APRENDIZAJE AUTOMÁTICO
Trabajo-2

Valoración: 18 puntos

Fecha de entrega: 26 Abril

Cuestionario (8 puntos): estará disponible más adelante

Ejercicios: 10 puntos (Justificar las respuestas en todos los casos)

Ejercicio.-1 (4 puntos): Usar la base de datos Weekly (ISLR) . Este conjunto de datos presenta predicciones semanales (1089) del mercado de valores durante 21 años (1990-2010).

1. Analizar la conducta de los datos usando resúmenes numéricos y gráficos de los mismos. ¿Se observa algún patrón de interés? En caso afirmativo dar una posible interpretación del mismo.

2. Usar la base de datos completa para ajustar un modelo de Regresión Logística usando Direction como variable respuesta y las variables Volume y Lag-1 a Lag-5 como predictores. Usar la función summary() para mostrar los resultados. ¿Alguno de los predictores es estadísticamente significativo? En caso afirmativo, ¿cuál? Justificar la respuesta

3. Calcular la matriz de confusión y el porcentaje total de predicciones correctas. Explicar lo que la matriz de confusión nos dice acerca de los errores cometidos por regresión logística. Justificar las respuesta

4. Ahora ajustar un modelo de regresión logística a los datos entre 1990 y 2008 usando Lag2 como el único predictor. Calcular la matriz de confusión y la fracción global de predicciones correctas para el periodo 2009 y 2010. Justificar las respuesta

5. Repetir (4) usando LDA, QDA y KNN con K=1. ¿Cuál de estos métodos parece dar los mejores resultados? Justificar las respuesta

6. Experimente con diferentes combinaciones de predictores, incluyendo transformaciones de las variables e interacciones entre ellas, en cada uno de los métodos (RLG, LDA, QDA, KNN) (si se desea, pueden usarse técnicas de selección de variables) . Muestre las variables, método y matriz de confusión que da los mejores resultados sobre los datos de test (2009-2010). . Justificar las respuestas adecuadamente.

7. Repetir el punto anterior usando un modelo de ajuste de validación cruzada con 10 particiones. Comparar con los resultados obtenidos en el punto anterior. Justificar las respuestas adecuadamente.

Ejercicio.-2 (3 puntos)
En este ejercicio desarrollaremos un modelo para predecir si un coche tiene un consumo de carburante alto o bajo usando la base de datos Auto.

a) Crear una variable binaria, mpg01, que será igual 1 si mpg contiene un valor por encima de la mediana , y 0 si mpg contiene un valor por debajo de la mediana. La mediana se puede calcular usando la función median() . (Nota: puede resultar útil usar la función data.frame() para unir en un mismo conjunto de datos mpg01 y las otras variables de Auto)

b) Explorar los datos gráficamente para investigar la asociación entre mpg01 y las otras características. ¿Cuáles de las otras características parece más útil para predecir mpg01? El uso de Scatterplots y boxplots (ver el libro para recordar concepto) puede resultar útil para contestar la cuestión. Justificar la respuesta.

c) Definir un conjunto de validación dividiendo los datos en un conjunto de entrenamiento (70%) y otro de test (30%):

    a. Ajustar un modelo LDA a los datos de entrenamiento y predecir mpg01 usando las variables que en (b) resultaron más asociadas con mpg01. ¿Cuál es el error de test en el modelo? Justificar la respuesta

    b. Ajustar un modelo QDA a los datos de entrenamiento y predecir mpg01 usando las variables que en (b) resultaron más asociadas con mpg01. ¿Cuál es el error de test en el modelo? Justificar la respuesta

    c. Ajustar un modelo de regresión Logística a los datos de entrenamiento y predecir mpg01 usando las variables que en (b) resultaron más asociadas con mpg01. ¿Cuál es el error de test en el modelo? Justificar la respuesta

    d. Ajustar un modelo KNN a los datos de entrenamiento y predecir mpg01 usando solamente las variables que en (b) resultaron más asociadas con mpg01. ¿Cuál es el error de test en el modelo? ¿Cuál es el valor de K que mejor ajusta los datos? Justificar la respuesta

d) Repetir los experimentos a-d del punto anterior pero usando Validación Cruzada de 5-particiones para evaluar el error de test. Comparar con los resultados obtenidos en el punto anterior.

Ejercicio.-3 (3 puntos)

Usar la base de datos Boston para ajustar un modelo que prediga si dado un suburbio este tiene una tasa de criminalidad por encima o por debajo de la mediana. Comparar los modelos encontrados por RLG, LDA y QDA.

1. Encontrar en cada caso el subconjunto óptimo de variables predictoras usando Validación Cruzada.
    a. Escribir una función para el cálculo del error del test.
2. Calcular los modelos y valorar sus resultados

Los BONUS solo se tendrán en cuenta si se ha obtenido al menos el 50% de los puntos en los ejercicios obligatorios. (NO HAY BONUS)

Informe a presentar

Para este trabajo como para los demás proyectos debe presentar un informe escrito con sus valoraciones y decisiones adoptadas en cada uno de los apartados de la implementación. Incluir los gráficos generados y el código R que haya desarrollado para resolver los ejercicios. También deberá incluirse una valoración sobre la calidad de los resultados encontrados. (hacerlo en pdf, MS Word o en texto plano)

Normas de la entrega de Trabajos: EL INCUMPLMIENTO DE ESTAS NORMAS SIGNIFICA PERDIDA DIRECTA DE 1 PUNTO CADA VEZ QUE SE DETECTE UN INCUMPLIMIENTO.

1. Cada contestación del cuestionario siempre incluirá la correspondiente pregunta.
2. El código se debe estructurar en un único script R con distintas funciones o apartados, uno por cada apartado de la práctica.
3. El path que se use en la lectura de imágenes o cualquier otro fichero de datos debe ser siempre “imagenes/nombre_fichero”
4. Todos los resultados numéricos serán mostrados por pantalla. No escribir nada en el disco.
5. La práctica deberá poder ser ejecutada de principio a fin sin necesidad de ninguna selección de opciones. Para ello fijar al comienzo los parámetros por defecto que se consideren óptimos.
6. La práctica debe de ejecutarse de principio a fin sin errores.
7. El código debe estar obligatoriamente comentado explicando lo que realizan los distintos apartados y/o bloques.
8. Poner puntos de parada para mostrar imágenes o datos por consola.
9. Todos los ficheros a entregar juntos se podrán dentro de un fichero zip, cuyo nombre debe ser Apellido1_P[1-3].zip.
10. ENTREGAR SOLO EL CODIGO FUENTE, NUNCA LOS DATOS.

Forma de entrega: Subir el zip al Tablón docente de CCIA.