Modelado Probabilístico Generativo Gaussiano de 2 Clases con Estimación de Parámetros por Máxima Verosimilitud (Marzo 2012)

Iván López Espejo

Deducción de la estimación por máxima verosimilitud de los parámetros correspondientes a las distribuciones de probabilidad de un modelo probabilístico generativo de clasificación gaussiano de 2 clases e iguales matrices de covarianza con el fin de calcular posteriormente el hiperplano de decisión para la clasificación de muestras entrantes al sistema.

I. Introducción

En este trabajo se deduce la estimación por máxima verosimilitud de los parámetros correspondientes a las distribuciones de probabilidad de un modelo probabilístico generativo de clasificación gaussiano de $K = 2$ clases e iguales matrices de covarianza. Una vez estimados estos, es posible calcular el hiperplano de decisión en base a la función logística sigmoidal. Calculado este, es posible clasificar una nueva muestra de entrada al sistema en una de las dos clases posibles sin más que estudiar la probabilidad de pertenencia a cada una de las clases a partir de la anterior función logística.

Para completar el trabajo, a partir de los resultados obtenidos, se realiza una implementación en MatLab con la que llevar a cabo una experimentación práctica.

II. Desarrollo

Partimos de poseer un conjunto de N muestras agrupables en dos clases según sendas distribuciones gaussianas. Para cada muestra, además, se tiene un valor de etiqueta, t_n, que indica a priori la pertenencia de dicha muestra a una clase, es decir,

$$\{\boldsymbol{x}_n, t_n\}, \qquad n = 1,2,\dots,N,$$

$$t_n = \begin{cases} 0 & \Leftrightarrow \boldsymbol{x}_n \in \mathtt{C}_0 \\ 1 & \Leftrightarrow \boldsymbol{x}_n \in \mathtt{C}_1 \end{cases},$$

donde, como se ve, t_n toma el valor 0 si la muestra $\boldsymbol{x}_n$ pertenece a la clase 0 ($\mathtt{C}_0$) o el valor 1 si pertenece a la clase 1 ($\mathtt{C}_1$). La probabilidad a priori de la primera clase es $P(\mathtt{C}_1) = \Pi$, de tal forma que, como sólo se poseen dos clases, la probabilidad a priori de la segunda es su complementario, es decir, $P(\mathtt{C}_0) = (1 - \Pi)$. Dado que hemos dicho que la densidad de probabilidad condicional de las muestras dada la clase se modela como una distribución gaussiana, las probabilidades conjuntas de las muestras y las clases se expresan, a partir de la Regla de Bayes, como

$$P(\boldsymbol{x}_n, \mathtt{C}_0) = P(\boldsymbol{x}_n|\mathtt{C}_0)P(\mathtt{C}_0) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})(1 - \Pi),$$

$$P(\boldsymbol{x}_n, \mathtt{C}_1) = P(\boldsymbol{x}_n|\mathtt{C}_1)P(\mathtt{C}_1) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\Pi.$$

Sea $\boldsymbol{t} = (t_1, t_2, \dots, t_N)^T$ el vector de etiquetas de pertenencia a una clase para las N muestras, la función de verosimilitud se expresa como

$$p(\boldsymbol{t}|\Pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\Pi]^{t_n} [\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})(1 - \Pi)]^{(1-t_n)},$$

donde, como se puede observar, sólo las probabilidades conjuntas de las muestras que pertenecen a su clase pesan en el anterior baremo. Los parámetros que deseamos estimar a continuación por máxima verosimilitud son las probabilidades a priori de cada clase, las medias de cada una de ellas y su matriz de covarianza. Para simplificar el cálculo, optimizamos sobre el logaritmo de la función de verosimilitud, pues su monotonía no cambia tras este tipo de composición:

$$\log p(\boldsymbol{t}|\Pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} t_n \log(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\Pi) + \\ + \sum_{n=1}^{N} (1 - t_n) \log\big(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})(1 - \Pi)\big),$$

donde la distribución normal multivariada de dos clases se expresa como

$$\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right\},$$

y, en términos logarítmicos como

$$\log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\log\left(2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right) - \\ -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}).$$

