Clasificación y validación cruzada

Estos ejercicios son para trabajar con lo visto sobre clasificación y selección de modelos utilizando validación cruzada. El conjunto de datos a estudiar es un conjunto de imágenes llamado MNIST. Cada imagen representa un dígito escrito a mano.

Descargar el dataset en formato csv.

Fecha de entrega: **14 de junio 2023.**
Les compartiremos un formulario para subirlo.

Ejercicios

1. Realizar un análisis exploratorio de los datos. Ver, entre otras cosas, cantidad de datos, cantidad y tipos de atributos, cantidad de clases de la variable de interés (el dígito) y otras características que consideren relevantes. ¿Cuáles parecen ser atributos relevantes? ¿Cuáles no? Se pueden hacer gráficos para abordar estas preguntas.
2. Construir un dataframe con el subconjunto que contiene solamente los dígitos 0 y 1.
3. Para este subconjunto de datos, ver cuántas muestras se tienen y determinar si está balanceado entre las clases.
4. Ajustar un modelo de knn considerando pocos atributos, por ejemplo 3. Probar con distintos conjuntos de 3 atributos y comparar resultados. Analizar utilizando otras cantidades de atributos.
5. Para comparar modelos, utilizar validación cruzada. Comparar modelos con distintos atributos y con distintos valores de k (vecinos). Para el análisis de los resultados, tener en cuenta las medidas de evaluación (por ejemplo, la exactitud) y la cantidad de atributos.
6. Trabajar nuevamente con el dataset de todos los dígitos. Ajustar un modelo de árbol de decisión. Analizar distintas profundidades.
7. Para comparar y seleccionar los árboles de decisión, utilizar validación cruzada con k-folding.
8. Les daremos un conjunto de test el día de la entrega, para que puedan evaluar sus modelos y reportar la performance.

Entrega

Preparar los siguientes archivos. Les compartiremos un formulario para subirlos.

- Un archivo llamado `digitos_nombregrupo.py` con el código principal. Este archivo puede complementarse con otros archivos `.py` donde figure parte del código, y que sean importados y utilizados desde el archivo principal.

  Como siempre, ordenar el código de la siguiente manera:
  - al inicio, una descripción que contemple: el nombre del grupo, los nombres de lxs participantes, contenido del archivo y cualquier otro dato relevante que considere importante.
  - luego la sección de los imports
  - luego la carga de datos
  - luego las funciones propias que hayan definido
  - y finalmente, el código que no está dentro de funciones

  El código debe estar modularizado (separando bloques con #%%) para permitir su ejecución por fragmentos.
  Todo lo que figure en el informe debe deducirse de los resultados del código.

- Un informe en pdf llamado `informe_tp2_nombregrupo.pdf` de aproximadamente 3 carillas.

  Ordenar el informe de la siguiente manera:
  - Breve introducción al problema donde se muestre el análisis exploratorio realizado.
  - Explicación sobre los experimentos realizados, incluyendo los gráficos que consideren convenientes.
  - Conclusiones, incluyendo los resultados relevantes de los modelos desarrollados.