PR2.2. ALGORITMOS DE APRENDIZAJE SUPERVISADO: KNN

Descripción de la tarea

Realizar estudios de aprendizaje supervisado utilizando el algoritmo de los k vecinos más cercanos (knn). Para ello en esta tarea se utilizarán tres datasets.

1. https://www.kaggle.com/yasserh/wine-quality-dataset . Clasificación de la calidad del vino según alguno de sus parámetros (ph, densidad, sulfatos, alcohol, azúcar, etc.). El dataset contiene una serie de datos sobre unos vino y su calidad (de 0 a 10). Este problema es de clasificación puesto que consiste en clasificar el vino en concreto en su clase de calidad (valores discretos del cero al 10)

2. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction . Clasificación binaria. El objetivo es clasificar entre presencia de enfermedad del corazón o no a partir de una serie de atributos (edad, sexo, colesterol, etc.). Para ello el dataset contiene una serie de datos pacientes etiquetados si tienen enfermedad o no.

3. https://www.kaggle.com/mssmartypants/paris-housing-price-prediction. Predicción de precios de las casas de París. El dataset contiene una serie de datos de casas, como por ejemplo número de habitaciones, año de construcción, piscina, garaje, etc. y el precio por el que se han vendido. El objetivo es, dada una serie de parámetros, predecir el precio que debe tener la casa.

Para cada uno de ellos realizar un documento de google colab con los siguientes epígrafes y tareas:

- Importación de librerías necesarias
- Preproceso
  - Importación de los datos del dataset
  - Mostrar las primeras y últimas filas del dataframe importado
  - Mostrar parámetros estadísticos de los datos (media, desviación típica, cuartiles, etc.)
  - Mostrar un mapa de calor que indique la correlación entre variables
  - Seleccionar las características a tener en cuenta en el estudio
  - Separar datos entre datos de entrada y etiquetas (resultados)
  - Separar datos entre entrenamiento y prueba (usando un 75% para entrenamiento y 25% para test)

- Entrenamiento y predicción
  - Elegir, instanciar (eligiendo unos valores concretos, por ejemplo k=3 y w='uniform') y entrenar el modelo
  - Realizar una predicción con los datos de prueba
- Evaluación
  - Para los problemas de clasificación
    - Mostrar el porcentaje de elementos correctamente clasificados
    - Mostrar la predicción realizada (imprimir la variable con la predicción)
    - Representar gráficamente la clasificación obtenida (matriz de confusión)
  - Para los problemas de regresión
    - Mostrar el error cuadrático medio (mean_squared_error)
    - Mostrar el error absoluto medio (mean_absolute_error)
    - Representar gráficamente los valores predichos con los valores reales.

```
# x axis for plotting
import numpy as np
import matplotlib.pyplot as plt
xx = np.stack(i for i in range(y_test.shape[0]))
plt.plot(xx, y_test, c='r', label='data')
plt.plot(xx, y_pred, c='g', label='prediction')
plt.axis('tight')
plt.legend()
plt.show()
```

- Optimización de hiperparámetros
  - Calcula la combinación de parámetros óptima (uniform o distance; valor de k). Para ello realiza ejecuciones con cada uno de los valores uniform y distance para los valores de k desde 1 a 30.
  - Cada ejecución anterior se deberá hacer usando validación cruzada (por ejemplo n_splits = 5). Con ello obtendremos una medida de bondad del modelo (accuracy_score o mean_absolute_error)), como lo ejecutaremos 5 veces, calcularemos la media de esas 5 ejecuciones.
  - Finalmente los parámetros elegidos serán los que den mejor media de esas medidas anteriormente nombradas.
  - Una vez obtenidos esos parámetros óptimos los aplicaremos al problema en cuestión y mostraremos los resultados.

Documentación a entregar. Se entregará en moodle centros un archivo (cuaderno de júpiter o google colab) para cada uno de los ejercicios con el desarrollo de las tareas anteriormente mencionadas. Además se entregará un fichero pdf con unas conclusiones (para cada ejercicio) que describan todas las fases del proceso de aprendizaje supervisado que se han ido desarrollando en esta tarea (explicación de qué se ha realizado)

Evaluación

Los criterios de evaluación de la tarea son a, b, c, d, e, f, g, h del RA3. Aplica algoritmos de aprendizaje supervisado, optimizando el resultado del modelo y minimizando los riesgos asociados.

