Universidad Veracruzana

Maestría en Inteligencia Artificial

Visión por Computadora

Tarea 6. Aplicación de LDA a la base de datos de pingüinos palmer en MATLAB para el caso de dos clases y para el caso multiclase.

Ángel García Báez

Profesor: Dr. Héctor Acosta Mesa y Dra. Adriana Laura López Lobato

April 11, 2025

Contents

1 Objetivo de la práctica **2**

2 Metodología **5**

3 Resultados para dos clases **10**
3.1 Vectores de medias 10
3.2 Matriz de dispersión dentro de clases 10
3.3 Matriz de dispersión entre clases 11
3.4 Valores y vectores propios 12
3.5 Proyección de las nuevas componentes 13
3.6 Proyección en 1D 14

4 Resultados para tres clases **15**
4.1 Vectores de medias 15
4.2 Matriz de dispersión dentro de clases 15
4.3 Matriz de dispersión entre clases 16
4.4 Valores y vectores propios 17
4.5 Proyección de las nuevas componentes 18
4.6 Proyección en 2D 19

5 Conclusiones **20**

6 Referencias **21**

7 Anexos **22**
7.1 Implementación de la exploración y el LDA en MATLAB . . 22

1 Objetivo de la práctica

Se tiene la base de datos de pingüinos palmer, la cual representa las mediciones de 3 especies de pingüinos en distintas islas, a lo largo de distintos años, la cual tiene la siguiente estructura:

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3,750	male	2,007
Adelie	Torgersen	39.5	17.4	186	3,800	female	2,007
Adelie	Torgersen	40.3	18.0	195	3,250	female	2,007
Adelie	Torgersen						2,007
Adelie	Torgersen	36.7	19.3	193	3,450	female	2,007
Adelie	Torgersen	39.3	20.6	190	3,650	male	2,007
Adelie	Torgersen	38.9	17.8	181	3,625	female	2,007

Figure 1: Base de datos de los pingüinos palmer (primeros 5 casos)

La base esta compuesta por 344 observaciones y 8 variables (4 variables categóricas o de etiqueta y 4 variables numéricas continuas). Para efectos del desarrollo del documento, se tomaran en cuenta unicamente las 4 variables numéricas continuas (bill_length, bill_depth, flipper_length y body_mass) junto con la variable categórica de species para hacer el coloreado en los gráficos.

La problemática que se desea abordar y por la cual se quiere aplicar LDA es la siguiente: Se desea poder proyectar la información de las 4 dimensiones en un espacio de menor dimensionalidad, esto con la finalidad de poder observar gráficamente como se están comportando los datos.

Dada la naturaleza del LDA que es capaz de representar los datos en $k-1$ dimensiones, donde k es la cantidad de grupos (en este caso, especies de pingüinos), se plantea el proyectar las aproximaciones en 1D y en 2D para el caso donde se tienen unicamente 2 especies y para el caso donde se tienen 3 especies respectivamente.

Con el objetivo de evidenciar lo difícil que es ver como se están comportando los datos a continuación se muestran los gráficos de dispersión tomando subconjuntos de las variables:

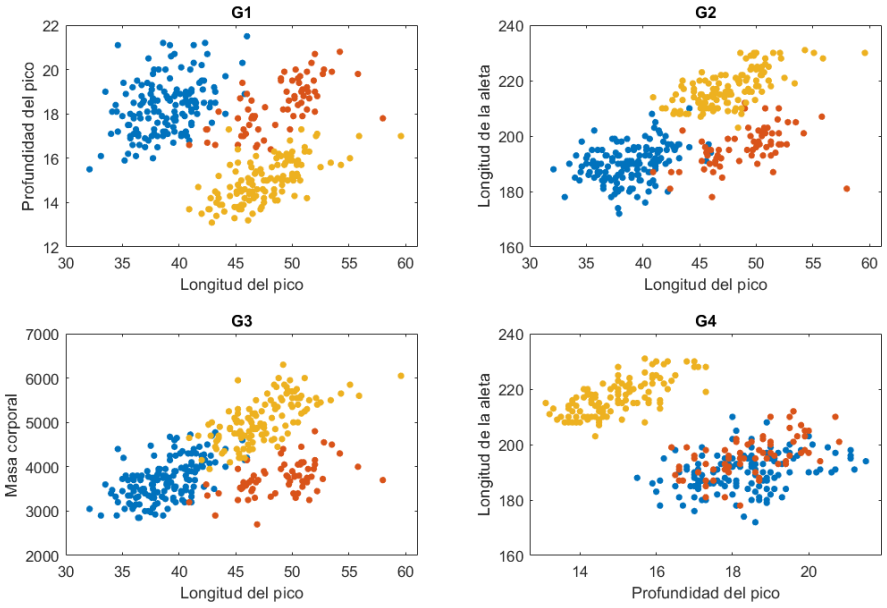

Figure 2: Gráficos de dispersión bivariados.

