Universidad
Rey Juan Carlos

Práctica 1
Análisis de datos

Curso 2017-2018

En esta práctica se pretende profundizar en el análisis de técnicas supervisadas lineales y en técnicas no supervisadas. En concreto, se empleará el método de regresión lineal simple y múltiple así como el algoritmo k-means, vistos en clase de teoría. Para el desarrollo de esta práctica puede utilizar el lenguaje de programación que considere más oportuno.

1. Análisis exploratorio

Para el desarrollo de esta práctica se utilizará las bases de datos*: winequality-white.csv y Facebook_metrics.txt.*

Indique el número de muestras así como el número de características que representan cada muestra. ¿Considera que la características son relevantes? Justifique su respuesta y describe la información que aporta cada característica.

Lleve a cabo un análisis exploratorio de los datos (missing values, estadísticos tales como la media, el máximo o el mínimo, diagrama de cajas).

Obtenga para cada característica su histograma. ¿Qué representa el eje-x y el eje-y de cada histograma? Describe y compare los resultados obtenidos.

Represente y comente distintos gráficos de dispersión. ¿Qué información obtiene?

2. Regresión lineal simple

La regresión lineal permite establecer la relación que existe entre una variable dependiente y una o varias variables explicativas. Esta técnica se emplea cuando la relación que existe entre las variables es lineal.

2.2 Regresión lineal

Indique, para cada base de datos, cuál es la variable dependiente, cuáles son las variables que más influyen en dicha variable, el valor del coeficiente de correlación y el porcentaje de error. A la vista de los resultados, justifique para qué base de datos se obtiene mejores resultados.

2.3 Regresión lineal. Selección de características

En este apartado vamos a analizar como cambian: el valor de los estimadores, el coeficiente de correlación y el porcentaje de error cuando cambiamos el número de variables explicativas. Explique detalladamente el proceso que ha seguido y justifique qué características le resultan más relevantes en esta aplicación.
Nota: es posible que tenga que repetir este apartado con distinto número de variables para poder extraer conclusiones válidas.

Por último, ejecute el método de clasificación denominado Simple Linear Regression. Como ya se ha visto en teoría, este método se caracteriza porque solo se considera una variable independiente para explicar el valor de la variable dependiente. Justifique qué característica consideraría en este caso, e indique el coeficiente de correlación y el porcentaje de error. ¿Mejoran los resultados?

2.4 Aprendizaje no supervisado. Algoritmo k-medias

El análisis de clusters consiste en dividir los datos en grupos de objetos (o clusters) basándose simplemente en la información contenida en los datos que describen estos objetos y las relaciones que existen entre ellos. El clustering puede ser visto como una clase de clasificación, en la cual, se realiza un etiquetado de los objetos con las etiquetas de los clusters. Es por esta razón, a menudo se refiere al clustering como clasificación no supervisada. Recordar que todos los conjuntos de datos tienen un atributo que representa la clase. Para tareas de clustering, este atributo no tiene interés, por tanto, lo vamos a ignorar excepto para la fase de evaluación.

Ejecute de nuevo varias vece el algoritmo cambiando el valor de parámetro k (número de clusters). Obtenga el SSE (la suma de los errores cuadráticos medios) para cada valor , así como la media y la desviación típica de cada grupo. A la vista de los resultados, ¿cuál sería el mejor valor para el parámetro numCluster?

2.5 Aprendizaje no supervisado. Algoritmo k-medias. Selección de características

Seleccione un subconjunto de características. Justifique qué características le resulten más relevantes en esta aplicación y ejecute el algoritmo utilizando solamente ese subconjunto. ¿Mejora el SSE obtenido?