.UBAfiuba
FACULTAD DE INGENIERÍA

Trabajo Práctico 2

Aprendizaje Estadístico
Segundo cuatrimestre de 2020

Clasificación

Romero, Adrián

Índice

1. Introducción

El objetivo del presente trabajo es estudiar y comparar las clasificaciones que se obtienen al utilizar Linear Discriminant Analyisis (LDA) y Quadratic Discriminant Analysis (QDA). Para ello se clasificarán las observaciones de flores del set de datos iris en 3 grupos: setosa, versicolor y virginica, a partir de mediciones realizadas sobre la longitud y el ancho de los pétalos y de los sépalos de las flores.

El desarrollo del trabajo estará dividido en tres secciones:

- **Explicación de LDA y QDA**: se explicarán brevemente los clasificadores para poder comprender sobre qué fundamentos se basan las secciones siguientes.
- **Análisis inicial**: se analizará el set de datos que se utilizará a lo largo del trabajo y el cumplimiento de las hipótesis que suponen los modelos de LDA y QDA. Además se buscará qué features son mejores para lograr una clasificación óptima.
- **Clasificación y análisis de métricas**: se utilizarán distintos clasificadores para clasificar a las observaciones y para cada uno de ellos se evaluarán las métricas de accuracy, precision, recall y se utilizarán técnicas de resampleo como validación cruzada (cross-validation).

Finalmente, en la conclusión, se resumirán los aspectos más destacados del trabajo.

2. Explicación de LDA y QDA

Los algoritmos de LDA y QDA son algoritmos de aprendizaje supervisado que buscan resolver un problema de clasificación.

Supongamos que tenemos k clases distintas, un conjunto de observaciones con sus etiquetas y una nueva observación x$\in \mathbb{R}^p$ que buscamos clasificar.

Se define la variable aleatoria G = clase de x. De manera que P(G = j)=π_j con j = 1,2,...,k es la probabilidad a priori de que la observación x provenga de la clase j. Se puede estimar como: la proporción de observaciones cuya etiqueta es j.

Si suponemos conocida a la distribución de las variables aleatorias X|G=j, es decir que suponemos conocer la distribución de una observación sabiendo de que clase proviene, entonces podemos calcular P(G = j| X = x) la probabilidad a posteriori de que que la observación sea de la clase j con la regla de Bayes, pues:

$$\text{P(G = j| X= x)} = \frac{\pi_j \cdot f_{X|G=j}(x)}{\sum_{n=1}^{k} f_{X|G=n}(x)}$$

Y será posible asignar a x la clase cuya probabilidad a posteriori sea mayor, según la regla de clasificación de Bayes.

Es decir, asignaremos x a la clase i si: P(G = i|X = x) >P(G = j|X = x) $\forall$ j = 1,2,...,k $\neq$ i (se observa que al plantear esta inecuación los denominadores se simplifican)

Si además X|G=j $\sim$N$_p(\mu_j, \Sigma_j)$, entonces la condición anterior puede escribirse como: asignaremos x a la clase i si:

$$Q_i(x) = -\tfrac{1}{2}\log|\Sigma_i| - \tfrac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \log\pi_i > Q_j(x) \ \forall \text{ j = 1,2,...,k} \neq \text{i}$$

En particular, si suponemos que las matrices de covarianza Σ_i son iguales para cada clase, entonces la condición se simplifica y se asignará x a la clase i si:

$$L_i(x) = \mu_i^T\Sigma^{-1}(x - \tfrac{1}{2}\mu_i) + \log\pi_i > L_j(x) \ \forall \text{ j = 1,2,...,k} \neq \text{i}$$

Las funciones dadas por $L_i(x) - L_j(x)$ y $Q_i(x) - Q_j(x)$ se llaman funciones discriminantes y el conjunto de puntos $Q_{ij} = \{x \in \mathbb{R}^p : Q_i(x) - Q_j(x) = 0\}$ y $L_{ij} = \{x \in \mathbb{R}^p : L_i(x) - L_j(x) = 0\}$ conforman una frontera de decisión. A un lado, la probabilidad de que la una observación sea de la clase i es mayor a que sea de la clase j y del otro ocurre lo opuesto. Estas fronteras de decisión son funciones cuadráticas en x en el caso de QDA y funciones lineales en x en el caso de LDA.