Se puede observar que para algunos pares de variables, se alcanza a ver una separación clara de los datos por especies, sin embargo, al verlo desde otro par de variables, la cosa se vuelve difusa de distinguir, como es el caso del gráfico G4, en donde se esta mostrando el comportamiento de las variables de la profundidad del pico y el largo de la aleta.

Por otro lado, se hizo la propuesta de modelarlos en 3 dimensiones, para observar como se comportan los datos en el espacio:

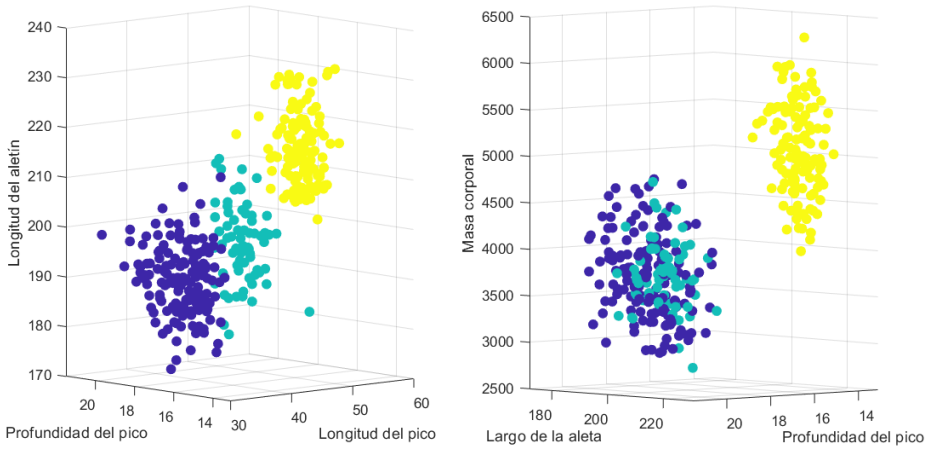

Figure 3: Gráficos de dispersión multivariados.

En el primer gráfico se logra apreciar una separación distinguible entre los grupos de pingüinos dada su visualización mediante las variables de profundidad del pico, longitud del pico y longitud de la aleta. Por otro lado, en el gráfico de la derecha, se observa como un grupo esta perfectamente diferenciado del resto pero los 2 grupos restantes se encuentran sobre puestos uno con el otro, lo que los hace difíciles de separar usando las variables de profundidad del pico, largo de la aleta y masa corporal.

Es por ello que se quiere usar LDA para proyectar la información y reducir la dimensionalidad de los datos para observar mejor su comportamiento en 1 y 2 dimensiones.

2 Metodología

El discriminante lineal es una técnica multivariada de aprendizaje supervisado que permite reducir la dimensionalidad de datos cuantitativos continuos que se encuentren etiquetados por una variable de clase, donde el resultado son $k-1$ combinaciones lineales donde proyectar los datos Johnson and Wichern (2007).

Cabe mencionar que esto se hace bajo supuestos fuertemente ligados a que los conjuntos de datos por grupo siguen una distribución normal multivariada, tienen una matriz de varianzas y covarianzas similar distinta de la matriz nula y que son independientes entre sí.

Para ejecutar la técnica, es necesario tener identificados a la matriz de características X y al vector de etiquetas o clases Y:

$$X = \{x_1, x_2, \ldots, x_n\}$$
$$Y = \{y_1, y_2, \ldots, y_k\}$$

Donde:

- X Es la matriz de datos de tamaño $N \times P$.
- Y Es el vector que contiene las clases de tamaño $N \times 1$.

Una vez identificados los elementos fundamentales para dar inicio a la técnica, se procede a calcular la matriz de dispersión entre clases como sigue:

$$S_w = \sum_{i=1}^{k} (X_i - \mu_i)(X_i - \mu_i)'$$

Donde:

- S_w Es la matriz de dispersión entre clases de tamaño $P \times P$.
- X_i Es la matriz de de características X de la clase i-ésima de tamaño $N_i \times P$.
- μ_i Es el vector de medias de la matrix X_i de tamaño $1 \times P$.
- k Es la cantidad de clases en el conjunto de datos.

Posteriormente, se procede a calcular la dispersión total presente en los datos como sigue:

$$S_t = (X - \mu)(X - \mu)'$$

Donde:

- S_t Es la matriz de dispersión de todo el conjunto de datos sin considerar clases de tamaño $P \times P$.
- X Es la matriz de de características X de tamaño $N \times P$.
- μ Es el vector de medias de la matrix X de tamaño $1 \times P$.

Una vez hecho el calculo de la matriz de dispersión global S_t, se obtiene por diferencia la matriz de dispersión dentro de clases como sigue:

$$S_b = S_t - S_w$$

Donde:

- S_b Es la matriz de dispersión entre clases de tamaño $P \times P$.
- S_t Es la matriz de de dispersión de los datos de tamaño $P \times P$.
- S_w Es la matriz de de dispersión dentro de clases de tamaño $P \times P$.

