferente a la de los datos que se usan para predecir.

Marco teórico

Teniendo en cuenta que en los problemas de clasificación tenemos:

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

La probabilidad conjunta $\mathbb{P}(Y, \mathbf{X})$ se puede escribir como $\mathbb{P}(Y|\mathbf{X})\mathbb{P}(\mathbf{X})$ o como $\mathbb{P}(\mathbf{X}|Y)\mathbb{P}(Y)$. El *dataset shift* aparece cuando las distribuciones conjuntas de entrenamiento y de prueba son diferentes, es decir, cuando $\mathbb{P}_{tr}(Y, \mathbf{X}) \neq \mathbb{P}_{tst}(Y, \mathbf{X})$. En el caso particular de $\mathbb{P}_{tr}(\mathbf{X}|Y) = \mathbb{P}_{tst}(\mathbf{X}|Y)$ y $\mathbb{P}_{tr}(Y) \neq \mathbb{P}_{tst}(Y)$, se habla de *prior probability shift*. El problema de cuantificación se trata de un caso donde $\mathbb{P}_{tst}(Y)$ es desconocido. Además, la mayoría de los métodos de cuantificación propuestos asumen que $\mathbb{P}_{tr}(\mathbf{X}|Y) = \mathbb{P}_{tst}(\mathbf{X}|Y)$, por lo que están dentro de los casos de *prior probability shift*.

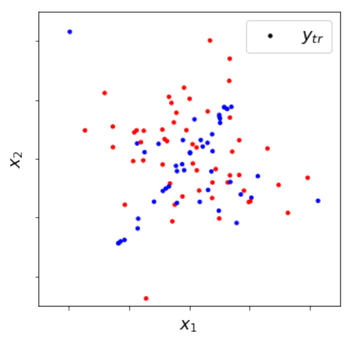

Figura 3: Muestra de entrenamiento

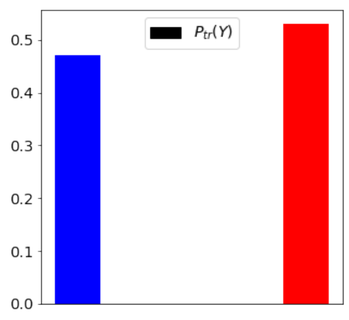

Figura 4: Prevalencia de clases en muestra de entrenamiento

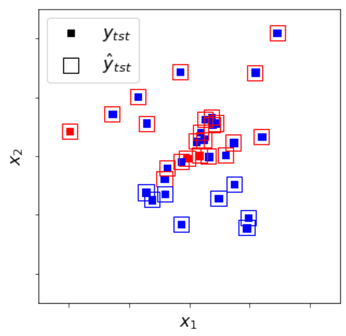

Figura 5: Clasificación en muestra de prueba. Para el modelo, las y_{tst} son desconocidas.

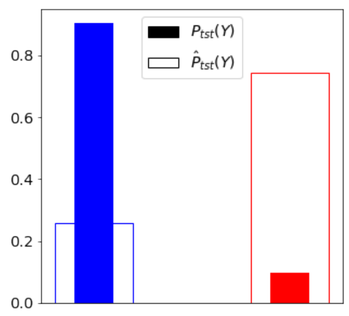

Figura 6: Prevalencia de clases verdadera y cuantificación en muestra de prueba

El problema de clasificar y contar

Bajo el supuesto de *prior probability shift*, la estimación $\hat{p}$ obtenida por el enfoque *CC* depende sólo de las características del 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):

