



Licenciatura en Actuaría

Materia: *Estadística Multivariada*

Actividad: *Tarea 2*

21 / 03 / 24

Nombre(s): *Emilio Valencia*
Yair Castello
Abel Torres

UNIVERSIDAD La Salle
Facultad de Negocios
Estadística Multivariada
Examen Final

1. LDA Y QDA

Ejercicio 1. Supongamos que $X_1 \sim N_r(\mu_1, \Sigma_{X,X})$ y $X_2 \sim N_r(\mu_2, \Sigma_{X,X})$ 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.

Ejercicio 2. Diga, explicando sus conclusiones, si las siguientes afirmaciones son falsas o verdaderas. Justifique adecuadamente su respuesta.

- (1) Si el límite de decisión de Bayes es lineal, ¿esperamos que LDA o QDA funcionen mejor en el conjunto de entrenamiento?
- (2) Si el límite de decisión de Bayes no es lineal, ¿esperamos que LDA o QDA funcionen mejor en el conjunto de entrenamiento?
- (3) 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 qué?
- (4) 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.

Ejercicio 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 $X_5 = X_1 X_2$. Dibuja las fronteras de LDA. Toma una base de datos de cualquier API y verifica que, efectivamente, se tienen las fronteras dibujadas.

Ejercicio 4. Considere algún conjunto de datos tomado de Kaggle¹. **Nota: El conjunto de datos debe ser distinto para todos.** Define $X_5 = X_1/X_2$ y $X_6 = X_3/X_4$.

- (1) ¿Qué mide X_5 y X_6 ?
- (2) Toma la siguiente transformación: $\log(X_5)$ y $\log(X_6)$. Grafica los datos transformados y realiza un LDA.
- (3) Realiza lo anterior para QDA.
- (4) Generaliza para n cualquiera.

Ejercicio 5. Suponga que $X_1 \sim N_n(\mu_1, \Sigma_1)$ y que $X_2 \sim N_n(\mu_2, \Sigma_2)$. Suponga que $\Sigma_1 \neq \Sigma_2$. Demuestra que la log-verosimilitud del cociente de discriminación es de la forma

$$Q(x) = \beta_0 + \beta^T x + x^T O x$$

en donde tendrás que encontrar a β_0 , β y a O .

Ejercicio 6. Suponga que se coleccionan datos de un cierto grupo con variables aleatorias:

¹<https://www.kaggle.com>

- (1) X_1 horas de estudio.
- (2) X_2 promedio sobre 5.
- (3) X_3 recibe una calificación de 10.

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

- (1) ¿Qué significan tales estimaciones para $(\beta_0, \beta_1, \beta_2)$.
- (2) Estima la probabilidad de que un estudiante que estudia 40 horas a la semana y tenga promedio de 3.5, obtenga una calificación de 10.
- (3) ¿Cuántas horas debe estudiar para obtener un 50% de posibilidades de obtener un 10.

Ejercicio 7. Suponga que una observación en el k -ésimo grupo se toma a partir de una distribución $N(\mu_k, \sigma^2)$. Demuestra que el clasificador bayesiano asigna una observación a la clase para la cual la función discriminante se maximiza.

Ejercicio 8. Supongamos que deseamos predecir si una determinada acción emitirá un dividendo este año ("Sí" o "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.

Ejercicio 9. Consideremos los datos de The Insurance Company Benchmark, que pueden descargarse de

kdd.ics.uci.edu/databases/tic

. Hay 86 variables sobre datos de uso de productos y datos sociodemográficos derivados de los códigos postales de los clientes de una compañía de seguros. Hay un conjunto de aprendizaje ticdata2000.txt de 5,822 clientes y un conjunto de prueba ticeval2000.txt de 4,000 clientes. Los clientes del conjunto de aprendizaje se clasifican en dos clases, dependiendo de si contrataron una póliza de seguro de caravana. El problema es predecir quién del conjunto de prueba estaría interesado en comprar una póliza de seguro de caravana. Utilice cualquiera de los métodos de clasificación en los datos de aprendizaje y, a continuación, aplíquelos a los datos de prueba. Compare sus predicciones para el conjunto de prueba con las del archivo tictgts2000.txt y calcule la tasa de error del conjunto de prueba. ¿Qué variables son más útiles para predecir la contratación de un seguro de caravana?