Una vez determinadas las matrices de dispersión entre clases y dentro de clases, el siguiente paso es maximizar el criterio de Fisher, el cual busca que las clases esten lo más separadas posibles y que la varianza dentro de las clases sea la más pequeña posible como se muestra a continuación:

$$MAX \quad J(W) = \frac{|W^T S_b W|}{|W^T S_w W|}$$

Donde:

- $J(W)$ Es el criterio de Fisher.
- W Es la matriz de vectores propios de tamaño $P \times P$.
- S_b Es la matriz de dispersión entre clases de tamaño $P \times P$.
- S_w Es la matriz de de dispersión dentro de clases de tamaño $P \times P$

Para encontrar los valores de la matriz W que logran maximizar el criterio de fisher, se resuelve como un problema de valores propios generalizados como se muestra:

$$S_b W = \lambda S_w W$$

Donde:

- W Es la matriz de vectores propios de tamaño $P \times P$.
- S_b Es la matriz de dispersión entre clases de tamaño $P \times P$.
- S_w Es la matriz de de dispersión dentro de clases de tamaño $P \times P$
- λ Es el vector de valores propios de tamaño $1 \times P$.

El resultado esperado son P valores propios que indican cuanta varianza acumula cada una de las componentes en el nuevo sistema de coordenadas donde se van a proyectar los datos originales y P vectores propios que serán los ejes sobre los cuales se proyecten los datos.

Para hacer la proyección de los datos centrados en el nuevo sistema, basta con aplicar la siguiente operación matricial con los vectores propios encontrados:

$$Z = XW$$

Donde:

- Z Son las proyecciones de las características X sobre los vectores W.
- X Es la matriz de características de tamaño $N \times P$.
- W es la matriz de vectores propios de tamaño $P \times P$.

Finalmente, se calcula la variabilidad explicada por cada componente mediante los valores propios como sigue:

$$VE = 100 * \frac{\lambda_i}{\sum_{i=1}^{P} \lambda_i}$$

Donde:

- VE Es la variabilidad explicada por el i-ésimo valor propio.
- λ_i Es el i-ésimo valor propio.

Cabe mencionar que los valores propios conservan la siguiente propiedad:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_P$$

Nota importante: Debido a que el discriminante lineal toma en cuenta la cantidad de clases con las que se esta trabajando, logra reducir aun más el número de dimensiones tal que solo se van a proyectar $k-1$ dimensiones como resultado. Pese a que operativamente se generen P valores y vectores propios, la realidad es que solo $k-1$ valores propios van a ser distintos de 0, esto implica que los vectores propios asociados a esos $k-1$ valores propios distintos de 0 van a ser los proyectados y que concentran la información original en una menor dimensionalidad.

Como ejemplo, si se tienen 2 clases y 4 variables, el sistema va a calcular 4 valores propios y 4 vectores propios, pero solo $2-1$ valores propios van a ser distintos de 0, por lo que toda la variabilidad explicada recae en un solo vector propio, por lo que la salida de los datos sera su proyección en 1 dimensión.

Otro ejemplo, si se tienen 3 clases y 4 variables, el sistema va a calcular 4 valores propios y 4 vectores propios, pero solo $3-1$ valores propios van a ser distintos de 0, por lo que toda la variabilidad explicada recae en dos vectores propios, por lo que la salida de los datos sera su proyección en 2 dimensiones.

A continuación, se aplico un breve pre-procesamiento de los datos:

- Se identificaron y eliminaron las filas con valores faltantes (2 filas eliminadas)
- Se guardo en X las variables numéricas de longitud del pico, profundidad del pico, largo de la aleta y masa corporal.
- Se guardo en Y las etiquetas de la especie a la que pertenecen los individuos.

Posteriormente, se aplico LDA para reducir dimensionalidad en 2 casos:

- Para los datos recortados unicamente con 2 especies de pingüinos (Adelie y Chinstrap)
- Para el caso de la base de datos completa que contempla las 3 especies de pingüinos (Adelie, Chinstrap y Gentoo).

3 Resultados para dos clases

A continuación se muestran los resultados obtenidos para los datos de los pingüinos palmer en matlab cuando se trabajan con 2 clases.

3.1 Vectores de medias

Se muestra el vector de medias obtenido para las variables de Longitud del pico, profundidad del pico, longitud de la aleta y la masa corporal para la especie Adelie, Chinstrap y el vector de medias general:

$$\begin{array}{cc} \mu_{Adelie} = & \{38.7914 \quad 18.3464 \quad 189.9536 \quad 3700.662\} \\ \mu_{Chinstrap} = & \{48.8338 \quad 18.4206 \quad 195.8235 \quad 3733.088\} \\ \mu = & \{41.90959 \quad 18.36941 \quad 191.77626 \quad 3710.73059\} \end{array}$$

Se observa la evidente diferencia de las escalas entre las variables, siendo la masa corporal la que tiene valores más altos respecto al resto de variables en su media para los 3 vectores.

