Cuantificación: Estimación de Prevalencias por Clase

Taller práctico

[Pablo González, Olaya Pérez-Mon]

[Universidad de Oviedo]

June 26, 2025

Objetivo del taller

- Poner en práctica los conceptos vistos en la sesión teórica: **Cuantificación**, *estimación de prevalencias por clase mediante aprendizaje supervisado.*
- Entender cuando necesitamos la cuantificación.
- Introducir herramientas para implementar cuantificadores de forma sencilla (quantificationlib).
- Aprender a entrenar cuantificadores y evaluarlos.

¿Qué es la cuantificación?

- En lugar de predecir etiquetas individuales, queremos estimar **la proporción de ejemplos por clase** en una **bag** de ejemplos no etiquetada.
- Ejemplo: estimar qué porcentaje de opiniones sobre un producto en Amazon son positivas, sin etiquetarlas una a una.
- Otras aplicaciones: control de calidad, análisis de opinión, medicina, biología, etc.

Asunción de cambio de distribución

- Suponemos que existe un **cambio de distribución** entre los datos de entrenamiento y los de test, concretatemente **Prior Probability Shift**:

$$p_{\text{train}}(x \mid y) = p_{\text{test}}(x \mid y) \quad \text{pero} \quad p_{\text{train}}(y) \neq p_{\text{test}}(y)$$

- Es decir: cambia la prevalencia de las clases, pero no el concepto de como son los ejemplos de cada clase.

Dataset: Amazon Reviews (LeQua)

- Opiniones de productos en Amazon.
- Dos clases: **positivas** y **negativas**.
- Se han preprocesado y convertido a vectores (bag-of-words o embeddings).
- Cada bolsa contiene un conjunto de opiniones con una distribución concreta de clases.

Opinión positiva

"Me encanta este producto, llegó rápido y es de muy buena calidad. Lo recomiendo sin duda."

Opinión negativa

"El producto no funcionaba, muy mala experiencia. No lo volveré a comprar."

¿Qué es quantificationlib?

- Librería Python para cuantificación:
  `https://github.com/AICGijon/quantificationlib`
- Proporciona implementaciones de los cuantificadores más relevantes:
  - **CC**: Classify Count (solución trivial).
  - **AC**: Adjusted Count.
  - **DFy**: Métodos basados en ajuste de distribuciones.
  - **EMQ**: Expectation Maximization.
- Permite entrenar y evaluar cuantificadores de forma sencilla.

Resumen del taller

- Explorar distintas estrategias de cuantificación.
- Evaluar su rendimiento bajo distintos cambios de distribución.
- Usar un dataset real y una herramienta práctica: `quantificationlib`.
- Comparar con soluciones triviales con cuantificadores más complejos y evaluar sus diferencias.

¡Manos a la obra!

Abramos el notebook:
https://colab.research.google.com/github/pglez82/quantificationlib_lab/blob/master/lab/lab.ipynb