FACULTAD DE NEGOCIOS
LICENCIATURA EN ACTUARÍA

Estadística Multivariada.

Andrea Cañas Caboada.

Emmanuel Rivera Ló

Tarea 2° Parcial: Análisis Discriminante

Grupo: ACT-600

Titular: Dr. Julio César Galindo López.

Fecha de Entrega: Jueves 21 de Marzo del 2024.

Semestre 2024-I

Universidad La Salle México

Facultad de Negocios

Licenciatura en Actuaría

Estadística Multivariada.
Tarea 2° Parcial: Análisis Discriminante.

Andrea Cañas Taboada.
Emmanuel Rivera López
Titular: Dr. Julio César Galindo López

January-June 2024
Semestre 2024-I

1. Análisis Discriminante

1. Supongamos que $X_1 \sim \mathbb{N}_r(\mu_1, \Sigma_{X,X})$ y $X_2 \sim \mathbb{N}_r(\mu_2, \Sigma_{X,X})$ son independientes. Considere el estadístico

$$\frac{\mathbb{E}[a^t X_1] - \mathbb{E}[a^t X_2])^2}{\text{Var}[a^t X_1 - a^t X_2]}$$

como función de a. Demuestra que $a \propto \Sigma_{X,X}^{-1}(\mu_1 - \mu_2)$ maximiza el estadístico usando multiplicadores de Lagrange.
Demostración: Notése que el cociente dado es el criterio de Fisher que permitirá elegir un vector a, que permita maximizar el ratio de ***la varianza entre clases*** respecto a ***la varianza dentro de las clases;*** así, sin pérdida de generalidad, que el término del denominador puede expresarse en términos de la matriz de dispersión dentro de las clases de $\mathbf{x}$ cómo sigue:

$$\begin{aligned}\text{Var}[a^{Tr}X_1 - a^{Tr}X_2] &= a^t \text{Var}[X_1 - X_2] = a_{Tr}\Sigma_{X,X} \\ &= \sum_{i \in \mathscr{C}_k} (a^{Tr}\mathbf{x}^{(i)} - a^{Tr}\mu_k)^2 \\ &= \sum_{i \in \mathscr{C}_k} a^{Tr}(\mathbf{x}^{(i)} - \mu_k)(\mathbf{x}^{(i)} - \mu_k)^{Tr}a \\ &= a^{Tr}\mathscr{S}_k a\end{aligned}$$

Por lo tanto, el denominador se puede reescribir de la forma:

$$\begin{aligned}\text{Var}[a^{Tr}X_1 - a^{Tr}X_2] &= a^{Tr}\mathscr{S}_k a = a^{Tr}\mathscr{S}_1 a + a^{Tr}\mathscr{S}_2 a \\ &= a^T \mathscr{S}_W a\end{aligned}$$

dónde $\mathscr{S}_W$ denota la covarianza entre clases; de manera similar, el término del numerador correspondiente a la diferencia entre las medias proyectadas, puede reescribirse en términos de la *matriz de dispersión entre clases; esto es*

$$\begin{aligned}\mathbb{E}[\mathbf{a}^t X_1] - \mathbb{E}[\mathbf{a}^t X_2])^2 &= (\widehat{\mu}_1 - \widehat{\mu}_2)^2 = (a^{Tr}\mu_1 - a^{Tr}\mu_2)^2 \\ &= (\mathbf{a}^{Tr}\mu_1 - \mathbf{a}^{Tr}\mu_2)(\mathbf{a}^{Tr}\mu_1 - \mathbf{a}^{Tr}\mu_2)^{Tr} \quad = \mathbf{a}^{Tr}(\mu_1 - \mu_2)(\mu_1 - \mu_2)\mathbf{a} \\ &= \mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}\end{aligned}$$

dónde $\mathscr{S}_B$ denota la varianza entre clases. Así el criterio de Fisher puede reescribirse de la forma

$$J(a) = J(a) = \frac{\mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}}{\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}}$$