3.2 Matriz de dispersión dentro de clases

Aplicando la definición para la matriz de dispersión dentro de las clases descrita previamente, se obtuvo la siguiente matriz:

$$S_w = \begin{bmatrix} 1811.1510 & 356.3029 & 1603.6456 & 144719.76 \\ 356.3029 & 308.4067 & 681.8716 & 65889.04 \\ 1603.6456 & 681.8716 & 9822.5578 & 328426.69 \\ 144719.7580 & 65889.0407 & 328426.6946 & 41439235.25 \end{bmatrix}$$

Se observa como las varianzas y covarianzas de la matriz son mayores en los pares donde esta involucrada la variable de masa corporal (variable 4), mientras que por otro lado, resultan menores en aquellas parejas donde se encuentra involucrada la variable ancho del pico (variable 2).

3.3 Matriz de dispersión entre clases

Aplicando la definición para la matriz de dispersión entre las clases descrita previamente, se obtuvieron las siguientes matrices (la matriz de dispersión global y la matriz de dispersión entre clases):

$$S_t = \begin{bmatrix} 6539.6099 & 391.2542 & 4367.4699 & 159987.5 \\ 391.2542 & 308.6650 & 702.3009 & 66001.9 \\ 4367.4699 & 702.3009 & 11438.0365 & 337350.8 \\ 159987.4658 & 66001.8950 & 337350.7991 & 41488533.1 \end{bmatrix}$$

$$S_b = S_t - S_w = \begin{bmatrix} 4728.4588 & 34.9513 & 2763.8242 & 15267.7078 \\ 34.9513 & 0.2583 & 20.4294 & 112.8543 \\ 2763.8242 & 20.4294 & 1615.4787 & 8924.1045 \\ 15267.7078 & 112.8543 & 8924.1045 & 49297.8596 \end{bmatrix}$$

3.4 Valores y vectores propios

Se construye la matriz que representa el criterio de Fisher a Maximizar:

$$S = S_w^{-1} S_b = \begin{bmatrix} 3.7501 & 0.0277 & 2.1919 & 12.1085 \\ -2.4144 & -0.0178 & -1.4112 & -7.7957 \\ 0.1823 & 0.0013 & 0.1065 & 0.5885 \\ -0.0103 & -0.0001 & -0.0060 & -0.0334 \end{bmatrix}$$

Posteriormente, se calculan los valores y vectores propios como resultado de la resolución de la matriz S como un problema de valores y vectores propios generalizado como sigue:

$$\lambda = \begin{bmatrix} 3.8054 & 0 & 0 & 0 \end{bmatrix}$$

$$W = \begin{bmatrix} 0.8401 & 0.2773 & -0.0506 & -0.0018 \\ -0.5409 & 0.8034 & -0.9958 & -1.0000 \\ 0.0408 & -0.5268 & 0.0766 & 0.0085 \\ -0.0023 & 0.0076 & 0.0041 & 0.0013 \end{bmatrix}$$

Se observa como el primer vector propio de la un mayor peso a la variable de la longitud del pico para hacer la proyección en el nuevo espacio, siendo la masa corporal la que menor peso tiene en este nuevo eje.

Haciendo las cuentas, a continuación se muestra cuanta variabilidad explica cada componente:

$$VE = \begin{bmatrix} 100 & 0 & 0 & 0 \end{bmatrix}$$

Recordando que solo vamos a tener 1 vector característico que proyecta a los datos y 1 solo valor característico distinto de 0 asociado a dicho vector, lo anterior se puede re-escribir dicho vector propio y vector propio como sigue:

$$W = \begin{bmatrix} 0.8401 & -0.5409 & 0.0408 & -0.0023 \end{bmatrix}, \quad \lambda = 3.8054, \quad VE = 100\%$$

El vector W proyecta en un solo nuevo eje a la matriz de características X, dicho vector proyector explica un 100% de la variabilidad de los datos.

3.5 Proyección de las nuevas componentes

A continuación se muestra una pequeña fracción de los datos al ser proyectados sobre el nuevo sistema creado con ayuda de los vectores propios:

ND1
21.44258
22.57016
24.55841
20.28623
21.18188
22.05072

Figure 4: Matriz de caracteristicas X proyectadas sobre 1 nuevo eje).

El nuevo sistema de proyección consta unicamente de un eje en donde se proyectaron los datos, debido a que solo el primer eje resulto explicar el 100% de la variabilidad de los datos.

