APRENDIZAJE AUTOMATICO

Cuestionario-T1: 6 puntos

Fecha entrega: 29 Abril

Justificar la contestación en todos los apartados
Incluir los enunciados en la contestación

1. (**0.5 puntos**) Decir cuáles de los siguientes casos son un problema de regresión o de clasificación e indicar si estaremos más interesado en inferencia o en predicción. Identificar también los valores de n (tamaño muestra) y p (número de predictores):
   a) Recopilamos un conjunto de datos de las 500 empresas españolas más grandes. Por cada compañía recogemos: beneficio anual, número de empleados, tipo de industria y sueldo del director. Estamos interesado en comprender que factores afectan al sueldo del director.
   b) Estamos considerando lanzar un nuevo producto y deseamos conocer si será un éxito o un fracaso. Para ello recogemos datos de 20 productos semejantes ya existentes en el mercado. Para cada producto medimos: a) si fue un éxito o un fracaso; b) precio del producto; c) presupuesto de marketing; d) precio de oferta inicial y otras diez variables más.
   c) Estamos interesados en predecir el % de variación del euro respecto de los porcentajes de variación semanales de los mercados europeos. Para ello recogemos datos semanales de todo el 2012. En cada semana medimos el % de cambio del euro, el % de cambio de la bolsa Alemana, el % de cambio de la Bolsa Inglesa y el % de cambio de la bolsa Francesa.

2. (**1 punto**) Identificar dos aplicaciones empresariales (no comentadas en clase) en las que considere que las técnicas de Aprendizaje Automático serán útiles. Describir brevemente cada una de ellas, el interés de la misma y el problema que se resuelve. Identificar algunas de las variables que considere más importantes al problema. Describir un caso en que regresión será la técnica a aplicar (decir además si el problema es más de inferencia o de predicción) y otro de clasificación.

3. (**0.5 puntos**) Describir las diferencias entre las aproximaciones supervisadas paramétricas y las no-paramétricas. ¿Cuáles son las ventajas de la aproximación paramétrica en regresión y clasificación? ¿Cuáles las desventajas? Justificar la respuesta.

4. (**0.5 puntos**) Si tenemos un problema de clasificación con dos variables predictoras y nos muestran las fronteras de decisión de un clasificador kNN para distintos valores de k ¿Cómo podemos saber si la frontera de decisión comienza a estar sobre-ajustada? Justificar la respuesta.

5. (**2 puntos**) Suponga que tenemos un conjunto de datos con 5 variables predictoras, X_1, X_2, X_3, X_4, X_5, de las cuales X_1 y X_2 son cuantitativas, X_3 es cualitativa con dos valores (0=hombre, 1=mujer), X_4 representa la interacción entre X_1 y X_2, y X_5 representa la interacción entre X_1, y X_3. La variable de salida representa el valor del salario de hombres y mujeres. Hemos ajustado un modelo por mínimo cuadrados y se han obtenido los siguientes coeficientes $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

   a) ¿Cuáles de las siguientes contestaciones es correcta y por qué?
      i) Para valores fijos de X_1 y X_2 los hombres ganan más en promedio que las mujeres.
      ii) Para valores fijos de X_1 y X_2 las mujeres ganan más en promedio que los hombres
      iii) Para valores fijos de X_1 y X_2 los hombres ganan más en promedio que las mujeres con tal que X_1 sea suficientemente grande.
   b) Predecir el salario de una mujer con X_1= 4.0 y X_2=110
   c) Dado que el coeficiente de X_4 es pequeño existe poca evidencia de un efecto de interacción entre X_1 y X_2, ¿Verdadero o Falso? Justificar la respuesta