Se observa entonces que para lograr predicciones sobre observaciones es necesario encontrar:

- $\hat{\mu_i}$, una estimación para la media de X|G=i.
- $\hat{\Sigma_i}$, una estimación para la matriz de covarianza de X|G=i (o simplemente $\hat{\Sigma}$ en el caso de LDA).
- $\hat{\pi_i}$, una estimación para la probabilidad a priori de pertenecer a la clase i.

Por último, la regla de clasificación de Bayes utilizando LDA y QDA logra encontrar las combinaciones lineales de los componentes de x, tales que la proyección de los datos de entrenamiento sobre esas combinaciones tienen varianza entre grupos máxima y varianza intra grupo mínima. Esto es claramente bueno para lograr una clasificación óptima pues busca aquellas direcciones, tales que al proyectar los datos, las observaciones que pertenecen a clases distintas queden proyectadas separadas y las observaciones que pertenecen a una misma clase queden proyectadas juntas.

3. Análisis Inicial

El set de datos que se utilizará en este proyecto es el conocido set iris. Consta de 150 observaciones de flores a las cuales se le midieron:

- Largo de sépalo (en cm.)
- Ancho de sépalo (en cm.)
- Largo de pétalo (en cm.)
- Ancho de pétalo (en cm.)

Hay 50 observaciones de la clase setosa, 50 de la clase versicolor y 50 de la clase virginica. A partir de esta información podemos estimar π_i, la probabilidad a priori de pertenecer a la clase i como:

$$\hat{\pi_1} = \hat{\pi_2} = \hat{\pi_3} = \frac{50}{150} = \frac{1}{3}$$

Uno de los supuestos de los algoritmos de QDA y LDA es que las x, condicionadas a que pertenecen a cierta clase, provienen de distribuciones normales p-variadas. Una forma de verificar este supuesto es realizar un qqplot para cada feature y según clase, esperando que los cuantiles muestrales normalizados se parezcan lo mas posible a los cuantiles teóricos de una distribución normal estándar. Para realizar esto primero se deben normalizar los valores, restándole la media y dividiendo por el desvío. Los qqplots obtenidos son los siguientes:

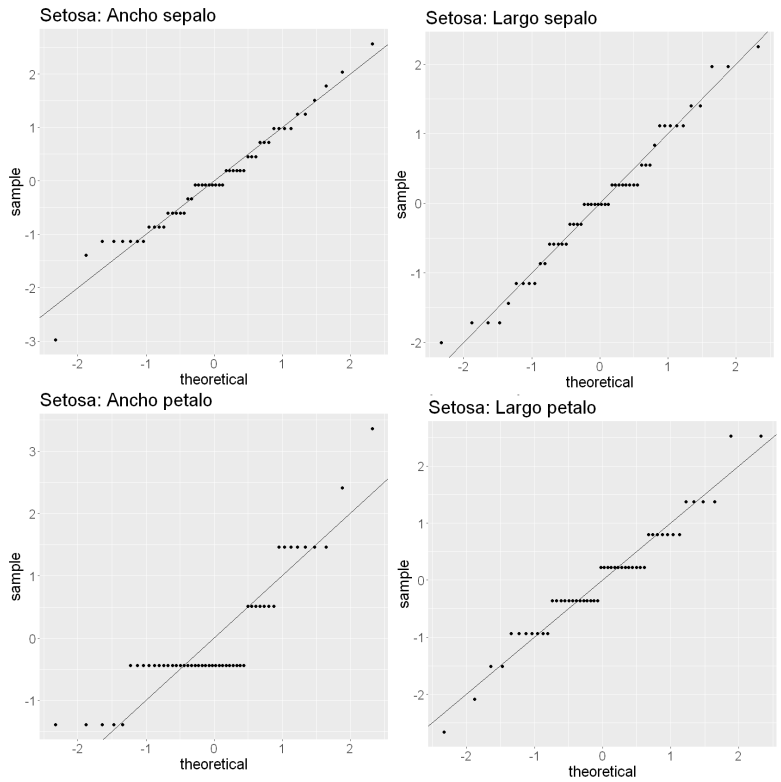

Figura 1: QQplots: ancho y largo de sépalo y pétalo de las flores de clase setosa

