# Imbalanced Data and Label Uncertaintly

#### Conjuntos de datos desbalanceados y etiquetado con incertidumbre

## Introducción

A medida que avanza la ciencia y la tecnología, la capacidad de recolectar datos se ha incrementado significativamente. Sin embargo, la presencia de datos desbalanceados, es decir, que las etiquetas de los datos no son proporcionales en el dataset, puede provocar que las técnicas de *machine learning* (ML) estándar vean afectado su rendimiento y su capacidad de predicción.

## Naturaleza del problema

### Datos desbalanceados

Cualquier conjunto de datos con una proporción desigual entre las clases se puede considerar desbalanceado, pero comúnmente se entiende por conjunto de datos desbalanceado cuando existe una diferencia significativa en la proporción de las etiquetas.

Suponiendo un conjunto de datos de mamografías (He & Garcia, 2009) con 10,923 “Negativos” (clase mayoritaria) y 260 “Positivos” (clase minoritaria) en relación con la presencia de cáncer, se desearía obtener un clasificador que determine la etiqueta sin errar en la predicción, con un 100% de acierto. En la realidad, el clasificador tenderá a clasificar con la etiqueta de la clase mayoritaria a prácticamente todo el conjunto de datos, obteniendo un grado de acierto cercano al 100%. Sin embargo, el grado de acierto en las observaciones que pertenecen a la clase contraria sería cercano a 0. En el ejemplo, 234 pacientes de cáncer de mama serían clasificados como sanos, lo que provocaría un error catastrófico para los pacientes y médicos, ya que el objetivo en medicina es detectar la clase minoritaria para poder ofrecer el tratamiento. Además, el tipo de error también es importante, ya que el coste de determinar erróneamente como cancerosa a una persona sana es mucho menor que el de diagnosticar como sana a una persona que realmente tiene cáncer. El clasificador ideal proporcionaría una alta precisión para la clase minoritaria sin afectar a la precisión de la clase mayoritaria.

Por esta razón, métricas convencionales como el *general* *accuracy* o el *error rate*, no proporcionan información adecuada para modelos entrenados con conjuntos de datos desbalanceados. Por tanto, es necesario utilizar métricas de evaluación más informativas como receiver *operating characteristic curve* (ROC), *precision-recall* y curvas de coste.

La presencia de datos desbalanceados puede deberse a la naturaleza de los datos en el espacio (desbalanceo intrínseco) o a factores como el tiempo o el almacenamiento de los datos y no está directamente relacionado con la naturaleza del espacio de los datos (desbalanceo etrínseco).

La primera de ellas es el desbalanceo intrínseco, debido a la naturaleza de los datos en el espacio.

## Referencias

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.