3.6 Proyección en 1D

Debido a que el primer y único eje de la proyección de los datos es el único que se tiene para cuando hay 2 clases, se procede a proyectarlo en un espacio unidimensional para observar como se comporta:

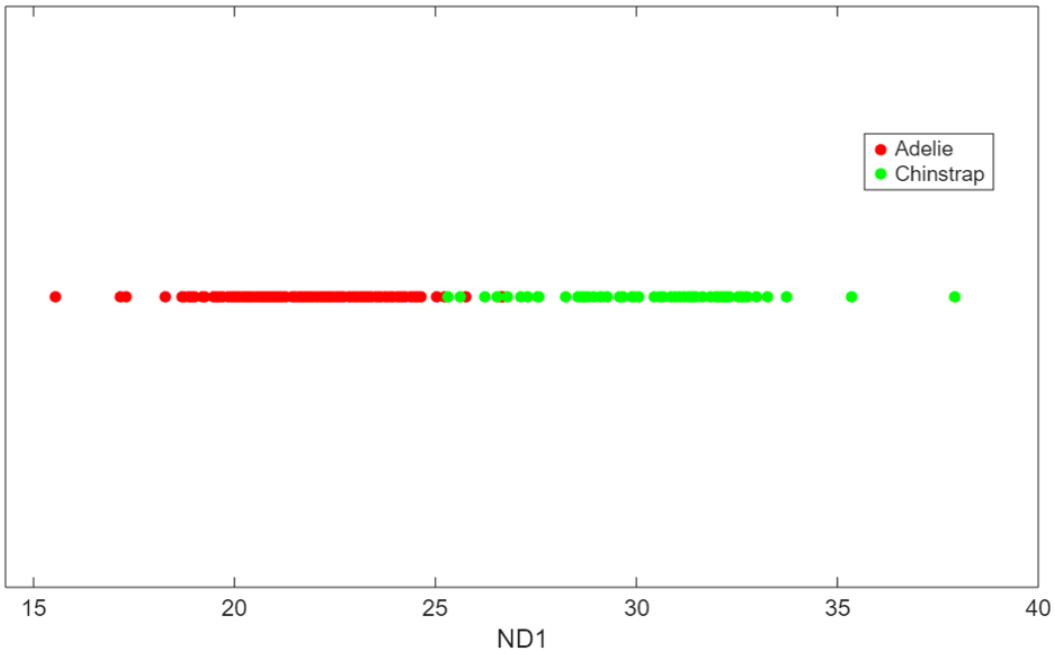

Figure 5: Proyección en 1D sobre el eje 1.

El gráfico de la proyección unidimensional del discriminante lineal muestra un traslape entre las especies, siendo que Algunos Adelie tienen caracteristiacs que los hacen parecer Chinstrap y biceversa.

4 Resultados para tres clases

A continuación se muestran los resultados obtenidos para los datos de los pingüinos palmer en matlab cuando se trabajan con las 3 clases del conjunto de datos.

4.1 Vectores de medias

Se muestra el vector de medias obtenido para las variables de Longitud del pico, profundidad del pico, longitud de la aleta y la masa corporal para la especie Adelie, Chinstrap, Gentoo y el vector de medias general:

$$\begin{array}{cccccc} \mu_{Adelie} = & \{38.7914 & 18.3464 & 189.9536 & 3700.662\} \\ \mu_{Chinstrap} = & \{48.8338 & 18.4206 & 195.8235 & 3733.088\} \\ \mu_{Gentoo} = & \{47.5049 & 14.9821 & 217.1870 & 5076.016\} \\ \mu = & \{43.92193 & 17.15117 & 200.91520 & 4201.75439\} \end{array}$$

Se observa la evidente diferencia de las escalas entre las variables, siendo la masa corporal la que tiene valores más altos respecto al resto de variables.

4.2 Matriz de dispersión dentro de clases

Aplicando la definición para la matriz de dispersión dentro de las clases descrita previamente, se obtuvo la siguiente matriz:

$$S_w = \begin{bmatrix} 2969.8881 & 593.6636 & 3215.733 & 271554.1 \\ 593.6636 & 425.8673 & 1230.383 & 109283.8 \\ 3215.7334 & 1230.3829 & 14953.257 & 608678.3 \\ 271554.1482 & 109283.7765 & 608678.321 & 72443483.2 \end{bmatrix}$$

Se observa como las varianzas y covarianzas de la matriz son mayores en los pares donde esta involucrada la variable de masa corporal (variable 4), mientras que por otro lado, resultan menores en aquellas parejas donde se encuentra involucrada la variable ancho del pico (variable 2).