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

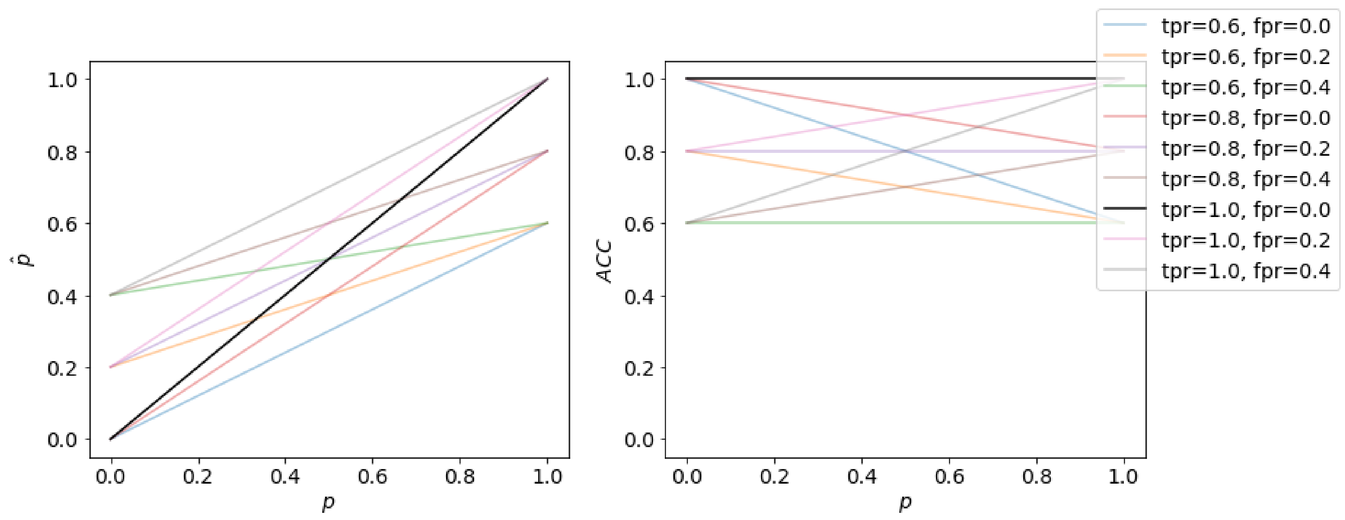

Figura 7: La línea negra representa el cuantificador y clasificador perfecto, respectivamente. Las otras líneas muestran las estimaciones teóricas de $\hat{p}$ resultantes de aplicar la ecuación **??**, y el *accuracy* correspondiente al clasificador, según se varían los valores de *tpr* y *fpr*

Mejora de la clasificación

Los métodos de cuantifacación también se pueden emplear para suplir los defectos de clasificadores frente a cambios en las distribuciones de los datos. Por ejemplo, en el caso del clasificador óptimo de Bayes, dado por:

$$h(\mathbf{x}) = \arg\max_y p_{Y|\mathbf{X}=\mathbf{x}}(y) = \arg\max_y \frac{p_{\mathbf{X}|Y=y}(\mathbf{x}) p_Y(y)}{p_{\mathbf{X}}(\mathbf{x})} \tag{2}$$

la decisión depende de $p_Y(y)$, que suele ser estimado con los datos de entrenamiento. Para mejorar el rendimiento del clasificador, se debería usar $\hat{p}_Y(y) = p_{tst}$, pero como p_{tst} suele ser desconocido, se puede usar un método de cuantificación para estimarlo.

Métodos

Algunos de los principales métodos de cuantificación son:

- **Clasificar y Contar (CC):** El *baseline* ya mencionado, cuyo estimador es:

$$\hat{p}^{CC}_{tst}(c = 1) = \frac{\#\{\mathbf{x} \in \mathbf{X}_{tst} | h_{tr}(\mathbf{x}) = 1\}}{\#\mathbf{X}_{tst}} \tag{3}$$

- **Clasificar, Contar y Ajustar (ACC):** Corrige las estimaciones de **CC** mediante las estimaciones de *tpr* y *fpr*:

$$\hat{p}^{ACC}_{tst}(c = 1) = \frac{\hat{p}^{CC}_{tst}(c = 1) - \hat{fpr}}{\hat{tpr} - \hat{fpr}} \tag{4}$$

- **Clasificar y Contar Probabilístico (PCC):** Utiliza los *scores* del clasificador, en vez de las clases predichas:

$$\hat{p}^{PCC}_{tst}(c = 1) = \frac{1}{m} \sum_{i=1}^{m} s(\mathbf{x}_i, y = 1) \tag{5}$$

- **Clasificar, Contar y Ajustar Probabilístico (PACC):** Combina las ideas de **ACC** y **PCC**:

$$\hat{p}^{PACC}_{tst}(c = 1) = \frac{\hat{p}^{PCC}_{tst}(c = 1) - \hat{fp}_{pa}}{\hat{tp}_{pa} - \hat{fp}_{pa}} \tag{6}$$

siendo tp_{pa} el promedio de los *scores* para individuos de clase positiva, y fp_{pa} para los de clase negativa.

- **Selección de Umbrales:** Busca un umbral que reduzca la varianza en las estimaciones de *tpr* y *fpr*:
  - MAX: selecciona el umbral que maximiza $tpr - fpr$.
  - X: busca obtener $fpr = 1 - tpr$.
  - T50: elige el umbral con $tpr = 0{,}5$, asumiendo que los positivos conforman la clase minoritaria.
  - Median Sweep (MS): para todos los umbrales que modifiquen los posibles valores de *fpr* y *tpr*, obtiene las prevalencias y calcula su mediana.

Objetivo y trabajo futuro

El objetivo de este trabajo es resumir el estado del arte en el área y evaluar mediante simulaciones los principales modelos propuestos. Como trabajo futuro, se sugiere proponer nuevos métodos de cuantificación, tanto para estimación puntual como por intervalos.

maximiliano.marufo@gmail.com