

Machine Learning

Objetivo

El objetivo de la práctica es simple: abordar un problema de Machine Learning realista siguiendo la metodología y buenas prácticas explicadas durante las clases teóricas. Por tanto, en estas instrucciones no se especifican los pasos exactos que el alumno tiene que llevar a cabo para realizar esta tarea con éxito; es parte del trabajo aplicar las técnicas de procesamiento/transformación de variables que mejor se adecúen al problema, identificar los modelos que proporcionen prestaciones óptimas, las variables potencialmente más relevantes y la métrica adecuada para contrastar los distintos modelos. Aún así, se proporciona una pequeña guía de los pasos necesarios. Las posibilidades son amplias, así que es recomendable abordar una aproximación incremental: comenzar por soluciones sencillas para progresivamente aumentar la complejidad de las técnicas utilizadas.

A diferencia de los datasets utilizados en las clases, este está compuesto por datos reales, es decir, precisa de un análisis y limpieza mayores. Por el mismo motivo no se pretende obtener unos resultados espectaculares, es suficiente con que sean decentes; se valorará mucho más que el proceso seguido tenga sentido y no contenga errores graves de concepto.

Conjunto de datos

El conjunto de datos escogido es el dataset adjunto, extraído de Airbnb mediante técnicas de scraping.

Tarea

Es un problema de regresión: tenéis que predecir el precio del airbnb utilizando los datos disponibles.

1. Preparación de datos: División train/test
2. Análisis exploratorio, por ejemplo:
 - a. Head, describe, dtypes, etc.
 - b. Outliers
 - c. Correlación
3. Preprocesamiento:
 - a. Eliminación de variables, mediante selección (random forest/Lasso), alta correlación, alto porcentaje de missings, o el método que se considere oportuno.
 - b. Generación de variables
4. Modelado:
 - a. Cross validation

-
- b. Evaluación; mejor si lo hacéis de más de un modelo, porque así podéis comparar entre ellos.
 - 5. Conclusión: escrita, no numérica; un par de líneas es más que suficiente.

Modo de entrega

Hay que realizar la práctica en Python y subirla en un repositorio a GitHub o Drive. No basta con subir el código; hay que explicar lo que se ha hecho de forma suficientemente detallada, preferiblemente con gráficas y/o comentarios en markdown (o en el propio código Python, no hay problema). La estructura del proyecto es indiferente, puede ser en un archivo .py o en cuadernos de Jupyter .ipynb.