Ejercicio 1. Supongamos que $X_1 \sim N_r(\mu_1, \Sigma_{X,X})$ y $X_2 \sim N_r(\mu_2, \Sigma_{X,X})$ 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.

$$\mathbb{E}[a^t X_1] = a^t \mathbb{E}[X_1] = a^t \mu_1$$

$$\mathbb{E}[a^t X_2] = a^t \mathbb{E}[X_2] = a^t \mu_2$$

$$\text{Var}(a^t X_1 - a^t X_2) = \text{Var}(a^t X_1) + \text{Var}(a^t X_2) = 2 \text{Var}(a^t X_1) = a^t \Sigma_{X,X} a$$

Por lo tanto el estadístico es:

$$\frac{(a^t \mu_1 - a^t \mu_2)^2}{a^t \Sigma_{X,X} a}$$

Utilizando Lagrange

$$L(a, \lambda) = (a^t \mu_1 - a^t \mu_2)^2 - \lambda(a^t a - 1)$$

$$\frac{\partial L}{\partial a} = 2(a^t \mu_1 - a^t \mu_2)(\mu_1 - \mu_2) - 2\lambda a$$

$$2(a^t \mu_1 - a^t \mu_2)(\mu_1 - \mu_2) - 2\lambda a = 0$$

$$2(a^t \mu_1 - a^t \mu_2)(\mu_1 - \mu_2) = 2\lambda a$$

$$(a^t \mu_1 - a^t \mu_2)(\mu_1 - \mu_2) = \lambda a$$

Entonces nos queda

$$a \propto \lambda^{-1}(\mu_1 - \mu_2)$$

$$a \propto (\mu_1 - \mu_2) \rightarrow \text{maximiza}$$

Ejercicio 2. Diga, explicando sus conclusiones, si las siguientes afirmaciones son falsas o verdaderas. Justifique adecuadamente su respuesta.

- (1) Si el límite de decisión de Bayes es lineal, ¿esperamos que LDA o QDA funcionen mejor en el conjunto de entrenamiento?
- (2) Si el límite de decisión de Bayes no es lineal, ¿esperamos que LDA o QDA funcionen mejor en el conjunto de entrenamiento?
- (3) 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 qué?
- (4) 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.

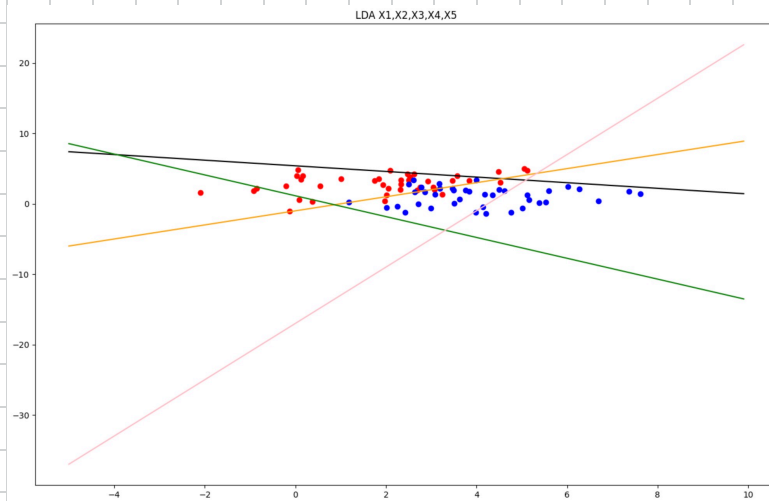
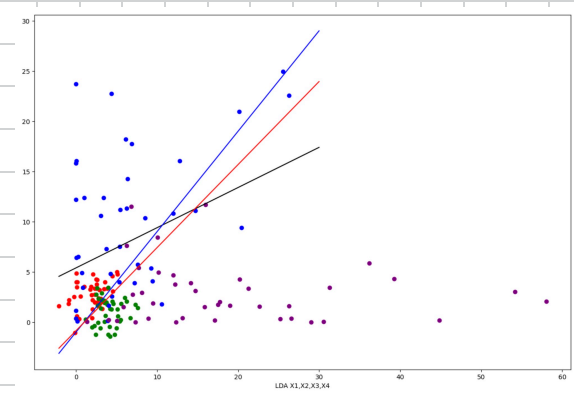
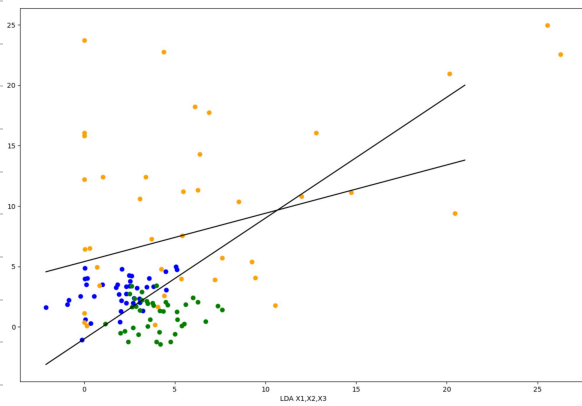
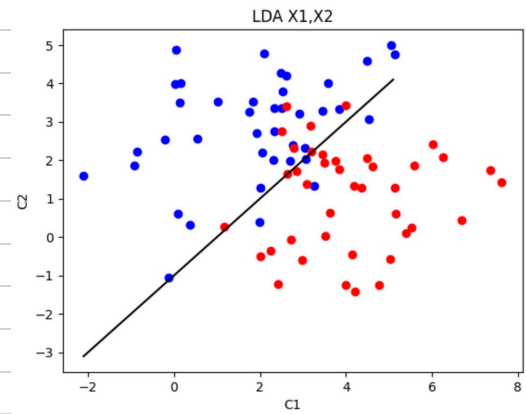
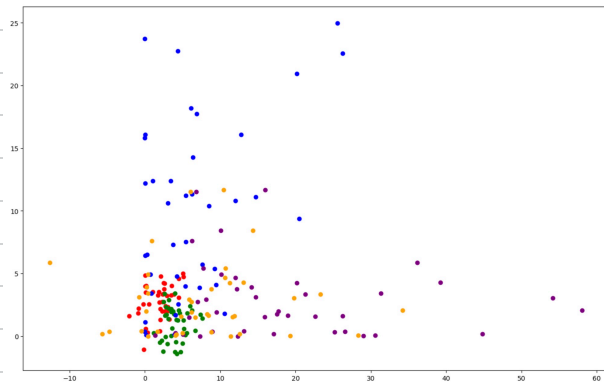
(1) El LDA funcionará mejor para una frontera lineal, ya que no es necesario el cuadrático.

