TÉCNICAS DE INTELIGENCIA ARTIFICIAL

TEMA 3 - CLASIFICACIÓN

EJERCICIOS

CONCEPTUALES

EJERCICIO 1

Usando un poco de álgebra, demostrar que la siguiente expresión

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

es equivalente a esta otra

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

En otras palabras, la función logística y la representación logit son equivalentes para un modelo de regresión logística.

EJERCICIO 2

Clasificar una observación en la clase para la cual

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)}$$

es más alta es equivalente a clasificar una observación en la clase para la cual

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

es más alta. Demostrar que esto es realmente así. En otras palabras, si asumimos que la observaciones en la clase k vienen de una distribución $N(\mu_k, \sigma^2)$, el clasificador de Baues asigna una observación a la clase para la cual la función discriminante se maximiza.

EJERCICIO 3

Analizamos ahora las diferencias entre ALD y ACD.

1. Si el umbral de decisión de Bayes es lineal, ¿esperamos que el ALD o el ACD rinda mejor en el conjunto de datos de entrenamiento? ¿Y en el de test?
2. Si el umbral de decisión de Bayes es no-lineal, ¿esperamos que el ALD o el ACD rinda mejor en el conjunto de datos de entrenamiento? ¿Y en el de test?

3. En general, a medida que aumenta el tamaño muestral, n, ¿esperamos que la precisión en test del ACD respecto al del ALD mejore, empeore, o no varíe? ¿Por qué?

4. Verdadero o falso: incluso si el umbral de decisión de Bayes para un cierto problema fuera lineal, probablemente obtendremos una tasa de error de test más baja usando el ACD en lugar del ALD porque el ACD es lo suficientemente flexible como para modelar un umbral de decisión lineal. Justifica la respuesta.

EJERCICIO 4

Supongamos que recopilamos un conjunto de datos de un grupo de estudiantes en una clase de inteligencia artificial. Registramos las siguientes variables: $X_1 =$ horas de estudio $X_2 =$ NM (nota media), e $Y =$ obtener calificación A. Ajustamos un modelo de regresión logística, y obtenemos $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0{,}05$, y $\hat{\beta}_2 = 1$.

1. Estimar la probabilidad de que un estudiante que estudia 40 h y tiene una nota media de 3.5 obtenga una A.

2. ¿Cuántas horas tiene que estudiar el mismo estudiante del apartado anterior para tener un 50 % de probabilidad para obtener una A?

EJERCICIO 5

Dividimos una base de datos en dos particiones del mismo tamaño, entrenamiento y test. Después, probamos dos tipos de clasificadores. Primero, usamos regresión logística y obtenemos una tasa de error de entrenamiento del 20 %, y una tasa de error de test del 30 %. Luego usamos KNN con $K = 1$ y obtenemos una tasa de error media (entre entrenamiento y test) del 18 %. En base a estos resultados, qué método deberíamos usar para clasificar una nueva observación? ¿Por qué?

APLICADOS

EJERCICIO 6

Este ejercicio está relacionado con el uso de la base de datos Weekly que es parecida a Smarket, pero contiene 1089 retornos semanales de 21 años, desde principios de 1990 hasta el final de 2010.

1. Producir resúmenes numéricos y gráficos de la base de datos Weekly. ¿Existe algún patrón?
2. Utiliza toda la base de datos para ajustar un modelo de regresión logística para predecir Direction en base a las cinco variables lag y Volume. ¿Es alguno de los predictores estadísticamente significativo? En caso afirmativo, identifícalos.
3. Calcula la matriz de confusión y el porcentaje de predicciones correctas. Examina la matriz de confusión y explica lo que ésta indica sobre los tipos de errores que el modelo de regresión logística comete.
4. Ahora ajusta un modelo de regresión logística usando como datos de entrenamiento las observaciones desde 1990 hasta 2008, y utiliza Lag2 como único predictor. Calcula la matriz de confusión y el porcentaje de predicciones correctas para los datos de test (observaciones desde 2009 a 2010).
5. Repite el apartado 4 usando el ALD.
6. Repite el apartado 4 usando el ACD.
7. Repite el apartado 4 usando KNN con $K = 1$.
8. ¿Cuál de los métodos parece obtener mejores resultados?
9. Experimenta con diferentes combinaciones de los predictores (posibles transformaciones o interacciones) para cada uno de los métodos. Reporta las variables, método, y la matriz de confusión que parecen obtener los mejores resultados en los datos de test. Deberías de experimentar con diferentes valores de K para el clasificador KNN.

EJERCICIO 7

En este ejercicio desarrollarás un modelo para predecir si un determinado coche tiene alta o baja autonomía usando la base de datos Auto:

1. Crear una variable binaria, mpg01, que contenga un 1 si mpg es superior al valor mediano, y un 0 si mpg es menor que la mediana.
2. Utiliza boxplots para analizar la asociación entre mpg01 y las otras características. ¿Qué predictores parecen ser más útiles a la hora de predecir mpg01?
3. Divide los datos en conjunto de entrenamiento (70 %) y de test (30 %). Para cada valor posible de la variable year cuantifica el número de coches diferentes que hay y asigna en orden de aparición el 70 % a entrenamiento y el resto a test.

4. Utiliza el ALD en los datos de entrenamiento para predecir mpg01 usando las variables que parecían estar más asociadas con mpg01 en el apartado 2. ¿Cuál es el error de test del modelo?

5. Utiliza el ACD en los datos de entrenamiento para predecir mpg01 usando las variables que parecían estar más asociadas con mpg01 en el apartado 2. ¿Cuál es el error de test del modelo?

6. Utiliza la regresión logística en los datos de entrenamiento para predecir mpg01 usando las variables que parecían estar más asociadas con mpg01 en el apartado 2. ¿Cuál es el error de test del modelo?

7. Utiliza KNN en los datos de entrenamiento, con diferentes valores de K, para predecir mpg01. Utiliza solo las variables que parecían estar más asociadas con mpg01 en el apartado 2. ¿Cuál es el error de test del modelo? ¿Qué valor de K parece obtener mejores resultados en test?