Sustituyendo la anterior expansión en el desarrollo logarítmico de la función de verosimilitud, esta resulta

$$\log p(\boldsymbol{t}|\Pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} t_n \log \Pi - \sum_{n=1}^{N} t_n \log\left(2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right) - \\ -\frac{1}{2}\sum_{n=1}^{N} t_n (\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_1) + \sum_{n=1}^{N} (1 - t_n) \log(1 - \Pi) - \\ -\sum_{n=1}^{N} (1 - t_n) \log(2\pi|\boldsymbol{\Sigma}|^{1/2}) - \frac{1}{2}\sum_{n=1}^{N} (1 - t_n)(\boldsymbol{x}_n - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_0).$$

Derivando con respecto a Π la anterior expresión e igualando a 0 optimizamos para calcular por máxima verosimilitud la probabilidad a priori de la primera clase y, por ende, la de la segunda:

$$\frac{\partial \log p(\boldsymbol{t}|\Pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\partial \Pi} = 0 \Rightarrow \frac{1}{\Pi}\sum_{n=1}^{N} t_n = \frac{1}{1-\Pi}\sum_{n=1}^{N}(1-t_n).$$

La anterior derivada es trivial, pues sólo depende de dos términos logarítmicos en la sumatoria de la función log-verosimilitud. Vamos a continuación a renombrar las sumatorias en ambos miembros de la anterior derivada. La suma de las etiquetas t_n nos proporciona la cantidad de muestras pertenecientes a la primera clase, es decir:

$$\sum_{n=1}^{N} t_n = N_1.$$

De otro lado, la suma del complementario, $(1-t_n)$, nos proporciona la cantidad de muestras pertenecientes a la clase 0, es decir:

$$\sum_{n=1}^{N}(1-t_n) = N_0.$$

La suma de ambos términos nos proporciona la cantidad total de muestras, N, de la forma:

$$N = N_0 + N_1.$$

Sustituyendo lo anterior en el resultado de la derivada parcial de la función log-verosimilitud con respecto a Π, llegamos a que la estimación ML de Π es

$$\frac{N_1}{\Pi} = \frac{N_0}{1-\Pi} \Rightarrow \frac{1}{\Pi} - 1 = \frac{N_0}{N_1} \Rightarrow \Pi^{-1} = \frac{N_0}{N_1} + 1 \Rightarrow$$
$$\Rightarrow \Pi = \frac{\mathrm{N}_1}{\mathrm{N}_0 + \mathrm{N}_1} = \frac{\mathrm{N}_1}{\mathrm{N}}.$$

Como era de esperar, la probabilidad a priori de C_1 es el número de muestras pertenecientes a dicha clase sobre el total, siendo la probabilidad a priori de la clase C_0 su complementario, es decir,

$$P(\mathrm{C}_0) = (1-\Pi) = 1 - \frac{\mathrm{N}_1}{\mathrm{N}} = \frac{\mathrm{N}-\mathrm{N}_1}{\mathrm{N}} = \frac{\mathrm{N}_0}{\mathrm{N}},$$

o, en otras palabras, el número de muestras pertenecientes a C_0 sobre el total.

A continuación, llevamos a cabo un procedimiento análogo al anterior para obtener la estimación de la media de la clase 1. Comenzamos derivando la función log-verosimilitud con respecto a $\boldsymbol{\mu}_1$:

$$\frac{\partial \log p(\boldsymbol{t}|\Pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}_1} = 0 \Rightarrow$$
$$\Rightarrow -\frac{\partial}{\partial \boldsymbol{\mu}_1}\left(\frac{1}{2}\sum_{n=1}^{N} t_n\,(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_1)\right) = 0.$$

Como vemos, sólo un término de la sumatoria (el del exponente de la gaussiana correspondiente) depende de la media de la clase 1. Para mayor facilidad de cálculo, expandimos su argumento:

$$(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_1) =$$
$$= \boldsymbol{x}_n{}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n - \boldsymbol{x}_n{}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1{}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n + \boldsymbol{\mu}_1{}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1.$$