6. (**1.5 puntos**) Tenemos un conjunto de datos de 100 observaciones con una única variable predictor y una respuesta cuantitativa. Ajustamos a dichos datos un modelo de regresión lineal $Y = \beta_0 + \beta_1 X + \varepsilon$ y un modelo de regresión cúbico $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

   a) Supongamos que la verdadera relación entre X e Y es lineal, es decir $Y = \beta_0 + \beta_1 X + \varepsilon$. Considerar la suma de los residuos de los datos de entrenamiento (RSS) tanto para el modelo lineal como para el modelo cúbico. ¿Deberíamos esperar que en general un valor fuera menor que el otro, que fueran iguales, o no hay suficiente información para establecer una opinión por adelantado?
   b) Contestar lo mismo del punto anterior pero considerando la sumas RSS de los datos de test.
   c) Supongamos que la verdadera relación entre X e Y es no lineal, pero no conocemos como de lejos está de ser lineal. Consideremos las sumas RSS de entrenamiento para el modelo lineal y el cúbico ¿Deberíamos esperar que en general un valor fuera menor que el otro, que fueran iguales, o no hay suficiente información para establecer nada por adelantado? Justificar la contestación.
   d) Contestar lo mismo del punto anterior pero considerando la sumas RSS de los datos de test.

Los BONUS solo se tendrán en cuanta si se alcanzado al menos el 50% de los puntos en las preguntas obligatorias.

BONUS.1 (**1.5 puntos**) (Justificar la contestación en todos los apartados)

1. Calcular las ecuaciones de ajuste de los coeficientes de regresión lineal simple en el caso de datos normalizados

2. Deducir las expresiones para β_0 y β_1 con datos no normalizados a partir de los valores estimados con datos normalizados en regresión lineal simple.
3. ¿Bajo qué condiciones el coeficiente β_1 de la recta de regresión de Y sobre X será igual al de la recta de regresión de X sobre Y?

BONUS.2 (**2.5 puntos**) (Justificar la contestación en todos los apartados)

Usar la base de datos Carseats (ISLR library). (Justificar las respuestas en todos los casos)

1. Ajustar un modelo de regresión múltiple para predecir Sales usando Price, Urban, y US.
2. Dar una interpretación de cada coeficiente del modelo.
3. Escribir el modelo en forma de ecuación, tratando las variables cualitativas de forma correcta.
4. ¿Para qué predictores se puede rechazar la hipótesis nula $H_0 : \beta_j = 0$?
5. De acuerdo a la respuesta a la cuestión anterior, ajustar un modelo más pequeño que solo use la variables par a las cuales hay evidencia de asociación con la variable salida
6. ¿Cómo de bien se ajustan los datos a los modelos de los puntos (1) y (5) ?
7. Usando el modelo de (5) obtener los intervalos de confianza al 95% para los coeficientes.
8. ¿Existe evidencia de puntos outliers o puntos de alta influencia en el modelo del punto (5)?

BONUS.3 (**1.5 puntos**) (Justificar la contestación en todos los apartados)

Usar la base de datos cars (ISLR library) .

1. Usar lm para ajustar un modelo de regresión lineal múltiple con mpg como variable respuesta..
2. Evaluar si hay algún problema reseñable en el ajuste del modelo lineal. Usar gráficos de ajuste.
3. Usar los símbolos * y : para ajustar modelos de regresión lineal con efectos de interacción. Evaluar los resultados.
4. Aplicar distintas transformaciones a las variables, tales como log(X), $\sqrt{X}$, X^2. Valorar los resultados

BONUS-1 (Justificar las respuestas en todos los casos)

1.- Considerar las ecuaciones de las estimaciones de los parámetros del modelo de regresión lineal simple:

a) ¿Qué transformación deberían sufrir los datos para que el modelo de regresión ajustado tuviera término independiente igual a cero (β_0=0)?
b) ¿Cómo se afecta el coeficiente β_1 por dicha transformación?
c) ¿Puede extraerse alguna propiedad relevante del modelo a partir de la transformación?
d) ¿Bajo qué condiciones el coeficiente de regresión de Y sobre X es igual al coeficiente de regresión de X sobre Y?
e) Generar una muestra de tamaño 100 en donde la estimación del coeficiente de regresión de Y sobre X es diferente al de regresión de X sobre Y. Mostrar un gráfico con dicha muestra.
f) Generar una muestra de tamaño 100 en donde la estimación del coeficiente de regresión de Y sobre X es igual al de regresión de X sobre Y. Mostrar un gráfico con dicha muestra.