Para evaluar la práctica se puntúan los siguientes apartados:

- Clasificación de vinos (3,33 puntos)
  - Preproceso; Entrenamiento y predicción; Evaluación; Optimización; Conclusiones.
- Clasificación presencia de enfermedad de corazón (3,33 puntos)
  - Preproceso; Entrenamiento y predicción; Evaluación; Optimización; Conclusiones.
- Predicción del precio de las casas en Paris (3,33 puntos)
  - Preproceso; Entrenamiento y predicción; Evaluación; Optimización; Conclusiones.

Rúbrica por cada problema

	100%	**75%**	**50%**	**25%**	**0%**
Preproceso	Proporciona los datos etiquetados al modelo, importándolos y realizando todas la modificaciones necesarias.	Proporciona los datos etiquetados al modelo, hay algunos errores menores en la importación, pero los datos se adaptan	Proporciona los datos etiquetados al modelo, hay algunos errores en la importación o no se realizan modificaciones	Proporciona los datos etiquetados al modelo, hay errores en la importación y no se realizan modificaciones	No realiza la importación de datos, o la realiza de forma errónea
Entrenamiento y predicción	Realiza el entrenamiento del modelo y la predicción con los datos de test	Realiza el entrenamiento del modelo y la predicción con los datos de test. Pero existen algunos errores menores	Realiza el entrenamiento correctamente pero no la predicción	Realiza el entrenamiento pero no la predicción y existen algunos errores en el entrenamiento	No realiza el entrenamiento del modelo, ni la predicción o presentan errores
Evaluación	Se realiza una evaluación del modelo con los datos. Además se representa de diversas formas, gráficamente en texto	Se realiza una evaluación del modelo con los datos. Sólo se presenta un gráfico	Se realiza una evaluación del modelo con los datos. Se muestra un gráfico pero con algunos errores menores	Se realiza una evaluación del modelo con los datos. Sólamente en texto	No se realiza la evaluación o presenta errores importantes
Optimización	Se realiza una optimización de hiperparámetros o una comparativa entre distintos modelos de la misma familia.	Se realiza una optimización de hiperparámetros o una comparativa pero existen algunos errores menores o no está suficientemente detallado	Se realiza una optimización de hiperparámetros o Hay algún error en los los valores óptimos o en la comparativa.	Se realiza una optimización de hiperparámetros o una comparativa. Existen errores importantes, no se consiguen los valores óptimos o la comparativa es escasa o irrelevante	No se realiza la optimización ni se prueban diversos modelos de la misma familia
Conclusiones	Realiza unas conclusiones que describen todas las fases del proceso de solucionar un problema con aprendizaje supervisado	Realiza unas conclusiones que describen la mayoría de las fases del proceso de solucionar un problema con aprendizaje	Realiza unas conclusiones que describen al menos la optimización de parámetros y la evaluación de los resultados obtenidos	Realiza unas conclusiones que describen alguna de las fases del proceso de solucionar un problema con aprendizaje	No realiza conclusiones o son erróneas o irrelevantes

Refuerzo y ampliación

Refuerzo.

1. Realiza el estudio de aprendizaje supervisado (usando knn) para el problema de clasificación de los distintos tipos de flores iris según el ancho y el alto del pétalo y sépalo. Para ello utiliza el dataset iris (https://archive-beta.ics.uci.edu/dataset/53/iris)

Ampliación.

- Realiza un estudio de aprendizaje supervisado con el siguiente dataset (para predecir el precio de ordenadores) https://www.kaggle.com/muhammetvarl/laptop-price
- Estudia cómo podrías cambiar el problema de clasificación de la calidad de los vinos a un problema de regresión en el que obtengamos como resultado valores continuos.
- Consulta los siguientes enlaces que contienen datasets para machine learning https://archive-beta.ics.uci.edu/datasets y https://www.kaggle.com/datasets.
    - Elige algún/algunos datasets y plantea y resuelve un problema de clasificación y otro de regresión usando el algoritmo de los k vecinos más cercanos (knn).
- Realiza una implementación sencilla del algoritmo, para ello asumirá que todos los vecinos importan los mismo, es decir w=uniform, por lo que la variable que tomará es k (el número de vecinos a considerar)
    - Recordemos el algoritmo:
        - Calcular la distancia entre el nuevo ejemplo y todas las muestras de entrenamiento.
        - Ordenar por la distancia y determinar los K vecinos más cercanos.
        - Determinar el resultado
            - Clasificación → mayoría simple de las clases de los vecinos.
            - Regresión → media de los valores de los vecinos.