UNIVERSIDADE
DE VIGO

Departamento de Estatística
e Investigación Operativa

Minería de Datos, 2008-2009
Un Problema de Clasificación en R

Planteamiento y Análisis de un Problema de Clasificación

Generación de los datos

1. Cargar el paquete **MASS** (instalarlo antes si no estuviese incluido en la instalación)**.**

2. Crear dos clases $\mathcal{C}_1$, $\mathcal{C}_2$ cada una con 200 puntos $\mathbf{x} \in \mathbb{R}^2$ aleatorios de tal forma que las densidades condicionadas $\mathsf{p}(\mathbf{x}|\mathcal{C}_j)$, $j = 1, 2$ sean una mixtura de gausianas, es decir:

$$\mathsf{X}|\mathcal{C}_j \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_{j,1}, \boldsymbol{\Sigma}_{j,1}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_{j,2}, \boldsymbol{\Sigma}_{j,2}), j = 1, 2$$

   siendo:

$$\boldsymbol{\mu}_{1,1} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \boldsymbol{\mu}_{1,2} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \boldsymbol{\mu}_{2,1} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \boldsymbol{\mu}_{2,2} = \begin{bmatrix} 5.5 \\ 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{j.i} = \sigma_{j,i} \cdot \mathbf{I}_{2\times 2} \text{ con: } \begin{cases} \sigma_{1,1} = \sigma_{2,2} = 0.75 \\ \sigma_{2,1} = \sigma_{1,2} = 0.25 \end{cases}$$

   Para ello, utilizar la función **mvrnorm** del paquete **MASS** para generar vectores aleatorios gausianos.

3. Utilizar la codificación $\mathsf{Y} \in \{-1, 1\}$ para las clases. Es decir, las clases son: $\mathcal{C}_1 = \{\mathbf{x} : \mathsf{Y}(\mathbf{x}) = -1\}$ y $\mathcal{C}_2 = \{\mathbf{x} : \mathsf{Y}(\mathbf{x}) = +1\}$.

   Otra notación equivalente para las clases será su etiqueta: $\mathsf{Y} = -1$ (para la clase $\mathcal{C}_1$) e $\mathsf{Y} = +1$ (para la clase $\mathcal{C}_2$).

4. Dividir aleatoriamente el conjunto total de datos en dos muestras de tamaño 200: una para entrenamiento y otra para test.

5. Graficar en $\mathbb{R}^2$ el conjunto de datos de entrenamiento utilizando colores diferentes para las clases.

Regla de Bayes. Frontera y Estimación del Error

1. Cargar el paquete **mvtnorm**.

2. **Probabilidades a posteriori de las clases**. Calcular las probabilidades a posteriori de las clases utilizando la función **dmvnorm** de dicho paquete, mediante el teorema de Bayes:

$$\mathsf{p}(\mathcal{C}_j|\mathbf{x}) = \frac{\mathsf{p}(\mathbf{x}|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j)}{\mathsf{p}(\mathbf{x})}$$

   donde supondremos que las clases tienen la misma probabilidad de ocurrir.

3. **Error de Bayes**. Obtener las predicciones de las clases según la regla de Bayes:

$$\mathsf{g}(\mathbf{x}) = \begin{cases} +1 & \text{si } \mathsf{p}(\mathcal{C}_1|\mathbf{x}) < \mathsf{p}(\mathcal{C}_2|\mathbf{x}), \text{ es decir si } \mathsf{p}(\mathcal{C}_1|\mathbf{x}) < 0.5 \\ -1 & \text{si } \mathsf{p}(\mathcal{C}_1|\mathbf{x}) \geq \mathsf{p}(\mathcal{C}_2|\mathbf{x}), \text{ es decir, si } \mathsf{p}(\mathcal{C}_1|\mathbf{x}) \geq 0.5 \end{cases}$$

   y la tasa de error de clasificación que dicha regla comete en las muestras de entrenamiento y de test.

4. **Frontera de Bayes**.

   (a) Generar una rejilla de datos en $[0, 8] \times [0, 6]$ a intervalos de longitud $\Delta = 0.1$ en ambas coordenadas.

   (b) Calcular las probabilidades a posteriori sobre los puntos de dicha rejilla de igual manera que en el punto 2 anterior (basta calcular las de la clase $\mathcal{C}_1$).

