Práctica de la primera parte de Aprendizaje Estadístico 2024/2025

Master en Tecnologías de Análisis de Datos Masivos: Big Data

2024-10-07

Contents

Este es el enunciado de la primera práctica de las dos que componen el bloque de laboratorio de la Asignatura de Aprendizaje Estadístico, curso 2024/2025 del Máster Universitario en Tecnologías de Análisis de Datos Masivos: Big Data de la UMU.

En esta práctica se debe llevar a cabo un análisis exploratorio de los datos y demostrar las siguientes destrezas:

- Capacidad para manejar RStudio y sus rutinas de uso.
- Análisis de asociación entre variables.
- Capacidad para descubrir efectos ocultos en los datos.
- Capacidad para visualizar datos y resultados.
- Una mínima capacidad crítica para analizar resultados y extraer conclusiones.

Para entregar la realización de la práctica, se da dos opciones:

- Entregar un fichero .R junto el html o pdf generado por `Compile report`, función del menú de `File`.
- Entregar un fichero Markdown ejecutable, junto con el html o pdf correspondiente a la compilación de dicho Markdown.

Importante: Todas las respuestas deben estar comentadas.

Análisis a realizar correspondiente a la Parte 1 de la asignatura

Los datos de este práctica se utilizan en el libro "The Elements of Statistical Learning" y provienen de un estudio de Stamey et al. (1989) que examinó la correlación entre el nivel de antígeno específico de la próstata (PSA) y una serie de medidas clínicas, en 97 hombres que estaban a punto de recibir una prostatectomía radical. Las variables descritas en inglés son:

`lcavol`: log of cancer volume, measured in milliliters (cc). The area of cancer was measured from digitized images and multiplied by a thickness to produce a volume.

`lweight`: log. to the base e of the prostate weight, measured in grams.

`age`: The age of the patient, in years.

`lbph`:log to the base e of the amount of benign prostatic hyperplasia (BPH),

`svi`: seminal vesicle invasion, a 0/1 indicator of whether prostate cancer cells have invaded the seminal vesicle.

`lcp`: log to the base e of the capsular penetration, which represents the level of extension of cancer into the capsule.

`gleason`: Gleason score, a measure of the degree of aggressiveness of the tumor.

`pgg45`: Percentage of Gleason scores that are 4 or 5.

`lpsa`: log to the base e of prostate specific antigen (PSA), a concentration measured in ng/m.

Los datos forman parte del paquete ElemStatLearn, que se puede descargar desde:

https://cran.r-project.org/src/contrib/Archive/ElemStatLearn/

Otra opción es descargar los datos desde:

https://hastie.su.domains/ElemStatLearn/data.html

A continuación, se detallan las preguntas.

0. Carga los datos y elimina la variable TRAIN.

1. Exploración de los datos (0.25 puntos)

- ¿Cuántas variables hay?
- ¿De qué clase son?
- ¿Hay una variable que correspondiente al identificador de paciente? Si es así, elimínala.
- ¿Hay valores nulos en alguno de los ficheros?
- ¿Están estandarizadas las variables? En este punto del análisis, ¿es necesario normalizarlas?
- ¿Por qué crees que algunas variables están es escala logarítmica?

2. Análisis de variable categóricas (0.25 puntos)

Modifica la clase las variables svi, gleason y age a categóricas. Una vez realizado este paso, proporciona información sobre la distribución de sus valores.

3. Análisis de frecuencias (1 punto)

- ¿Qué porcentaje de pacientes con la puntuación de Gleason igual a 7, presenta índice igual `svi` igual a 0?
- ¿Qué porcentaje de pacientes con índice `svi` igual a 0 tiene la puntuación de Gleason igual a 7?
- Estas dos variables, ¿son independientes?

4. Regresión lineal simple. (1 punto)

Analiza si la variable `lpsa` depende linealmente de la variable `lcavol`. Al menos debe quedar reflejado:

- Interpretación del modelo lineal calculado.
- El plot de los datos junto a la recta de regresión.
- Intervalos de confianza para los coeficientes del modelo con una confianza de 0.95.
- Definición de RSE y su valor.
- Estudio de la eficacia del modelo.

- Interpretación. El modelo lineal calculado, ¿cómo lo interpretas? Concretamente, ¿cómo a través del modelo lineal llegas a otro que relacionan las variables cavol y psa (sin los log. neperianos)?

5. Regresión lineal múltiple. Correlación (1.5 puntos)

Mediante el uso de un plot del tipo `corrplot`, observa las posibles correlaciones entre todas las variables y coméntalas.

Con todas las variables menos `age`, `pgg45` y `gleason`, estudia el modelo de regresión lineal para predecir la variable `lpsa` en función de las demás. Analiza el `summary` y saca conclusiones. Estudia la eficacia del modelo.

¿A qué conclusiones llegas? ¿Podrías justificar la regulación?

6. Modelo de Ridge y Lasso (2 puntos)

Siguiendo con la predicción de la variable `lpsa` y las variables mencionadas en el apartado anterior, aplica los métodos de Ridge y Lasso a un conjunto de entrenamiento (training sample). Para cada método, representa en un plot las estimaciones de los parámetros del modelo en función del logaritmo de λ. ¿Cuál es el mejor λ estimado tras la validación cruzada? ¿Qué diferencias aprecias respecto al modelo lineal múltiple? Representa el MSE en función del logaritmo de lambda. Predice `lpsa` con el conjunto de testeo. Compara los resultados con el modelo lineal múltiple.

¿A qué conclusiones llegas? ¿Qué modelo te parece mejor? ¿Crees que los modelos lineales podrías ser utilizados?

7. LDA (1 punto)

Aplica LDA para clasificar la variable `svi` basándose en las variables `lcavol`, `lcp` y `lpsa`. ¿Es un buen modelo? Cuanta más información se proporcione, mejor.

8. Regresión logística (1 punto)

Aplica regresión logística para clasificar la variable `svi` basándose en las variables `lcavol`, `lcp` y `lpsa`. ¿Es un buen modelo? ¿Cómo lo interpretas? ¿Cuál es la probabilidad de `svi`=1 con `lcavol`=2.8269, `lcp`=1.843 y `lpsa`=3.285. Cuanta más información se proporcione, mejor.

9. PCA-PCR (2 puntos)

Análisis de componentes principales de las variables `lcavol, lweight, lbph, lcp, svi`. Al menos debe quedar reflejado:

- Un biplot y summary del modelo. Coméntalos.
- Justificar la desviación estándar del primer componente principal.
- Tras la proporción de varianza acumulada, ¿cuáles son las componentes principales que reflejan el 80% de la varianza total de los datos?

Además, aplica PCR para predecir lpsa teniendo en cuenta las variables `lcavol, lweight, lbph, lcp, svi`. ¿Qué conclusiones podrías sacar?¿Este modelo es mejor que el del apartado 5?