Módulo 4. Parte III. Aprendizaje con R

Enunciados de Ejercicios

Autor: Raquel Dormido Canto

Actualizado Enero 2023

Contenido

EJERCICIO 1: Regresión lineal

En el fichero **Venecia.txt** tenemos almacenadas 50 observaciones anuales del nivel del mar en Venecia. Los datos corresponden a los años 1931-1981. Estos datos son datos reales publicados en la referencia: Smith R. L, "Extreme value theory based on the r largest annual events", Journal of Hydrology, 86 (1986).

Calcular y analizar un modelo de regresión lineal simple para estos datos entre las variables `año` y `nivel`. Para ello:

a) Dibujar la recta de ajuste del modelo junto con la correspondiente nube de puntos.
b) Obtener un resumen del ajuste.
c) ¿Cuánto vale el coeficiente de correlación al cuadrado en este caso?
d) ¿Cuánto valen los estimadores de todos los parámetros del modelo?
e) Calcula un intervalo de confianza con un nivel del 90%
f) Calcular y representar los intervalos de confianza y de predicción al 95% del nivel medio para los años comprendidos entre 1940 y 1970.
g) Calcular los residuos estandarizados frente a los valores ajustados y verificar la hipótesis de normalidad.

****

EJERCICIO 2: Regresión lineal múltiple

En el fichero **granizados.txt** tenemos almacenadas una serie de datos que se pretenden analizar para identificar qué factores son los más influyentes a la hora de consumir granizados. En concreto los datos corresponden a mediciones realizadas en una familia durante 30 semanas (entre el 18 de marzo de 1953 y el 11 de julio de 1953) del consumo por semana de granizado por persona (`y`). Aparte de `y` otras variables almacenadas en el fichero, que se pensaba que podían tener alguna influencia sobre el consumo son:

`p`: precio de una pinta de granizado
`i`: ingresos semanales de la familia
`temp`: temperatura media de la semana
`SEMANA`: número de semana

Teniendo en cuenta esta información queremos realizar lo siguiente:

a) Representar gráficamente el consumo de granizados en función de las semanas.
b) Determinar la matriz de correlación de las variables `y`, `p`, `i` y `temp`.
c) ¿Cuál es la variable que parece tener más influencia en `y`?
d) Realizar un ajuste lineal de `y` sobre `p`, `i` y `temp`. Dibujar la evolución predicha por nuestro modelo. Analizar si todas las variables son igual de significativas en el modelo.
e) ¿Cuánto vale la varianza residual y R2 ?
f) Realizar un ajuste lineal de `y` sobre `i` y `temp`. ¿Cuánto vale en este caso la varianza residual y R2?

****

EJERCICIO 3: KNN

Uno de los usos del Machine Learning está relacionado con la detección de tumores oncológicos, de manera que, es posible construir modelos para detectar si un tumor es maligno o no o para predecir el crecimiento anormal de las células.

En el fichero **tumor.csv** se almacenan los datos de 100 pacientes con 10 variables por paciente. 9 de estas variables son numéricas (una de ellas es la variable `id` que se puede eliminar), y se refieren a datos medidos sobre tumores de los pacientes. Estas 8 variables son las siguientes: `radius` (radio), `texture` (textura), `perimeter` (perímetro), `area` (área), `smoothness` (grado de suavidad), `compactness` (grado de compactibilidad), `symmetry` (grado de simetría), fractal dimensión (dimensión fractal). La otra variable es una variable categórica: `diagnosis_result`. Esta variable puede tomar dos valores M (maligno) o B (benigno), en función del tipo de tumor.

El objetivo de este ejercicio es, siguiendo las pautas del ejemplo desarrollado para el algoritmo KNN, aplicar dicho algoritmo a los datos y construir un modelo que tenga la variable `diagnosis_result` como variable objetivo, esto es, que determine los resultados de la diagnosis basándose en las 8 variables numéricas mencionadas anteriormente. Para ello resolver las siguientes cuestiones:

a) Cargar los datos y normalizarlos.
b) Crear los conjuntos de datos de entrenamiento y prueba. Dividir para esto los datos en dos partes, de manera que, de las 100 observaciones, los datos del 1 al 65 los tomamos como datos de entrenamiento y los datos del 66 al 100 como datos de prueba.
c) Construir el clasificador
d) Evaluar el modelo

****

EJERCICIO 4: k-means

Con el mismo fichero **Libro1.txt** que se ha utilizado en la explicación del ejemplo del algoritmo k-means:

a) Realizar un clustering de los datos utilizando el algoritmo k-means para *k* = 3, 4 y 5 grupos.
b) Realizar una representación gráfica de las agrupaciones obtenidas.
c) Calcular la suma de distancias de los datos a los centroides de los clusters en cada caso.
d) Añadir al conjunto de datos una columna adicional que se llame grupo y que contenga el número del cluster al que ha sido asignado cuando *k*= 5 grupos.

****

EJERCICIO 5: Algoritmo Jerárquico Aglomerativo

En el fichero **proteinas.txt** están los datos del año 1973 correspondientes al consumo de proteínas en 25 países europeos correspondientes a nueve grupos de alimentos. Los nombres de las variables son los siguientes:

1. Country: País
2. RdMeat: Carne roja
3. WhMeat: Carne Blanca
4. Eggs: Huevos
5. Milk: Leche
6. Fish: Pescado
7. Cereal: Cereales
8. Starch: Feculentos
9. Nuts: Frutos secos, y aceites
10. FrVeg: Frutas y verduras

a) Realizar una clasificación jerárquica de los países en base a su consumo de proteínas según las distintas fuentes de alimentación. Utilizar el método de Ward (`D.Ward`), y especificar que los casos se etiqueten con la variable `Country`.

   Nota: habría que trabajar con los datos tipificados para evitar que unas pocas variables dominen el análisis. Para tipificar un dataframe `proteinas.dat` y almacenar el resultado

en el dataframe `proteinas.tipif` podemos utilizar la instrucción `proteinas.tip=scale(proteinas.dat)`

b) Contestar, examinando el historial de agrupamientos, a las siguientes preguntas: ¿qué dos países se combinan primero? ¿En qué consiste la segunda etapa? ¿Cuándo es la primera vez que se forma un agrupamiento con más de dos países?
c) Examinar el dendograma: si queremos quedarnos con tres grupos, realizar la lista de los países que pertenecen a cada grupo. ¿Y con 4 grupos?
d) Realizar el análisis en componentes principales. Guardar las puntuaciones de los países según la primera componente principal. Ordenar los países por orden creciente de estas puntuaciones. ¿El orden obtenido parece guardar relación con los grupos obtenidos en el apartado anterior? ¿Cómo podemos explicar esta relación?

****