prog_datasci_7_analisis_entrega

May 21, 2020

1 Programación para *Data Science*

1.1 PEC 7 - Introducción al análisis de datos en Python

En este Notebook encontraréis el conjunto de actividades evaluables como PEC de la asignatura. Veréis que cada una de ellas tiene asociada una puntuación, que indica el peso que tiene la actividad sobre la nota final de la PEC. Adicionalmente, hay un ejercicio opcional, que no tiene puntuación dentro de la PEC, pero que se valora al final del semestre de cara a conceder las matrículas de honor y redondear las notas finales. Podréis sacar la máxima nota de la PAC sin necesidad de hacer este ejercicio. El objetivo de este ejercicio es que sirva como pequeño reto para los estudiantes que quieran profundizar en el contenido de la asignatura.

Veréis que todas las actividades de la PEC tienen una etiqueta, que indica los recursos necesarios para llevarla a cabo. Hay tres posibles etiquetas:

- NM **Sólo materiales**: las herramientas necesarias para realizar la actividad se pueden encontrar en los materiales de la asignatura.
- EG **Consulta externa guiada**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, pero el enunciado contiene indicaciones de dónde o cómo encontrar la información adicional necesaria para resolver la actividad.
- EI **Consulta externa independiente**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, y el enunciado puede no incluir la descripción de dónde o cómo encontrar esta información adicional. Será necesario que el estudiante busque esta información utilizando los recursos que se han explicado en la asignatura.

Es importante notar que estas etiquetas no indican el nivel de dificultad del ejercicio, sino únicamente la necesidad de consulta de documentación externa para su resolución. Además, recordad que las **etiquetas son informativas**, pero podréis consultar referencias externas en cualquier momento (aunque no se indique explícitamente) o puede ser que podáis hacer una actividad sin consultar ningún tipo de documentación. Por ejemplo, para resolver una actividad que sólo requiera los materiales de la asignatura, puedéis consultar referencias externas si queréis, ya sea tanto para ayudaros en la resolución como para ampliar el conocimiento!

En cuanto a la consulta de documentación externa en la resolución de los ejercicios, recordad **citar siempre la bibliografía utilizada** para resolver cada actividad.

1.2 Ejercicio 1

Carga el conjunto de datos wine incorporado en la librería `sklearn`. Para resolver este ejercicio, sólo podéis utilizar la sección `data`.

En este conjunto de datos, hay vinos de distintas categorias, pero de momento no sabemos a qué categoría pertenence cada uno.

Dado este conjunto de datos, en un restaurante nos piden adivinar en cuántos grupos podríamos categorizar los vinos para así almacenarlos en bodegas distintas, intentando que esto requiera el mínimo número de salas pero que al mismo tiempo permita representar lo mejor posible las diferencias entre estos vinos.

1) El algoritmo K-Means nos permite agrupar los vinos en distintos clusters. No obstante, será necesario evaluar su rendimiento para separar las distintas categorías en relación al número de clusters utilizados. Así, podemos representar el número de habitaciones simplemente ajustando el número de clusters en el algoritmo K-means. ¿Cuál es el menor número de habitaciones que permiten almacenar los vinos con la mayor diferencia entre los vinos de cada habitación? Implementa la solución y muestra el resultado gráficamente. **(2 puntos)**

Nota:

En los ejercicios resueltos hemos visto como usar la funcion `fit` de K-means para separar los datos. Tened en cuenta que, una vez hecho el *fit*, la función `inertia_` nos permite cuantificar el rendimiento del algoritmo para separar los grupos (separarabilidad).

2) ¿Sería posible utilizar una técnica de análisis de datos **supervisada** para resolver este problema? ¿Por qué? **(1 punto)**

NM

[1]:
```
# Respuesta
```

Respuesta

1.3 Ejercicio 2

A partir de estos datos crea un dataframe, incluyendo una columna con el nombre *class*, que indique el tipo de vino, es decir la variable `target`. El resto de columnas corresponderán con las características o `feature_names`. Finalmente, muestra por pantalla la cabecera del *dataframe* y los descriptivos estadísticos del *dataset*.

(1 punto) NM

[2]:
```
### Respuesta
```

1.4 Ejercicio 3

Ahora que tenéis el conjunto de datos organizados, trazad la distribución de cada una de las características para cada clase de vino individualmente. Para ello podéis utilizar historgramas como vimos en el Notebook de teoría o podéis elegir utilizar un diagrama de violín de la librería Seaborn. Aseguraos de mostrar una sola figura con múltiples paneles, dedicando un panel a cada característica de los vinos (ej. magnesio). En cada panel se mostrarán las distribuciones de valores para cada clase de vino. **(1 punto)** NM

[3]:
```
# Respuesta
```

1.5 Ejercicio 4

El **análisis discriminante lineal (LDA por sus siglas en inglés)** es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características. Haciendo uso del teorema de Bayes, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores (características), pertenezca a cada una de las clases de la variable cualitativa, $P(Y = k|X = x)$. Finalmente se asigna la observación a la clase k para la que la probabilidad predicha es mayor.

En este ejercicio, tendréis que crear un modelo basado en LDA para clasificar las clases de vino según las características del conjunto de datos.

3.1- En primer lugar, tendréis que estandarizar los datos utilizando el método StandardScaler de sklearn.

3.2- A continuación, tendréis que dividir los datos en conjuntos de entrenamiento (*training*) y evaluación (*test*), con un tamaño de la muestra para test del 33%.

3.3- A continuación, entrenad el modelo LDA con la porción de datos para el entrenamiento y calculad la precisión del modelo. Tendrés que usar el classificador LDA de sklearn para entrenar el modelo y el método score( ) para evaluarlo. Escribid una breve explicación que indique lo que significa obtener tal precisión.

(3 puntos) EG

[6]:
```
# Respuesta

import numpy as np
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Eliminad la columna de clase del conjunto, ya que será la variable que␣
 ↪queremos predecir
X = wine_data_df.drop(['class'], axis=1)
y = wine_data_df['class']
```

1.6 Ejercicio 5

Ahora que tenéis el modelo entrenado, tendréis que aplicar la transformación del LDA al conjunto de datos de forma que os permita reducir su dimensionalidad a sólo 2 componentes. Una vez hecho esto, tendréis que representar la transformación resultante gráficamente utilizando un diagrama de dispersión. Para ello, deberéis generar dos figuras, una con los datos transformados del conjunto de entrenamiento (*training*) y otra del conjunto de evaluación (*test*).

Nota: No olvidéis colorear cada punto de datos segun la clase de vino a la que corresponde.

(2 puntos) EI

[4]:
```
# Respuesta
```

1.7 Ejercicio Opcional

Ahora que hemos aprendido algunas cosas sobre los tipos de vino y sus características, imagina que un amigo está tomando una copa de vino y te lee el siguiente etiquetado:

```
alcohol =                          14.23 malic_acid =                        1.71
ash =                               2.43 alcalinity_of_ash =                15.60
magnesium =                       127.00 total_phenols =                     2.80
flavanoids =                        3.06 nonflavanoid_phenols =              0.28
proanthocyanins =                   2.29 color_intensity =                   5.64
hue =                               1.04 od280_od315_of_diluted_wines =      3.92
proline =                        1065.00
```

¿Podrías adivinar qué clase de vino está tomando tu amigo? ¿Cuál es la probabilidad de equivocarte?

Nota: Utiliza el modelo LDA entrenado para adivinar la clase de vino.

EG

[6]:
```
# Respuesta
```