A continuación tenemos en cuenta las siguientes tres identidades de derivación matricial para resolver el cálculo:

$$\frac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A}^T,$$

$$\frac{\partial(\boldsymbol{x}^T\boldsymbol{A})}{\partial \boldsymbol{x}} = \boldsymbol{A},$$

$$\frac{\partial(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^T\boldsymbol{x}.$$

Aplicándolas, llegamos a que la anterior derivada finalmente puede ser expresada como

$$\sum_{n=1}^{N} t_n\left[-(\boldsymbol{x}_n{}^T \boldsymbol{\Sigma}^{-1})^T - \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + (\boldsymbol{\Sigma}^{-1})^T\boldsymbol{\mu}_1\right] = 0.$$

Puesto que la matriz de covarianza es una matriz simétrica ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$), siendo su inversa también simétrica, $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + (\boldsymbol{\Sigma}^{-1})^T\boldsymbol{\mu}_1 = 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$. Además, por la regla de trasposición del producto matricial, tenemos que $(\boldsymbol{x}_n{}^T\boldsymbol{\Sigma}^{-1})^T = (\boldsymbol{\Sigma}^{-1})^T\boldsymbol{x}_n$. Aplicando además de nuevo la propiedad de simetría, el primer par de monomios del corchete puede simplificarse a su vez como $-(\boldsymbol{x}_n{}^T\boldsymbol{\Sigma}^{-1})^T - \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n = -2\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n$. Teniendo esto en cuenta, la estimación ML de $\boldsymbol{\mu}_1$ resulta finalmente

$$\sum_{n=1}^{N} t_n 2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{x}_n) = 0 \Rightarrow \boldsymbol{\mu}_1 \sum_{n=1}^{N} t_n = \sum_{n=1}^{N} t_n\,\boldsymbol{x}_n \Rightarrow$$
$$\Rightarrow \boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n\,\boldsymbol{x}_n.$$

Como era de esperar, resulta la media muestral del conjunto de muestras de la clase 1.

Procedemos a continuación de forma análoga con el fin de estimar la media de la clase 0:

$$\frac{\partial \log p(\boldsymbol{t}|\Pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}_0} = 0 \Rightarrow$$
$$\Rightarrow -\frac{\partial}{\partial \boldsymbol{\mu}_0}\left(\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\,(\boldsymbol{x}_n - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_0)\right) = 0.$$

Aplicando las mismas identidades de derivación matricial que en el caso anterior, llegamos a

$$\sum_{n=1}^{N}(1-t_n)\,[-(\boldsymbol{x}_n{}^T\boldsymbol{\Sigma}^{-1})^T-\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0+(\boldsymbol{\Sigma}^{-1})^T\boldsymbol{\mu}_0]=0.$$

Simplificando a partir de las mismas consideraciones, llegamos a que la estimación ML de la media de la clase 0 es

$$\sum_{n=1}^{N}(1-t_n)2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0-\boldsymbol{x}_n)=0\Rightarrow\boldsymbol{\mu}_0\sum_{n=1}^{N}(1-t_n)=\sum_{n=1}^{N}(1-t_n)\,\boldsymbol{x}_n\Rightarrow$$
$$\Rightarrow\boldsymbol{\mu}_0=\frac{1}{N_0}\sum_{n=1}^{N}(1-t_n)\,\boldsymbol{x}_n,$$

es decir, de nuevo, la media muestral del conjunto de muestras pertenecientes a C_0.

