Práctica 3 - Aprendizaje Automático
Ajuste de Modelos Lineales

Límite de entrega: **29 de Mayo de 2022 a las 23:59 (PRADO)**

Valoración máxima: 12 puntos

Materiales a entregar: dos ficheros Python (.py o .ipynb), uno para cada uno de los problemas (regresión y clasificación) abordados, y un informe describiendo y analizando el trabajo desarrollado, los resultados obtenidos y las conclusiones extraídas. En caso de que se opte por entregar un Colab Notebook, el informe debe estar integrado en el mismo cuaderno (intercalando texto, código y resultados).

NORMAS DE DESARROLLO Y ENTREGA DE TRABAJOS

Es obligatorio presentar un informe con las valoraciones y decisiones adoptadas en el desarrollo de cada uno de los apartados. En dicho informe se incluirán los gráficos generados. También deberá incluirse una valoración sobre la calidad de los resultados encontrados. Sin este informe se considera que el trabajo NO ha sido presentado.

El incumplimiento de las normas que se listan a continuación puede implicar la pérdida de 2 puntos por cada norma incumplida:

- Cada ejercicio/apartado de la práctica debe quedar perfectamente identificado en el material entregado (código y memoria).
- Todos los resultados numéricos o gráficas serán mostrados por pantalla, parando la ejecución después de cada apartado.
- El codigo NO puede escribir nada a disco.
- El path que se use en la lectura de cualquier fichero auxiliar de datos debe ser siempre "datos/nombre_fichero". Es decir, se espera que el código lea de un directorio llamado "datos", situado dentro del directorio donde se desarrolla y ejecuta la práctica.
- Un código es apto para ser corregido si se puede ejecutar de principio a fin sin errores.
- No es válido usar opciones en las entradas. Para ello, se deben fijar al comienzo los valores por defecto que se consideren óptimos.
- El código debe estar obligatoriamente comentado explicando lo que realizan los distintos apartados y/o bloques.
- En caso de que haya más de un fichero (por ejemplo, *.py y *.pdf), estos se entregarán juntos dentro de un único fichero zip, sin ningún directorio que los contenga.
- Se entrega solamente el código fuente, y no los datos empleados.

Este trabajo se centra en el ajuste y selección del mejor predictor lineal para un conjunto de datos dado. Para ello, se recomienda el uso de la librería Scikit-Learn (https://scikit-learn.org/). Esta librería contiene funciones de alto nivel que pueden ser muy útiles para el desarrollo de la práctica. En cualquier caso, para cada función de Scikit-Learn que se use, debe de explicar por qué es necesario su uso, así como explicar su funcionamiento y el significado de todos sus parámetros. En relación con este punto, los valores por defecto en la librería no se consideran elecciones justificadas *a priori* y, al igual que en el resto de la práctica, decisiones sin justificación y resultados sin interpretación no serán considerados válidos.

Como mínimo se deben abordar y comentar las siguientes etapas en cada problema (clasificación y regresión):

1. Analizar y describir adecuadamente el problema a resolver. Identificar los elementos X, Y and f del problema, y describirlos en detalle.
2. Identificar qué conjuntos de hipótesis se emplearán y justificar dicha elección.
3. Si la base de datos define conjuntos de training y test, únalos en un solo conjunto y genere sus propios conjuntos de training y test. Describa y justifique el mecanismo de partición.
4. Justifique todos los detalles del preprocesado de los datos (codificación, transformación[1], normalización, etc). Es decir, todas las manipulaciones sobre los datos iniciales que nos permitan fijar el conjunto de vectores de características que se usarán en el entrenamiento.
5. Justifique la métrica de error a usar. Discutir su idoneidad para el problema.
6. Discuta todos los parámetros y el tipo de regularización usada en el ajuste de los modelos seleccionados. Justificar la idoneidad de la regularización elegida.
7. Selección de la mejor hipótesis para el problema. Discuta el enfoque seguido y el criterio de selección usado. ¿Cúal es su error E_{out}?
8. Construya las curvas de aprendizaje del modelo, y discuta la calidad del ajuste obtenido a la vista de la conducta de dichas curvas.
9. Suponga ahora que Ud. debe realizar este ajuste para una empresa que le ha proporcionado los datos, sin distinción entre training y test. ¿Cúal sería el mejor modelo que les propondría, y qué error E_{out} les diría que tiene? Justifique todas las decisiones.

Bases de datos a emplear en esta práctica:

- Problema de regresión: https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD
- Problema de clasificación: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing (fichero bank-full.csv).

Se recomienda desarrollar un código lo suficientemente general como para ser reutilizado en el desarrollo del Proyecto Final.

[1]Las transformaciones no-lineales de las variables pueden definirse a partir de las potencias y productos de potencias de las variables originales, conjuntos de polinomios ortogonales, etc. Si se usan transformaciones no polinómicas de las variable como log, $\sqrt{()}$, sin, etc, debe justificar el interés de las mismas.