4.3 Matriz de dispersión entre clases

Aplicando la definición para la matriz de dispersión entre las clases descrita previamente, se obtuvieron las siguientes matrices (la matriz de dispersión global y la matriz de dispersión entre clases):

$$S_t = \begin{bmatrix} 10164.2055 & -864.1738 & 17178.136 & 888506.8 \\ -864.1738 & 1329.8345 & -5528.616 & -254853.2 \\ 17178.1360 & -5528.6161 & 67426.541 & 3350125.9 \\ 888506.8421 & -254853.2018 & 3350125.877 & 219307697.4 \end{bmatrix}$$

$$S_b = S_t - S_w = \begin{bmatrix} 7194.317 & -1457.8374 & 13962.403 & 616952.7 \\ -1457.837 & 903.9672 & -6758.999 & -364137.0 \\ 13962.403 & -6758.9990 & 52473.284 & 2741447.6 \\ 616952.694 & -364136.9782 & 2741447.557 & 146864214.2 \end{bmatrix}$$

4.4 Valores y vectores propios

Se construye la matriz que representa el criterio de Fisher a Maximizar:

$$S = S_w^{-1} S_b = \begin{bmatrix} 3.2165 & -0.4000 & 4.6163 & 176.8832 \\ -12.5988 & 6.5009 & -49.9444 & -2631.6653 \\ 0.9866 & -0.5443 & 4.1384 & 219.9163 \\ 0.0072 & -0.0088 & 0.0611 & 3.4865 \end{bmatrix}$$

Posteriormente, se calculan los valores y vectores propios como resultado de la resolución de la matriz S como un problema de valores y vectores propios generalizado como sigue:

$$\lambda = \begin{bmatrix} 15.0192 & 2.3231 & 0 & 0 \end{bmatrix}$$

$$W = \begin{bmatrix} -0.0846 & -0.9982 & -0.0375 & 0.2808 \\ 0.9930 & -0.0502 & -0.9943 & -0.8166 \\ -0.0825 & 0.0322 & 0.0997 & -0.5043 \\ -0.0012 & 0.0041 & -0.0042 & 0.0062 \end{bmatrix}$$

Se observa como el primer vector propio tiene un mayor peso en la variable de la longitud del pico mientras que en el segundo vector propio, el mayor peso lo recibe la variable de la longitud del pico. Siendo la longitud de la masa corporal la que menos peso tiene en cada uno de los respectivos vectores.

Haciendo las cuentas, a continuación se muestra cuanta variabilidad explica cada componente:

$$VE = \begin{bmatrix} 86.6046 & 13.3954 & 0 & 0 \end{bmatrix}$$

Recordando que solo los primeros 2 vectores característicos proyectan los datos en un nuevo espacio de 2 dimensiones, esto se refleja en la variabilidad explicada con los valores propios, donde el primer nuevo eje explica el 86.6% la variabilidad de los datos y el segundo explica el 13.4%. A raiz de esto, solo importan los primeros 2 valores propios y vectores propios, por lo que se re-escriben los resultados como sigue:

$$W = \begin{bmatrix} -0.0846 & 0.9930 & -0.0825 & -0.0012 \\ -0.9982 & -0.0502 & 0.0322 & 0.0041 \end{bmatrix}, \quad \lambda = [15.0192, 2.3231] \quad VE = [86.6064, 13.3954]\%$$

4.5 Proyección de las nuevas componentes

A continuación se muestra una pequeña fracción de los datos al ser proyectados sobre el nuevo sistema creado con ayuda de los vectores propios significativos obtenidos:

ND1	ND2
-4.331378	-18.81124
-6.130711	-18.77994
-5.660691	-21.56748
-4.149236	-17.28586
-3.079534	-19.22536
-5.052664	-19.07748

Figure 6: Base de datos centrada y proyectada sobre los vectores propios (primeros 6 casos).

Se observa como en los 2 nuevos ejes encontrados, los datos son proyectados hacia valores negativos.

4.6 Proyección en 2D

Debido a que entre la primera y la segunda nueva dimensión se explica un 100% de la variabilidad, a continuación se muestran sus proyecciones en un gráfico de dispersión:

