

Hoja de Ejercicios 1: Aprendizaje Estadístico y Tomada de Decisiones

Fecha de Entrega: 10/10/2017 a las 19:00

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

Preguntas de revisión. Respóndelas lo más corto y preciso posible.

1. Busca un ejemplo de regresión donde la variable dependiente es cuantitativa. Elabora. ¿Qué es Y? ¿Qué es X? ¿Qué es $f(X)$? ¿Cómo estimarías $f(X)$ por K nearest neighbours? ¿Cómo se interpretaría esta estimación $f(X)$ en este caso? ¿Cómo lo estimarías con un modelo paramétrico? ¿Cómo se interpretaría esta estimación $\hat{f}(X)$ en este caso?
2. ¿Puedes construir datos de Y y X tal que el error irreducible = 0 y otros tal que es muy alto? ¿Cómo lo representarías en un gráfico? ¿Qué implicaría para la estimación?
3. Busca un ejemplo de un problema de clasificación donde la variable dependiente toma 2 valores. Elabora. ¿Qué es Y? ¿Qué es X? ¿Qué es $p(X)$? ¿Cómo estimarías $p(X)$ y $C(X)$ en este caso con el método de K nearest neighbours?
4. Busca un ejemplo de un problema de clasificación donde la variable dependiente toma más de 2 valores. Elabora.
5. Un amigo quiere visitarte en Noviembre en Palma. Pregunta por las fechas óptimas y te dice que le gusta el sol. ¿En qué sentido puede verse esto como un problema de clasificación? Elabora. ¿Cómo se podría estimar el problema?
6. Encuentra un ejemplo real de clasificación (simple) donde el error de Bayes es máximo. Y uno donde es mínimo. ¿Qué implica para la estimación?
7. En un problema de clasificación con 3 características describe como son las frontera de decisión de Bayes resultando de una estimación de K nearest neighbours. ¿Para qué sirven? ¿Cómo dependen del K?

Preguntas de datos

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 1-8 (fácil).

Problema 2.

Ejercicio 8 del Capítulo 2 del libro.

Problema 3.

Ejercicio 9 del Capítulo 2 del libro.

Problema 4.

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

Problema 5.

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`. Entregue 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 los apartados (b) y (c) para descubrir cuál elección de k en el KNN genera el menor error de prueba.