Práctica 4
Regresión y clústering

Objetivo

Desarrollar un Jupyter-Notebook por entregable (3 en total) que permitan implementar algoritmos para la obtención de relgas de asociación y patrones secuenciales.

Herramientas

- Lenguaje de programación: Python
- Librerías: Numpy, pandas, scikit-learn, matplotlib
- Entorno de gestion de librerías: Anaconda
- Editor: Jupyter

Información de la entrega

La entrega se realizará a través de la tarea LAB4 disponible en la página Canvas de la asignatura.
Consistirá en un fichero comprimido (.zip,.tar.gz) con nombre LAB04-GRUPOxx.zip que contendrá:

1. Un Jupyter-Notebook por cada entregable (archivos con extensión .ipynb)
2. La memoria del laboratorio se entregará integrada en el Notebook de manera que explique y complemente el código entregado
3. El código entregado tiene que ser funcional, correcto y completo.

Las entregas que no se ajusten exactamente a este formato NO SE EVALUARÁN.

Rúbrica

Código

El valor de cada entregable para la nota final de la práctica se indica en el enunciado, así como el valor de cada uno de los apartados.
Todos los aspectos de programación se dan por supuestos.
El código debe ser:

- Funcional: debe ejecutar sin errores y el resultado debe ser el esperable
- Original: el código no puede ser una copia de trabajos publicados en Internet o de otros compañeros. Grupos con código igual serán suspendidos.
- No redundante: se penalizará el código que no sea útil o redundante
- Comentado: es obligatorio incluir comentarios en el código, en su justa medida
- Gráficas: deben incluir todos los datos que sean necesarios

Memoria

La memoria estará incluida en los Jupyter-Notebook que se entreguen de manera que complementen el código entregado. La redacción debe ser clara y correcta ortográfica y gramaticalmente. Debe incluir la justificación de cada paso que se realice para la resolución de los problemas planteados.

Entregable 1 – Precio de los Inmuebles

Este entregable vale 3,5 puntos de la nota final de la práctica 4.

Si preguntamos por la casa de sus sueños a un posible comprador, nos la describirá sin tener en cuenta la altura de los techos o su proximidad a una carretera de circunvalación. Sin embargo, este tipo de características sí pueden influir el precio final de un inmueble.
La inmobiliaria Casas.SA quiere poder asesorar mejor a sus clientes cuando quieren comprar una casa en Ames, Iowa. Para eso, nos proporcionan un dataset (casas.csv) que contiene 79 variables que describen (casi) cada aspecto de las casas en esa ciudad.

1) Para empezar, cread un modelo en el que se tengan en cuenta todas las características de los inmuebles para predecir/estimar el precio de un inmueble. (0.75 puntos)
2) Haz un análisis de residuos del modelo lo más exhaustivo posible (0.5 puntos)
3) Crea un modelo que permita saber qué variables afectan más a la hora de predecir el precio del inmueble(0.75 puntos)

Vamos a crear ahora modelos más simples que nos permitan adelantar el precio de un inmueble.

4) Vamos a crear ahora un modelo más sencillo teniendo en cuenta sólo algo obvio: el tamaño del inmueble (LotArea), para hacer un forecast del precio del inmueble. Calculad el coeficiente r de pearson entre LotArea y el precio del inmueble y dibujad el modelo. (0.5 puntos)
5) Cread también un modelo que tenga en cuenta la variable que representa la cantidad de calle pegada a la propiedad (LotFrontage). Calculad el coeficiente r de pearson entre LotFrontage y el precio del inmueble y dibujad el modelo. (0.5 puntos)
6) Comparad los modelos entre ellos y con el modelo completo del apartado 1). ¿Los modelos sencillos se acercan al modelo completo? ¿Qué diferencias encuentras?¿Qué características nos habrían resultado mejores para crear modelos más simples de predicción? (0,5puntos)

DATA MINING

LAB 04

Regresión y clústering

Entregable 2 – Características de los Inmuebles

Este entregable vale 4 puntos de la nota final de la práctica 4.

La inmobiliaria Casas.SA quiere dar una asistencia completa a sus clientes. Para eso, necesita analizar más en profundidad las características de los inmuebles en la ciudad de Ames para poder agrupar los inmuebles y así poder enseñar a los clientes aquellos que les puedan interesar más según las características que deseen.
A partir del dataset de casas.csv, aplica un algoritmo para obtener los grupos de inmuebles.

1) Realiza todo el preprocesamiento que sea necesario para adaptar las variables que no sean unívocas del dataset y poder usar el algoritmo adecuado. (1 punto)
2) Utiliza varias configuraciones teniendo en cuenta el número de grupos a crear y cambiando la medida de distancia entre individuos. Crea una tabla donde se incluya toda la información y el número de iteraciones necesarias para llegar a la solución que presentas. Se considera la mejor solución a aquella que necesita menos iteraciones (1 punto)
3) Con la mejor configuración del apartado anterior, utiliza dos criterios para elegir el lugar inicial del punto central de los grupos. Dibuja cómo se van modificando los grupos y cómo van cambiando sus centroides en cada iteración. Obtén una conclusión acerca de dónde deberían situarse los centroides. (1 punto)
4) Estudia qué técnicas de postprocesamiento se podrían aplicar en base al error cometido en cada clúster (1 punto)

Entregable 3 –

Este entregable vale 2,5 puntos de la nota final de la práctica 4.

Una bodega quiere saber si los vinos que produce son diferentes, para eso ha mandado a un laboratorio el análisis químicos de 13 tipos de vinos. A partir de este análisis, recogido en el archivo vinos.csv se nos pide analizar los datos para ayudar a caracterizar los vinos antes de su venta.

1) Utiliza varias configuraciones para el modelo aplicando “single linkage” que más se adapte y teniendo en cuenta los tipos de distancias entre elementos. ¿Cuál es la k del modelo? (1,5 puntos)
2) Dibuja un dendograma con los clústeres obtenidos. Explica alguna de las relaciones interesantes que puedes encontrar. ¿Se pueden identificar claramente varios tipos de vinos? ¿qué características los definirían? (1 punto)

DATA MINING

LAB 04

Regresión y clústering