Finalmente, llevamos a cabo la estimación de la matriz de covarianza de ambas clases. De nuevo, derivamos la función log-verosimilitud, en este caso, respecto de $\boldsymbol{\Sigma}$. Teniendo en cuenta únicamente los términos dependientes de la matriz de covarianza, el problema se reduce a resolver la siguiente ecuación:

$$\frac{\partial\log p(\boldsymbol{t}|\Pi,\boldsymbol{\mu}_1,\boldsymbol{\mu}_0,\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}=0\Rightarrow$$
$$\Rightarrow\frac{\partial}{\partial\boldsymbol{\Sigma}}\Bigg(-\frac{1}{2}\sum_{n=1}^{N}t_n\log|\boldsymbol{\Sigma}|--\frac{1}{2}\sum_{n=1}^{N}t_n(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)-$$
$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\log|\boldsymbol{\Sigma}|-$$
$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)(\boldsymbol{x}_n-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_0)\Bigg)=0.$$

Para resolver la derivada del logaritmo del determinante de $\boldsymbol{\Sigma}$ hacemos uso de la siguiente identidad de derivación matricial:

$$\frac{\partial\log|\boldsymbol{A}|}{\partial\boldsymbol{A}}=(\boldsymbol{A}^{-1})^T.$$

Análogamente, aplicamos la siguiente identidad para resolver la derivada en los términos del argumento de la exponencial de las gaussianas:

$$\frac{\partial\boldsymbol{a}^T\boldsymbol{A}^{-1}\boldsymbol{b}}{\partial\boldsymbol{A}}=-(\boldsymbol{A}^{-1})^T\boldsymbol{a}\boldsymbol{b}^T(\boldsymbol{A}^{-1})^T.$$

Sustituyendo, la derivada resulta finalmente:

$$-\frac{1}{2}\sum_{n=1}^{N}t_n\boldsymbol{\Sigma}^{-1}+\frac{1}{2}\sum_{n=1}^{N}t_n\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\boldsymbol{\Sigma}^{-1}+$$
$$+\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_0)(\boldsymbol{x}_n-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}=0.$$

Para simplificar, multiplicamos por la derecha por la matriz de covarianza, resultando en

$$-\frac{N_1}{2}+\frac{\boldsymbol{\Sigma}^{-1}}{2}\sum_{n=1}^{N}t_n\,(\boldsymbol{x}_n-\boldsymbol{\mu}_1)(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T-\frac{N_0}{2}+$$
$$+\frac{\boldsymbol{\Sigma}^{-1}}{2}\sum_{n=1}^{N}(1-t_n)\,(\boldsymbol{x}_n-\boldsymbol{\mu}_0)(\boldsymbol{x}_n-\boldsymbol{\mu}_0)^T=0.$$

Multiplicando ambos términos de la ecuación por 2 y sacando factor común la matriz de covarianza inversa,

$$\boldsymbol{\Sigma}^{-1}\Bigg[\sum_{n=1}^{N}t_n\,(\boldsymbol{x}_n-\boldsymbol{\mu}_1)(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T+\sum_{n=1}^{N}(1-t_n)\,(\boldsymbol{x}_n-\boldsymbol{\mu}_0)(\boldsymbol{x}_n-\boldsymbol{\mu}_0)^T\Bigg]=$$
$$=N_0+N_1=N.$$

Finalmente, la estimación ML de la matriz de covarianza resulta

$$\boldsymbol{\Sigma}=\frac{1}{N}\Bigg[\sum_{n=1}^{N}t_n\,(\boldsymbol{x}_n-\boldsymbol{\mu}_1)(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T+\sum_{n=1}^{N}(1-t_n)\,(\boldsymbol{x}_n-\boldsymbol{\mu}_0)(\boldsymbol{x}_n-\boldsymbol{\mu}_0)^T\Bigg],$$

que de nuevo se traduce, como era de esperar, en la covarianza muestral.

