Clasificación y validación cruzada

Objetivo del Trabajo Práctico 02

Evaluar lo visto en clase sobre clasificación y selección de modelos, utilizando validación cruzada.

Enunciado

En el presente TP trabajaremos con el conjunto de datos de imágenes denominado **Fashion MNIST** [1]. Cada imagen del set de datos representa una prenda de ropa. En el link ubicado a pie de página pueden acceder a una descripción más detallada del dataset.

Para comenzar deben descargar del campus de la materia el conjunto de datos, el cual se encuentra en formato csv.

Fecha de entrega: **12 de noviembre de 2023 23:50hs.** Al igual que el TP-01, la entrega de este TP se realizará a través del campus de la materia.

Ejercicios

1. Realizar un ***análisis exploratorio*** de los datos. Entre otras cosas, deben analizar la cantidad de datos, cantidad y tipos de atributos, cantidad de clases de la variable de interés (la prenda de ropa) y otras características que consideren relevantes. Además se espera que con su análisis puedan responder las siguientes preguntas:
    a. ¿Cuáles parecen ser atributos relevantes para predecir el tipo de prenda? ¿Cuáles no? ¿Creen que se pueden descartar atributos?
    b. ¿Hay clases de prendas que son parecidas entre sí? Por ejemplo, ¿Qué es más fácil de diferenciar: remeras de pantalones o remeras de pullovers?
    c. Tomen una de las clases, por ejemplo los vestidos, ¿Son todos muy similares entre sí?
    d. Este dataset está compuesto por imágenes, esto plantea una diferencia frente a los datos que utilizamos en las clases (por ejemplo, el dataset de Titanic). ¿Creen que esto complica la exploración de los datos?

[1] **Fashion MNIST**. https://www.kaggle.com/datasets/zalando-research/fashionmnist

Importante: las respuestas correspondientes a los puntos 1.a, 1.b y 1.c deben ser justificadas en base a gráficos de distinto tipo.

2. Dada una imágen se desea responder la siguiente pregunta: **¿la imágen corresponde a un pantalón o a una remera?**
    a. A partir del dataframe original, construir un nuevo dataframe que contenga sólo al subconjunto de imágenes correspondientes a las remeras y a los pantalones.
    b. Sobre este subconjunto de datos, analizar cuántas muestras se tienen y determinar si está balanceado en función de la clase a predecir.
    Importante: Separar también los datos en conjuntos de train y test.
    c. 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.
    Importante: Para evaluar los resultados de cada modelo usar el conjunto de test generado en el punto anterior.
    d. Comparar modelos de KNN utilizando distintos atributos y 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.
3. ***(Clasificación multiclase)*** Dada una imágen se desea responder la siguiente pregunta: ***¿qué tipo de prenda es la que representa la imágen?***
    a. Vamos a trabajar con los datos correspondientes a los 10 tipos de prendas. Separar el conjunto de datos en train y test.
    b. Ajustar un modelo de árbol de decisión. Analizar distintas profundidades.
    c. Para comparar y seleccionar los árboles de decisión, utilizar validación cruzada con k-folding.
    Importante: Para hacer k-folding utilizar los datos del conjunto de train.
    d. ¿Cuál fue el mejor modelo? Evaluar cada uno de los modelos utilizando el conjunto de test. Reportar su mejor modelo en el informe.

Grupos

Los grupos deben estar conformados por 3 (y sólo 3) integrantes. Ni más, ni menos. Deberán i) registrar la conformación del grupo en la siguiente planilla, y ii) definir quién va a ser el encargado del envío (debe ser uno y sólo uno de los integrantes del grupo):

https://docs.google.com/spreadsheets/d/1g3SJ4PtmmfgtVFuVOxiWM16tuXoiH21u9lxHas2IiGE/edit?usp=sharing

Acerca de la entrega

Para la entrega deberán preparar los siguientes archivos:

- Un archivo llamado *fashion_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.
  - A continuación, la carga de datos.
  - Siguiendo, 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.

  Importante: Incluir un archivo README.txt con los requerimientos de bibliotecas utilizadas e instrucciones de cómo ejecutar el código.

- Un informe breve (no más de 10 carillas) en pdf llamado *informe_tp2_nombregrupo.pdf*.

  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.

Importante: ¡No deben entregar los archivos del dataset!