Hoja de Ejercicios 2: Aprendizaje Estadístico y Tomada de Decisiones

Fecha de Entrega: Preguntas 1-6 26/10/2017 a las 20:00

Preguntas 7-10 02/11/2017 a las 20:00

Trabajan en los mismos grupos de cuatro personas. Es suficiente entregar la solución una vez. No hace falta solucionar todo para sacar una buena nota.

Problema 1.
Swirl: descargue el package swirl en R: $install.packages('swirl')$. Cargue swirl $library(swirl)$ y utilízalo para aprender las funciones básicas de R. El swirl es un package de auto-aprendizaje en R, por lo tanto este ejercicio sólo es recomendable para los que nunca han programado en R. Dale $swirl()$ en R, y luego seleccione la opción 1: *R Programming: The basics of programming in R.* Haz los temas 1-13 (fácil). **No se tiene que entregar nada.**

Problema 2.
Haz el Tutorial sobre bucles en R en DataCamp: `https://www.datacamp.com/community/tutorials/tutorial-on-loops-in-r#gs.vWxo6aQ`.(fácil) **No se tiene que entregar nada.**

Problema 3.
Ejercicio 8 del Capítulo 2 del libro.

Problema 4.
Ejercicio 9 del Capítulo 2 del libro.

Problema 5.
Ejercicio 10 del Capítulo 2 del libro. Entregue el código y los resultados.

Problema 6.
Vamos a utilizar un conjunto de datos de 100 pacientes para implementar el algoritmo KNN. El conjunto de datos se ha elaborado teniendo en cuenta los resultados obtenidos generalmente por el examen rectal digital. Utilice los datos Pro.csv. Proporcione el código y los resultados de los siguientes apartados:

(i) Cargue los datos y analícelos. Haz una tabla con los pacientes que tuvieron un resultado *Malignant (M) y Benign (B).*

(ii) Separe la muestra entre datos de entrenamiento y de prueba para predecir el resultado del tratamiento.

(iii) Ajuste un modelo KNN y evalúe su desempeño.

(iv) Repite el apartado (iii) para descubrir cuál elección de k en el KNN genera el menor error de prueba (con el $k \in [1; 12]$).

Problema 7.
Cargue los datos del Problema 6 otra vez y obtén el k óptimo mediante una Validación Cruzada 10-Veces: descargue la *package KODAMA* y utilice la función "KNN.CV()", mire las páginas 7-8 de la documentación de la *package KODAMA*. (fácil)

Problema 8.
Cargue las *libraries ISLR* y *boot*. Cargue la base de datos *Wage*. Es una encuesta sobre sueldos en la región central atlántica de EE. UU. en 2009. Haz una regresión polinomial para predecir el salario (*wage*) utilizando sólo la variable experiencia (*age*). Utilice la validación cruzada de 5, 10 y n Veces ($LOOCV$) para encontrar el orden óptimo para el polinomio (considere $d \in [1; 10]$). Haz un gráfico del error de validación cruzada para cada orden del polinomio, para cada una de las 3 validaciones cruzadas. Proporcione el código y los resultados.

Problema 9.
Cargue las *libraries ISLR* y *boot*. Cargue la base de datos *Boston*. Haz una regresión polinomial para predecir la concentración de óxidos de nitrógeno en partes por 10 millones (*nox*) utilizando sólo la media ponderada de las distancias a cinco centros de empleo de Boston (*dis*). Utilice la validación cruzada de 5, 10 y n Veces ($LOOCV$) para encontrar el orden óptimo para el polinomio (considere $d \in [1; 10]$). Haz un gráfico del error de validación cruzada para cada orden del polinomio, para cada una de las 3 validaciones cruzadas. Proporcione el código y los resultados.

Problema 10.
Cargue la package "Lock5Data" y los datos "CommuteAtlanta".

(a) Dé una estimación de la media poblacional de "Distance". Llámela de $\hat{\mu}$.

(b) Dé una estimación del error estándar de $\hat{\mu}$. Interprete este resultado.

(c) Ahora estime el error estándar de $\hat{\mu}$ usando el Bootstrap con $B = 100,000$ muestras de boostrap. Compare los resultados con los del apartado (b).

(d) A partir de la estimación de Bootstrap del apartado (c), calcule un Intervalo de Confianza del 95% para la media poblacional de "Distance". Utilice el método del percentil. Compare los resultados con el intervalo de confianza obtenido utilizando la distribución Normal.

(e) Dé una estimación de la varianza poblacional de "Distance". Llámela de $\hat{\sigma}^2$.

(f) Estime el error estándar de $\hat{\sigma}^2$ usando el Bootstrap con $B = 100,000$ muestras de boostrap.

(g) Estime el percentil 25% poblacional de "Distance". Llámelo de $\hat{\mu}_{0.25}$.

(h) Estime el error estándar de $\hat{\mu}_{0.25}$ usando el Bootstrap con $B = 100,000$ muestras de boostrap. Interprete los resultados.

(i) A partir de la estimación de Bootstrap del apartado (h), calcule un Intervalo de Confianza del 95% para el percentil 25% poblacional de "Distance". Utilice el método del percentil.