De aquí, para maximizar el criterio, se planteará el siguiente problema de optimización equivalente al conciente previo:

$$\begin{aligned}&\text{máx. } \mathbf{a}^{Tr}\mathscr{S}_B a \\ &\text{sujeto a:} \\ &\qquad \mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a} = 1\end{aligned}$$

Aplicando la definición de multiplicadores de Lagrange, se tiene lo que sigue:

$$\begin{aligned}
\frac{d}{d\mathbf{a}}J(a) &= \frac{\partial J(a)}{\partial \mathbf{a}}\left[\frac{\mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}}{\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}}\right] = \mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}\frac{\partial J(a)}{\partial \mathbf{a}}[\mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}] - \mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}\frac{\partial J(a)}{\partial \mathbf{a}}[\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}] = 0 \\
&= [\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}]2\mathscr{S}_B\mathbf{a} - [\mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}]2\mathscr{S}_W\mathbf{a} = 0 \\
&= \frac{1}{\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}}[[\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}]2\mathscr{S}_B\mathbf{a} - [\mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}]2\mathscr{S}_W\mathbf{a}] = \frac{1}{\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}}[0] \\
&= \frac{[\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}]\mathscr{S}_B\mathbf{a}}{\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}} - \frac{[\mathbf{a}^{Tr}\mathscr{S}_B\mathbf{a}]2\mathscr{S}_W\mathbf{a}}{\mathbf{a}^{Tr}\mathscr{S}_W\mathbf{a}} = 0 \\
&= \mathscr{S}_B\mathbf{a} - J\mathscr{S}_W\mathbf{a} = 0 \\
&= \mathscr{S}_W^{-1}\mathscr{S}_B\mathbf{a} - J\mathbf{a} = 0
\end{aligned}$$

Despejando a $\mathbf{a}$, se tiene lo que sigue:

$$\begin{aligned}
\mathscr{S}_W^{-1}\mathscr{S}_B\mathbf{a} - J\mathbf{a} = 0 &\mapsto J\mathbf{a} = \mathscr{S}_W^{-1}\mathscr{S}_B\mathbf{a} \\
&\mapsto J\mathbf{a} = \mathscr{S}_W^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^{Tr}\mathbf{a} \\
&\mapsto J\mathbf{a} = \mathscr{S}_W^{-1}(\mu_1-\mu_2)\underbrace{(\mu_1-\mu_2)^{Tr}\mathbf{a}]}_{k\in\mathbb{R}} \\
&\mapsto J\mathbf{a} = k\mathscr{S}_W^{-1}(\mu_1-\mu_2) \mapsto \mathbf{a} = \frac{k\mathscr{S}_W(\mu_1-\mu_2)}{J}
\end{aligned}$$

De aquí, la solución del problema de optimización implica la relación de proporcionalidad de $\mathscr{S}_B \propto (\mu_1-\mu_2)$, por otro lado, no es difícil ver que $\mathscr{S}_W^{-1}(\mu_1-\mu_2) \propto \mathbf{a}$, por lo que esto demuestra que:

$$\mathbf{a} \propto \Sigma_{X,X}^{-1}(\mu_1-\mu_2)$$

■

2. Diga, explicando sus conclusiones, si las siguientes a rmaciones son falsas o verdaderas. Justifique adecuadamente sus respuestas.

a) Si el límite de decisión de Bayes es lineal, ¿esperamos que LDA o QDA funcionen mejor en el conjunto de entrenamiento?
Si el límite de desición Bayesiano es lineal, se esperaría un mejor funcionamiento del análisis por discriminación cuadrático (QDA), ya que este será mucho más manejable considerando el término cuadrático y tendería a un ajuste mucho más cercano; por otra parte, se puede inferir que el QDA no sería el ideal para el conjunto de prueba, ya que si se tiene un conjunto en el cuál las covarianzas de las clases sean 'similares' por el supuesto de ser lineales, implicaría que esté sujeto a un sobreajuste no deseado.