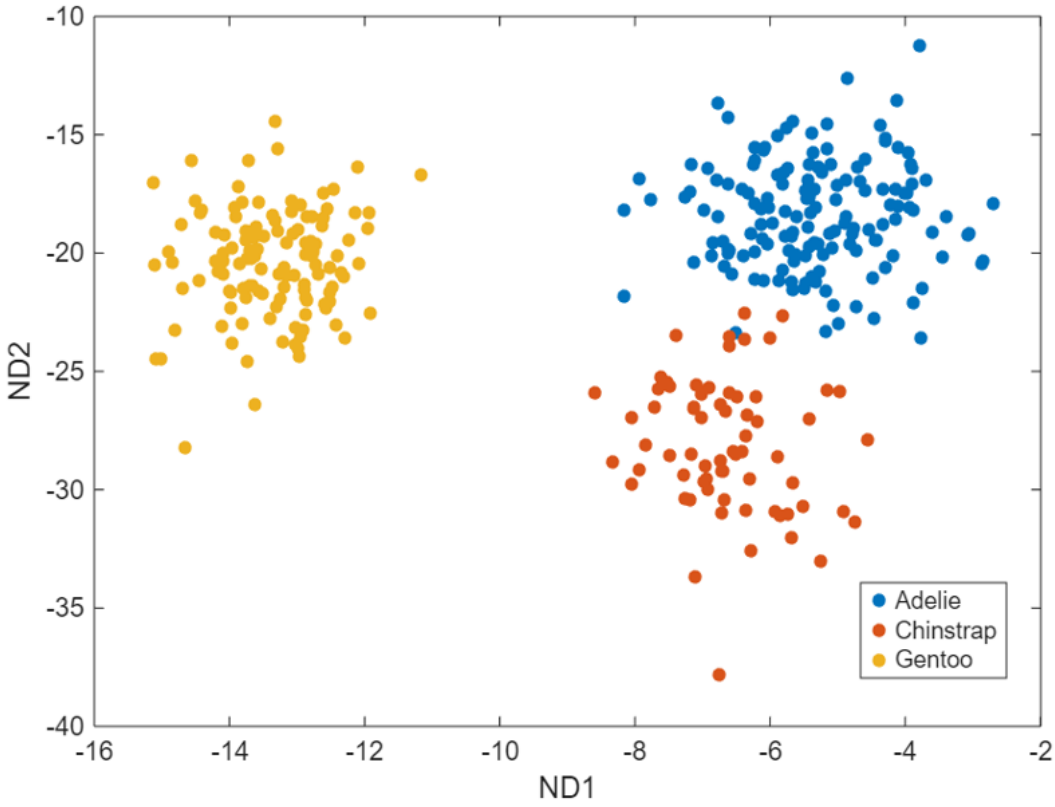

Figure 7: Proyección en 2D sobre los nuevos ejes.

El gráfico muestra una proyección de los datos en donde se observa claramente la separación de la especie Gentoo respecto a las demás, mientras que la separación entre Adelie y Chinstrap es más evidente que en las gráficas mostradas al inicio del documento. El discriminante cumple bien su función de reducir la dimensionalidad y proyectar los datos originales en un nuevo espacio donde sea más evidente la separabilidad de los mismos.

5 Conclusiones

Tras la exploración e implementación del algoritmo para la obtención del LDA, la tecnica reporta resultados bastante buenos al reducir la dimensionalidad de un conjunto de 4 dimensiones al lograr proyectarlo unicamente en 2, manteniendo entendible y explicable quienes son las variables de mayor peso en cada nueva dimensión para lograr dicha separabilidad.

Funciona bien a rasgos generales, puesto que como se observo en los resultados, la especie de Adelie y Chinstrap estan mezcladas entre sí, efecto que se logra atenuar mediante LDA al momento de visualizar los datos.

Por otro lado, aquí se trabajo el caso para 2 y 3 grupos, que son los que finalmente determinan cuantos ejes nuevos van a resultar del proceso, pero el problema se complica cuando se tienen 5 o más grupos, puesto que no es posible visibilizar los 4 tentativos nuevos ejes al mismo tiempo, por lo que seria necesario buscar o proponer alternativas que preserven la consistencia de los datos y puedan ser fácilmente representados en espacios de a lo más 3 dimensiones aunque se tengan 5 o más grupos.

Finalmente, la presencia de normalidad, homocedasticidad e independencia para los datos se asumió para este ejemplo, es importante la verificación de dichos supuestos debido a que la técnica clásica esta supuesta a ellos, en caso de no cumplirse, el proceso puede presentar problemas como matrices singulares no invertibles y de ultimas, que los resultados e inferencias obtenidos carezcan de robustez.

6 Referencias

References

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson, Upper Saddle River, NJ, 6th edition.

