Autor: María José
Gómez Silva

CAPÍTULO 3. PREPROCESADO.

Antes de usar los datos para el entrenamiento y la evaluación de un modelo de Machine Learning, es necesario *preparar y preprocesar* esos datos. Entre las tareas más comunes de la etapa de preprocesado se encuentran las de localizar y corregir errores o ruido, eliminar valores atípicos y aplicar las transformaciones necesarias para la normalización de los valores presentados por los datos.

En este capítulo se describirán todas las técnicas mencionadas para el preprocesamiento de los datos, además de otras para el aumento de su cantidad y variedad, y la corrección del sesgo cuando sea necesario.

Tras la preparación de los datos, es necesaria la selección de aquellos más influyentes en la predicción del modelo o incluso la generación de nuevos datos o características a partir de los existentes, proceso conocido como *"feature engineering"*.

3. 1 Ruido

Ruido es el término general que se emplea para referirse a las modificaciones indeseadas y, en general, desconocidas que sufre una señal o colección de datos durante su captura, su almacenaje o su tratamiento. En ocasiones, también nos referimos con el término ruido a la información no útil que acompaña a nuestros datos. Los mecanismos empleados para la eliminación o reducción del ruido en los datos se suelen denominar, de forma genérica, *filtros*.

La presencia de ruido en nuestros datos puede deberse a múltiples causas y factores, como errores en las mediciones (por error humano o saturación de los sensores), errores en el almacenaje, limitaciones en la representación numérica de los computadores, o interferencias en la captura de los datos, entre otros.

Un ejemplo muy visual, es el del ruido presente en las imágenes. La Figura 66 muestra a la izquierda una imagen con ruido *gaussiano*. El ruido *gaussiano* suele deberse a la acumulación de errores e interferencias en el proceso de captura, trasmisión y almacenamiento de los datos y se presenta como pequeñas desviaciones de carácter

aleatorio sobre el valor ideal de los datos. La Figura 1 muestra una segunda imagen, a la derecha, con ruido de *sal y pimienta*. Este ruido suele estar ocasionado por la saturación de los sensores y aparece en las imágenes con la presencia de píxeles totalmente blancos y otros totalmente negros.

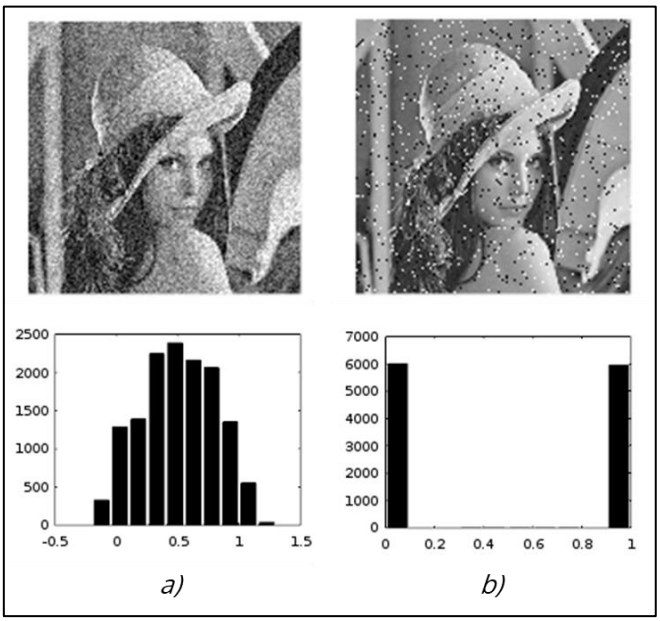

Figura 1. Imágenes con ruido a) gaussiano y b) de sal y pimienta e histogramas de los valores del ruido añadido a los valores de los píxeles de la imagen ideal.

Antes de emplear los datos recolectados en un algoritmo de Machine Learning es necesario su filtrado para que la presencia de ruido no dañe el rendimiento de los procesos posteriores.

En la práctica, tendremos que explorar los datos, e identificar el ruido o presencia de valores molestos a eliminar. Además, dependiendo del tipo de ruido, deberemos seleccionar el método de filtrado apropiado. Por ejemplo, el ruido *gaussiano* de la imagen anterior se puede suavizar mediante un filtro de media, que consiste en sustituir el valor de cada píxel por la media de los valores de los pixeles vecinos. Sin embargo, ese filtro tiene un resultado desastroso si se aplica a una imagen con ruido de sal y pimienta. En ese caso, en lugar del filtro de la media, hay que aplicar el filtro de la mediana.

En general, cuando tenemos una colección de datos en un *DataFrame*, deberemos aplicar diferentes procesos para eliminar cada tipo de ruido presente. Por ejemplo, pueden aparecer valores no medidos o desconocidos, que identificaremos como valores NaN. También, puede aparecer ruido como pequeñas fluctuaciones con respecto a la tendencia esperada de los datos, que suavizaremos con filtros paso bajo, como el de la media. O incluso, pueden aparecer valores atípicos, que identificaremos y descartaremos con técnicas de análisis estadístico. Todos estos procesos son descritos en las siguientes secciones.