b) Si el límite de decisión de Bayes no es lineal, ¿esperamos que LDA o QDA funcionen mejor en el conjunto de entrenamiento?
Si el límite NO es lineal, se esperaría un mejor rendidimiento del QDA para el conjunto de entrenamiento, recordando el hecho de que esta técnica permite considerar límites de decisión con una complejidad superior, y automáticamente se descartan los límites establecidos por el LDA.

c) En general, conforme el tamaño de la muestra n aumenta, esperamos que la precisión de la predicción de la prueba de QDA en relación con LDA mejore, disminuya o se mantenga sin cambios?, ¿Por que?
Sin pérdida de generalidad, se esperaría que la precisión de ambos métodos sea igualmente buena cuándo n tienda a valores muy grandes; no obstante, el QDA ayudaría a evitar un sobreajuste (en el caso de que los límites de decisión sean plausibles y no lineales). Más aún, se esperá que conforme aumente el tamaño de la muestra, se podría buscar un método alternativo para realizar el análisis.

d) Si el límite de decisión de Bayes para un problema determinado es lineal, probablemente lograremos una tasa de error de prueba superior usando QDA en lugar de LDA porque QDA es lo suficientemente flexible como para modelar un límite de decisión lineal.
Este enunciado es falso, por definición, se sabe que se debe de contar con una muestra relativamente grande, y en el caso de que se tengan 'pocos' elementos pertenecientes a la muestra, se esperaría que al emplear el QDA haya una probabilidad alta de que se presente un sobreajuste, implicando un error de prueba mucho mayor a comparación del LDA.

3. Supongamos que se tienen dos variables aleatorias X_1 y X_2. Tomemos $X_3 = X_1^2, X_4 = X_2^2$ y $X_5 = X_1X_2$. Dibuja las fronteras de LDA. Toma una base de datos de cualquier API y verique que, efectivamente, se tienen las fronteras dibujadas.

4. Suponga que se coleccionan datos de un cierto grupo con variables aleatorias:

a) X_1 las horas de estudio;

b) X_2 promedio sobre 5;

c) X_3 recibe una calificación sobre 10.

Suponga además que $\widehat{\beta}_0 = -6, \widehat{\beta}_1 = 0.005$ y $\widehat{\beta}_2 = 1$

a) ¿Que significan tales estimaciones para $(\beta_0, \beta_1, \beta_2)$.?
Solución: Siguiendo la filosofía del modelo de Logit $n-$dimensional, se tiene que los estimadores representan lo que sigue:
$\widehat{\beta}_0$: Se estima que si las horas de estudio así cómo el promedio obtenido sobre 5 son constantes, i.e. iguales a cero, y se hace una estimación al recibir una calificación de 10, entonces se obtendrá -6 cómo valor del estimador.
$\widehat{\beta}_1$: Se esperaría que el estudiante obtiviera una calificación de 10 respecto a que se promediará sobre 5; y establece que si se es constante en las horas de estudio, por cada hora adicional habrá un aumento del 0.005 en la calificación obtenida.
$\widehat{\beta}_2$: Se esperá obtener una calificación de 10 que cambie respecto al promedio sobre 5 manteniendo constancia en las horas de estudio; esto es, por cada punto obtenido sobre la calificación sobre 5 se esperará un aumento en la calificación.

b) Estime la probabilidad de que un estudiante que estudia 40 horas a la semana y tenga promedio de 35, obtenga una calificación de 10.
Solución: Siguiendo la filosofía del modelo de Logit $n-$dimensional, se considerará la función para la probabilidad de que el estudiante tenga éxito para obtener una calificación excelente; para este caso, la función del cálculo de probabilidades está dada por:

$$\mathbb{P}[\Pi_1|\mathbf{x}] = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 X_1+\widehat{\beta}_2 X_2}}{1+e^{\widehat{\beta}+\widehat{\beta}_1 X_1+\widehat{\beta}_2 X_2}}$$