(2) El QDA funcionará mejor porque permite modelar límites no lineales.

(3) Esperamos que mejore porque se pueden estimar diferentes matrices de covarianza para cada clase.

(4) El QDA puede tener mayor error al sobreajustar un modelo con límite lineal.

Ejercicio 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 verifica que, efectivamente, se tienen las fronteras dibujadas.



Ejercicio 6. Suponga que se coleccionan datos de un cierto grupo con variables aleatorias:

- (1) X_1 horas de estudio.
- (2) X_2 promedio sobre 5.
- (3) X_3 recibe una calificación de 10.

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

- (1) ¿Qué significan tales estimaciones para $(\beta_0, \beta_1, \beta_2)$.
- (2) Estima la probabilidad de que un estudiante que estudia 40 horas a la semana y tenga promedio de 3.5, obtenga una calificación de 10.
- (3) ¿Cuántas horas debe estudiar para obtener un 50% de posibilidades de obtener un 10.

1) Significa que con estos coeficientes se puede hacer regresión sobre esta probabilidad.

$$2) P(\pi_1 | (40, 3.5)) = \frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2}} = \frac{e^{-6 + 0.005(40) + 3.5}}{1 + e^{-6 + 0.005(40) + 3.5}} = 0.0911$$

$$3) P(\pi_1 | (X_1, 3.5)) = 0.5 = \frac{1}{1 + e^{6 - 0.005 X_1 - 3.5}} \Rightarrow e^{6 - 0.005 X_1 - 3.5} = 1$$

$$6 - 0.005 X_1 - 3.5 = 0$$

$$\Rightarrow 500 //$$

Ejercicio 7. Suponga que una observación en el k -ésimo grupo se toma a partir de una distribución $N(\mu_k, \sigma^2)$. Demuestra que el clasificador bayesiano asigna una observación a la clase para la cual la función discriminante se maximiza.

El clasificador asigna una observación a la clase con la mayor probabilidad posterior, esto es

$$p(\pi_k | x) = \frac{f_k(x) \pi_k}{\sum_i f_i(x) \pi_i} = \frac{\pi_k e^{\left\{ -\frac{1}{2\sigma^2} (x - \mu_k)^2 \right\}}}{\sum_{i=1}^K \pi_i e^{\left\{ -\frac{1}{2\sigma^2} (x - \mu_i)^2 \right\}}}$$

aplicando logaritmo natural

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

Es equivalente a encontrar k para lo cual se maximiza

$$\frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

//

Ejercicio 8. Supongamos que deseamos predecir si una determinada acción emitirá un dividendo este año ("Sí" o "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.

$$\begin{aligned}\bar{X}_1 &= 10 \\ \bar{X}_2 &= 0\end{aligned}$$

$$\sigma^2 = 36$$

$$\pi_1 = 0.8$$

$$\pi_2 = 0.2$$

$$p(\Pi_i | \mathbf{x}) = P(\mathbf{X} \in \Pi_i | \mathbf{X} = \mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2},$$

$$P(\pi_1 | X = 4) = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{4-10}{6}\right)^2} (0.8)}{\frac{1}{\sigma\sqrt{2\pi}} \left(e^{-\frac{1}{2}\left(\frac{4-10}{6}\right)^2} (0.8) + e^{-\frac{1}{2}\left(\frac{4-0}{6}\right)^2} (0.2) \right)} = 0.7518 \approx 75.18\%$$