La probabilidad de la primera clase dada la muestra se puede expresar en términos de la función logística como $P(\mathrm{C}_1|\boldsymbol{x}_n)=\sigma(\boldsymbol{w}^T\boldsymbol{x}_n+w_0)$, donde

$$\sigma(\boldsymbol{w}^T\boldsymbol{x}_n+w_0)=\frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}_n+w_0)}}.$$

Además, tenemos que $a=\boldsymbol{w}^T\boldsymbol{x}_n+w_0=\log\frac{P(\mathrm{C}_1|\boldsymbol{x}_n)}{P(\mathrm{C}_0|\boldsymbol{x}_n)}$, de tal forma que en el límite, cuando las dos clases son equiprobables, $a=0=\boldsymbol{w}^T\boldsymbol{x}_n+w_0$, lo que constituye nuestro hiperplano de decisión. Resolviendo la anterior ecuación, llegamos a la forma explícita del hiperplano de decisión:

$$\boldsymbol{w}^T\boldsymbol{x}+w_0=0\Rightarrow(w_1\;w_2)\begin{pmatrix}x\\y\end{pmatrix}+w_0=0\Rightarrow w_1x+w_2y+w_0=0\Rightarrow$$
$$\Rightarrow y(x)=-\frac{w_1x+w_0}{w_2},$$

donde $\boldsymbol{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)$ y $w_0=-\frac{1}{2}\boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+$ $+\boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\log\frac{P(\mathrm{C}_1)}{P(\mathrm{C}_0)}$.

Finalmente se ha llevado a cabo una implementación en MatLab que realiza todo el proceso: generación de muestras de dos clases según distribuciones gaussianas, estimación de sus parámetros por máxima verosimilitud, cálculo del hiperplano de decisión y clasificación de una nueva muestra de entrada. La figura 1 muestra un ejemplo de resultado del funcionamiento de las rutinas programadas.

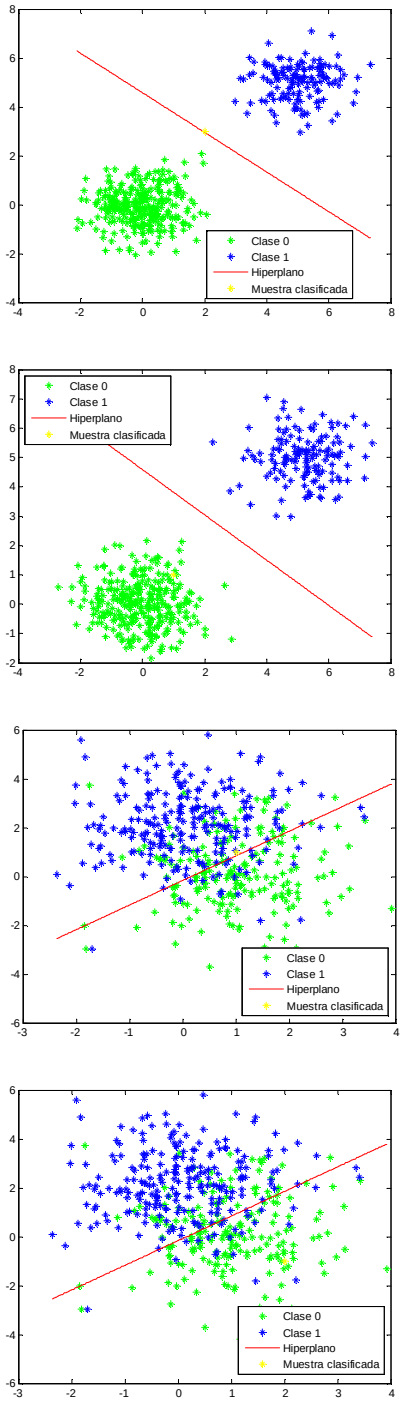

Fig. 1. De arriba a abajo: (a) dos clases con vectores de medias (0,0) y (5,5) y matriz de covarianzas diagonal con autovalores 0.8 y 0.6 donde se clasifica una muestra muy cercana al hiperplano de decisión, (b) igual caso que (a) pero donde se clasifica una muestra con claridad en la clase 0, (c) dos clases con vectores de medias (1,0) y (0,2) y matriz de covarianzas diagonal con autovalores 1 y 2 donde se clasifica una muestra muy cercana al hiperplano de decisión, (d) igual caso que (c) pero donde se clasifica una muestra más probable en clase 0.