Redefiniendo los valores de las estimaciones de $\widehat{\beta}_0$ y $\widehat{\beta}_1$, se tiene lo que sigue:

$$\begin{aligned}\mathbb{P}[\Pi_1|\mathbf{x}] &= \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 X_1+\widehat{\beta}_2 X_2}}{1+e^{\widehat{\beta}+\widehat{\beta}_1 X_1+\widehat{\beta}_2 X_2}} = \frac{e^{-6+(0.005)(40)+(1)(3.5)}}{1+e^{-6+(0.005)(40)+(1)(3.5)}} = \frac{e^{-6+0.2+3.5}}{1+e^{-6+0.2+3.5}} \\ &= \frac{e^{-2.3}}{1+e^{-2.3}} = \frac{0.1002588437}{1+0.1002588437} = 0.091122961010485613533930 47\end{aligned}$$

c) ¿Cuantas horas debe estudiar para obtener un 50 % de posibilidades de obtener un 10?
Solución: Se procederá de manera análoga con la fórmula de estimación de probabilidades para el modelo de Logit, solamente que note que $\mathbb{P}[\Pi_1|\mathbf{x}] = 1/2$, por lo que se procederá a despejar la ecuación para determinar el número de horas de estudio; esto es:

$$\begin{aligned}0.5 = \frac{e^{-6+0.005X_1+3.5}}{1+e^{-6+0.005X_1+3.5}} &\mapsto 0.5(1+e^{-6+0.005X_1+3.5}) = e^{-6+0.005X_1+3.5} \\ &\mapsto 0.5+0.5(1+e^{-6+0.005X_1+3.5}) = e^{-6+0.005X_1+3.5} \\ &\mapsto 0.5 = e^{-6+0.005X_1+3.5} - 0.5(1+e^{-6+0.005X_1+3.5}) \\ &\mapsto e^{-6+0.005X_1+3.5} = 1 \\ &\mapsto \log(e^{-6+0.005X_1+3.5}) = \log(1) \\ &\mapsto -6+0.005X_1+3.5 = 0 \mapsto 0.005X_1 = 6-3.5 \\ &\mapsto X_1 = \frac{2.5}{0.005} = 500\end{aligned}$$

5. Suponga que una observacion en el k-ésimo grupo se toma a partir de una distribución $\mathcal{N}(\mu_k, \sigma^2)$. Demuestra que el clasificador bayesiano asigna una observación a la clase para la cual la función discriminante se máximiza.
Demostración: Por hipótesis, para determinar el clasificador Bayesiano se debe de considerar la función de probabilidad a posteriori definida cómo:

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

Y la función de discriminación Bayesiana está definida cómo:

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

Por el hecho de que σ^2 es una constante, $p(\pi_i|\mathbf{x})$ se reduce a:

$$p_k(x) = \frac{\pi_k \exp\left\{-\dfrac{1}{2\sigma^2(\mathbf{x}-\mu_k)^2}\right\}}{\displaystyle\sum_{i=1}^{k} \pi_r \exp\left\{-\frac{1}{2\sigma^2(\mathbf{x}-\mu_k)^2}\right\}}$$

Maximizar la función de probabilidad a posteriori permite maximizar a su vez cualquier función monótona creciente definida por $p_k(X)$, por lo que se considerará maximizar convenientemente $\log[p_k(\mathbf{X})]$, así:

$$\log[p_k(X)] = \log(\pi_k) - \frac{1}{2\sigma^2}(\mathbf{x}-\mu_k)^2 - \log\left[\sum_{i=1}^{k} \pi_r \exp\left\{-\frac{1}{2\sigma^2(\mathbf{x}-\mu_k)^2}\right\}\right]$$

Recuerde además que ya que se está maximizando sobre k, y ya que el último término se mantiene invariante k puede ser omitido. Por lo tanto, quedará maximizar:

$$\begin{aligned} f &= \log[\pi_k] - \frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) \\ &= \log[\pi_k] - \frac{x^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \end{aligned}$$

Por lo tanto, se tiene que el clasificador bayesiano asigna una observación a la clase tal que la función efectivamente máximiza, y está dada por:

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

■

6. Supongamos que deseamos predecir si una determinada acción emitirá un dividendo este año (Sí ó No) basándose en X, el porcentaje de beneficio del año pasado. Examinamos un gran número de empresas y descubrimos que el valor medio de X para las empresas que emitieron dividendos fue de 10, mientras que el valor medio para las que no lo hicieron fue 0. Además, la varianza de X para estas dos conjuntos de empresas fue 36. Finalmente, el 80 % de las empresas emitieron dividendos. Suponiendo que X sigue una distribución normal, prediga la probabilidad de que una empresa emita un dividendo este año dado que su beneficio porcentual fue $X = 4$ el año pasado.

Solución: En primera instancia, se procederá a definir los eventos en términos de cómo siguen una v.a. normal; defínase a u cómo la v.a. que mide el número de empresas (D) que emitirán un dividendo, la cuál sigue una distribución normal con parámetros $u_{(D)} \sim \mathcal{N}(10, 36)$; en contraste, el valor u para las compañías (ND) que no emiten dividendos, las cuáles siguen una distribución normal con parámetros $u_{ND} \sim \mathcal{N}(0, 36)$, adenás, se tiene que la probabilidad de que se emitan dividendos es $\mathbb{P}[D] = 0.8$; dado esto, se debe de calcular la probabilidad de que cierta empresa empita un dividendo dado que su beneficio el año pasado fue 4; aplicando la definición del clasificador de Naive-Bayes, se tiene lo que sigue:

$$\begin{aligned} \mathbb{P}[\Pi|\mathbf{X}] &= \frac{\mathbb{P}[\Pi_1|\mathbf{X}=\mathbf{x}_1]\mathbb{D}}{\mathbb{P}[\Pi_1|\mathbf{X}=\mathbf{x}_1]\mathbb{P}[D] + \mathbb{P}[\Pi_2|\mathbf{X}=\mathbf{x}_2]\mathbb{P}[ND]} \\ &= \frac{\pi_D \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu_D^2)}{\sigma^2}\right\}}{\pi_D \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu_D^2)}{\sigma^2}\right\} + \pi_{ND} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu_{ND}^2)}{\sigma^2}\right\}} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\left[\pi_D \exp\left\{-\frac{1}{2}\frac{(x-\mu_D^2)}{\sigma^2}\right\}\right]}{\frac{1}{\sqrt{2\pi\sigma^2}}\left[\pi_D \exp\left\{-\frac{1}{2}\frac{(x-\mu_D^2)}{\sigma^2}\right\} + \pi_{ND} \exp\left\{-\frac{1}{2}\frac{(x-\mu_{ND}^2)}{\sigma^2}\right\}\right]} \\ &= \frac{\pi_D \exp\left\{-\frac{1}{2}\frac{(x-\mu_D^2)}{\sigma^2}\right\}}{\pi_D \exp\left\{-\frac{1}{2}\frac{(x-\mu_D^2)}{\sigma^2}\right\} + \pi_{ND} \exp\left\{-\frac{1}{2}\frac{(x-\mu_{ND}^2)}{\sigma^2}\right\}} = \frac{\left(\frac{4}{5}\right)\exp\left\{-\frac{(4-10)^2}{2(36)}\right\}}{\left(\frac{4}{5}\right)\exp\left\{-\frac{(4-10)^2}{2(36)}\right\} + \left(\frac{1}{5}\right)\exp\left\{-\frac{(4-0)^2}{2(36)}\right\}} \\ &= \frac{\left(\frac{4}{5}\right)\exp\left\{-\frac{36}{72}\right\}}{\left(\frac{4}{5}\right)\exp\left\{-\frac{36}{72}\right\} + \left(\frac{1}{5}\right)\exp\left\{-\frac{16}{72}\right\}} = \frac{(\frac{4}{5})e^{-1/2}}{(\frac{4}{5})e^{-1/2} + (\frac{1}{5})e^{-2/9}} \\ &= 0.75185245329752618430729702 5 \end{aligned}$$