7 Anexos

7.1 Implementación de la exploración y el LDA en MATLAB

```
%% Cargar la base de datos en formato CSV %%
ruta = "penguins.csv";
datos = readtable(ruta); % Leer el archivo CSV en una tabla
size(datos); % Ver el tamaño de los datos
% 1 etiqueta de clasificación (Species)
% 3 características categóricas (Isla, sexo y año de observación)
% 4 variables continuas: bill length, bill depth, fliper y bodymass
% Filtrar solo las columnas numéricas y eliminar filas con datos faltantes
indices = find(any(isnan(datos{:, 4:7}), 2)); % Buscar filas con valores NaN en las c
datos1 = datos;
datos1(indices, :) = []; % Eliminar esas filas
% Extraer las variables numéricas y la especie para el gráfico
% Extraer las variables numéricas y la especie para el gráfico
especie = datos1.species; % Columna 'species' con las etiquetas de especies
% Convertir la columna 'especie' en un vector de tipo categorical
especie = categorical(especie); % Convertir a tipo categorical, que gscatter entiende
datos1 = table2array(datos1(:, 4:7)); % Convertir las columnas 4 a 7 a un arreglo num
%% Gráficos descriptivos %%
% Graficar el gráfico de dispersión usando gscatter
% Longitud del pico y ancho del pico
subplot(2,2,1)
gscatter(datos1(:,1), datos1(:,2), especie,"filled");
xlabel('Longitud del pico'); % Etiqueta del eje x
ylabel('Profundidad del pico'); % Etiqueta del eje y
title('G1'); % Título del gráfico
% Longitud del pico y longitud de la aleta %
subplot(2,2,2)
gscatter(datos1(:,1), datos1(:,3), especie,"filled");
xlabel('Longitud del pico'); % Etiqueta del eje x
ylabel('Longitud de la aleta'); % Etiqueta del eje y
title('G2'); % Título del gráfico
% Longitud del pico y masa corporal
subplot(2,2,3)
gscatter(datos1(:,1), datos1(:,4), especie,"filled");
xlabel('Longitud del pico'); % Etiqueta del eje x
ylabel('Masa corporal'); % Etiqueta del eje y
title('G3'); % Título del gráfico
```

```
% Ancho del pico y Longitud de la aleta
subplot(2,2,4)
gscatter(datos1(:,2), datos1(:,3), especie,"filled");
xlabel('Profundidad del pico'); % Etiqueta del eje x
ylabel('Longitud de la aleta'); % Etiqueta del eje y
title('G4'); % Título del gráfico
%% Exploración en 3D %&
% Crear un gráfico 3D
subplot(1,2,1)
scatter3(datos1(:,1), datos1(:,2), datos1(:,3), 50, especie, 'filled');
% Ajustar etiquetas y título
xlabel('Longitud del pico');
ylabel('Profundidad del pico');
zlabel('Longitud del aletín');
title('');
subplot(1,2,2)
scatter3(datos1(:,2), datos1(:,3), datos1(:,4), 50, especie, 'filled');
% Ajustar etiquetas y título
xlabel('Profundidad del pico');
ylabel('Largo de la aleta');
zlabel('Masa corporal');
title('');
% Mostrar la leyenda de colores
% Guardar como archivo PNG

%% FUNCIÓN LDA %%
function resu = LDA(X, Y)
        % Asegúrate de que X es una matriz
        X = double(X);
        % Número de variables (columnas)
        nc = size(X, 2);
        % Clases únicas
        clas = unique(Y);
        nr = length(clas); % número de clases
        % Inicializar matrices
        medias = zeros(nr, nc);
        Sw = zeros(nc, nc);
        for i = 1:nr
                % Subconjunto para la clase k-ésima
                mini1 = X(Y == clas(i), :);
```

```
78                  % Vector de medias clase k
79                  medias(i, :) = mean(mini1, 1);
80                  % Matriz de covarianza clase k
81                  centered = mini1 - medias(i, :);
82                  Sn = centered' * centered;
83                  % Acumular la varianza dentro de clase
84                  Sw = Sw + Sn;
85          end
86          % Vector de medias global
87          m = mean(X, 1);
88          % Dispersión total
89          centered_global = X - m;
90          St = centered_global' * centered_global;
91          % Matriz de dispersión entre clases
92          Sb = St - Sw;
93          % Matriz S = inv(Sw) * Sb
94          S = inv(Sw) * Sb;
95          % Eigenvalores y eigenvectores
96          [V, D] = eig(S);
97          [eigenvalues, idx] = sort(diag(D), 'descend');
98          V = V(:, idx);
99          % Varianza explicada
100         VE = round(100 * eigenvalues / sum(eigenvalues), 4);
101         % Filtrar vectores con varianza explicada significativa
102         DS = VE > 0.0001;
103         SV = V(:, DS);
104         % Proyección de los datos
105         Z = X * SV;
106         % Renombrar las columnas
107         for i = 1:size(Z, 2)
108                 colnames{i} = ['ND', num2str(i)];
109         end
110         % Armar resultado
111         resu.varianza = VE;
112         resu.coeficientes = SV;
113         resu.proyecciones = array2table(Z, 'VariableNames', colnames);
114 end
115 %% Caso para 2 grupos (Adelie y Chinstrap)
116 X1 = datos1(Y ~= "Gentoo",:);
117 Y1 = especie(Y ~= "Gentoo");
```

```
%% Aplicar el LDA %%
salida1 = LDA(X1,Y1)
% Varianza explicada
salida1.varianza
% Vectores propios
salida1.coeficientes
% Proyecciones
Z1 = table2array(salida1.proyecciones);
%% Crear una dispersión en 1D (todos los puntos con la misma Y)
y = zeros(size(Z1, 1), 1); % Todos los puntos en Y=0
gscatter(Z1, y, Y1, 'rgb', '.', 15);
xlabel('ND1');
yticks([]); % Eliminar marcas del eje Y
ylabel('');
title('');
%% Caso para 3 grupos (Adelie, Chinstrap y Gentoo)
X2 = datos1;
Y2 = especie;
%% Aplicar el LDA %%
salida2 = LDA(X2,Y2)
% Varianza explicada
salida2.varianza
% Vectores propios
salida2.coeficientes
% Proyecciones
Z2 = table2array(salida2.proyecciones)
%% Gráficar en 2D %%
gscatter(Z2(:,1), Z2(:,2), Y2,"filled");
xlabel('ND1'); % Etiqueta del eje x
ylabel('ND2'); % Etiqueta del eje y
title(''); % Título del gráfico
```