   (c) Utilizar la función **contour** para graficar la frontera de bayes utilizando esas probabilidades. (No cerrar el gráfico pues sobre él iremos graficando fronteras de otras técnicas).

5. **Gráfico de la Probabilidad a Posteriori**. Cargar el paquete **TeachingDemos** y en una nueva ventana de gráficos, graficar en 3D las probabilidades a posteriori obtenidas en el punto 4b anterior, utilizando la función **persp**.

   Posteriormente, utilizar la función **rotate.persp** para rotar el gráfico.

   Entender la forma de la frontera obtenida en el punto 4c a la vista de este gráfico.

Análisis Discriminante Lineal

1. Cargar el paquete **MASS** si no está cargado.

2. Resolver el problema de clasificación mediante un discriminante lineal utilizando la función **lda** utilizando únicamente la muestra de entrenamiento.

3. Mediante la función **plot**, realizar un histograma de la variable discriminante (proyección) para ver la separación de las clases.

4. **Tasa de Error del Discriminante Lineal**. Obtener las predicciones de las clases según la regla discriminante y la tasa de error de clasificación que dicha regla comete en las muestras de entrenamiento y de test.

   Comparar ambas tasas con las obtenidas por la Regla de Bayes.

5. **Frontera del Discriminante Lineal**.

   (a) Obtener las probabilidades a posteriori que la regla discriminante lineal estima sobre cada punto de la rejilla del punto 4a anterior.

   (b) Siguiendo el procedimiento 4c seguido para la Regla de Bayes, graficar la frontera producida por el discriminante lineal sobre el mismo gráfico en el que se obtuvo la frontera de Bayes.

6. **Gráfico de la Estimación de la Probabilidad a Posteriori**. En una nueva ventana de gráficos, graficar en 3D la estimación de la probabilidad a posteriori producida por el discriminante lineal. Seguir un procedimiento análogo al utilizado en el punto 5 de la sección dedicada a la Regla de Bayes.

   Entender la forma de la frontera discriminante a la vista de este gráfico.

Regresión Logística

1. Resolver el problema de clasificación mediante un modelo de regresión logística utilizando la función **glm** incluida en el paquete **stats** que viene con R. Para ello, es necesario utilizar $\{0, 1\}$ como etiquetas de las clases en lugar de $\{-1, +1\}$.

2. **Tasa de Error de la Regresión Logística**. Calcular el error de clasificación de la regresión logística en la muestra de entrenamiento y en la de test.

3. **Frontera discriminante de la Regresión Logística.** Graficar la frontera producida por la regresión logística sobre el mismo gráfico en el que se obtuvo la frontera de Bayes y la frontera del discriminante lineal.

4. **Gráfico de la Estimación de la Probabilidad a Posteriori**. En una nueva ventana de gráficos, utilizando el mismo procedimiento que en el punto 5 de la pág. 2, graficar la estimación de las probabilidades a posteriori de las clases producida por la regresión logística.

   Entender la forma de la frontera logística a la vista de este gráfico.

   Evaluar la bondad de estas estimaciones comparando este gráfico con el de las probabilidades a posteriori teóricas obtenido en en el punto 5 de la pág. 2.

Clasificación mediante Regresión Lineal

1. Resolver el problema de clasificación mediante un modelo de regresión lineal utilizando la función **lm** incluida en el paquete **stats** que viene con R. Utilizar la misma codificación $\{0, 1\}$ que para la regresión logística.

2. **Tasa de Error de la Regresión Lineal**. Calcular la tasa de error de clasificación de la regresión en la muestra de entrenamiento y en la de test.

3. **Frontera discriminante de la Regresión Lineal.** Graficar la frontera producida por la regresión lineal sobre el mismo gráfico en el que se obtuvieron las fronteras de las técnicas anteriores.

4. **Gráfico de la Estimación de la Probabilidad a Posteriori**. En una nueva ventana de gráficos, utilizando el mismo procedimiento que en el punto 5 de la pág. 2, graficar la estimación de las probabilidades a posteriori de las clases producida por la regresión lineal.

   Analizar la calidad de estas estimaciones de la probabilidad a posteriori (han de ser valores en $[0, 1]$).