LISTA DE EJERCICIOS 2

Fecha de entrega: domingo 6 de febrero hasta la 1:00 pm

Todos los ejercicios son de la segunda edición del libro "An Introduction to Statistical Learning with Applications in R".

Escoja 4 de los 5 ejercicios para resolver. Cada uno tendrá un peso de 5 puntos.

Pregunta 1 (ISLR Sección 5.4 Ejercicio 2)

Derive la probabilidad de que una observación dada sea parte de una muestra bootstrap. Suponga que obtenemos una muestra bootstrap a partir de un conjunto de n observaciones.

(a) ¿Cuál es la probabilidad de que la primera observación del bootstrap no sea la j-ésima observación de la muestra original? Justifica tu respuesta.
(b) ¿Cuál es la probabilidad de que la segunda observación del bootstrap no sea la j-ésima observación de la muestra original?
(c) Argumenta que la probabilidad de que la j-ésima observación no esté en el inicio de muestra sea igual a $(1 - 1/n)^n$
(d) Cuando n = 5, ¿Cuál es la probabilidad de que la j-ésima observación esté en la muestra bootstrap?
(e) Cuando n = 100, ¿Cuál es la probabilidad de que la j-ésima observación esté en la muestra bootstrap?
(f) Cuando n = 10 000, ¿Cuál es la probabilidad de que la j-ésima observación esté en la muestra bootstrap?
(g) Cree una gráfica que muestre, para cada valor entero de n de 1 a 100 000, la probabilidad de que la j-ésima observación esté en la muestra de inicio. Comenta los resultados observados.
(h) Muestre numéricamente la probabilidad de que una la muestra bootstrap de tamaño n = 100 contiene la j-ésima observación, donde j = 4. Cree repetidamente muestras de bootstrap, y cada vez registramos si la cuarta observación está o no contenida en la muestra del bootstrap.

```
store=rep (NA , 10000)

for (i in 1:10000) {
store[i]=sum(sample (1:100 , rep =TRUE)==4) >0
}
mean(store)
```

Comente los resultados obtenidos.

Pregunta 2 (ISLR Sección 5.4 Ejercicio 3)

Ahora revisemos la validación cruzada k-fold

(a) Expliqué cómo se implementa la validación cruzada k-fold
(b) ¿Cuáles son las ventajas y desventajas de la validación cruzada k-fold en relación con
   i. El enfoque de conjunto de validación?
   ii. El enfoque de validación cruzada dejando uno afuera (LOOCV)?

Pregunta 3 (ISLR Sección 5.4 Ejercicio 8)

Ahora realizaremos una validación cruzada en un conjunto de datos simulados

(a) Genere un conjunto de datos simulados de la siguiente manera:

```
set.seed (1)
x <- rnorm (100)
y <- x - 2 * x^2 + rnorm (100)
```

En este conjunto de datos, ¿cuál es n y cuál es p? Escribe el modelo utilizado para generar los datos en forma de ecuación.

(b) Cree un diagrama de dispersión de X contra Y. Comenta lo que encuentres.
(c) Establezca una semilla aleatoria y luego calcule los errores LOOCV que resultado de ajustar los siguientes cuatro modelos usando mínimos cuadrados:

i. $Y = \beta 0 + \beta 1X + \epsilon$
ii. $Y = \beta 0 + \beta 1X + \beta 2X^2 + \epsilon$
iii. $Y = \beta 0 + \beta 1X + \beta 2X^2 + \beta 3X^3 + \epsilon$
iv. $Y = \beta 0 + \beta 1X + \beta 2X^2 + \beta 3X^3 + \beta 4X^4 + \epsilon$

Tenga en cuenta que puede resultarle útil utilizar la función data.frame() para crear un solo conjunto de datos que contenga X e Y
(d) Repita (c) usando otra semilla aleatoria e informe sus resultados. ¿Son tus resultados iguales a los que obtuviste en (c)? ¿Por qué?
(e) ¿Cuál de los modelos en (c) tuvo el error LOOCV más pequeño? ¿Es esto es lo que esperabas? Explica tu respuesta

Pregunta 4 (ISLR Sección 6.6 Ejercicio 9)

En este ejercicio, predeciremos el número de solicitudes recibidas utilizando las variables del dataset **College.** Se ha subido el data set a PAIDEIA.

(a) Divida el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.
(b) Ajuste un modelo lineal utilizando mínimos cuadrados en el conjunto de entrenamiento e informe el error de prueba obtenido.
(c) Ajuste un modelo rigde regression en el conjunto de entrenamiento, con λ elegido por cross-validation. Informe el error de prueba obtenido.
(d) Ajuste un modelo utilizando lasso en el conjunto de entrenamiento, con λ elegido por cross-validation. Informe el error de prueba obtenido.

Pregunta 5 (ISLR Sección 6.6 Ejercicio 11)

Ahora intentaremos predecir la tasa de criminalidad per cápita en el conjunto de datos de **Boston**. Usando **library(ISLR2)** pueden acceder a la base Boston.

1. Prueba los métodos de regresión explorados: ridge y lasso. Presente y discuta los resultados
2. Proponga un modelo (o conjunto de modelos) que parezca funcionar bien en este conjunto de datos y justifique su respuesta. Asegúrate de que estás evaluando el rendimiento del modelo utilizando el error del conjunto de validación, la validación cruzada o alguna otra alternativa razonable, en lugar del error de entrenamiento
3. ¿Su modelo elegido involucra todas las variables de la base de datos? ¿Por qué o por qué no?