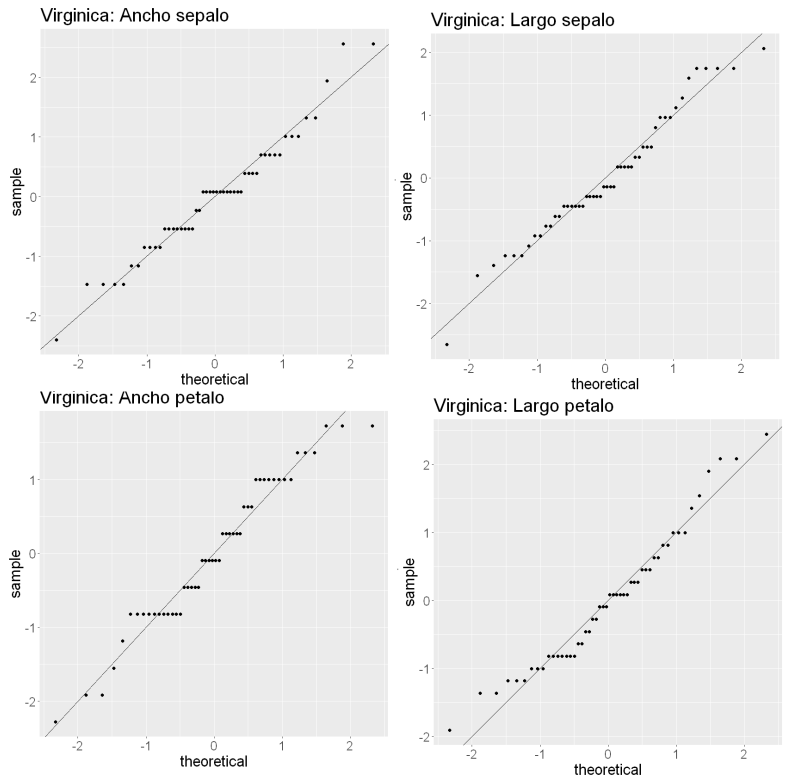

Figura 2: QQplots: ancho y largo de sépalo y pétalo de las flores de clase virginica

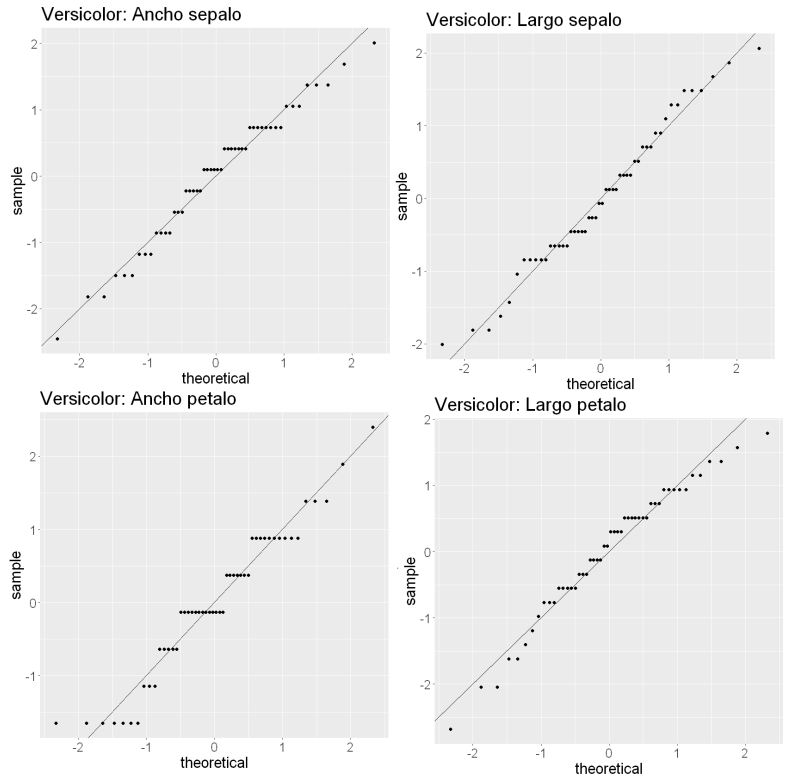

Figura 3: QQplots: ancho y largo de sépalo y pétalo de las flores de clase versicolor

La primera observación que resalta a primera vista es que los valores de las mediciones parecen estar discretizados. Sin embargo, dado que las variables que se están teniendo en cuenta son numéricas y continuas, el hecho de que ocurra esto puede deberse a limitaciones en la precisión de los instrumentos de medición y/o a redondeos efectuados.

Se puede observar que el supuesto de normalidad de las variables parecerá cumplirse pues los qqplots muestran que dentro de cada clase, los features se distribuyen de manera aproximadamente normal.

Con el objetivo de visualizar los datos de entrenamiento, se realizan scatter plots de a pares de features y boxplots e histogramas que permitan ver diferencias entre las clases:

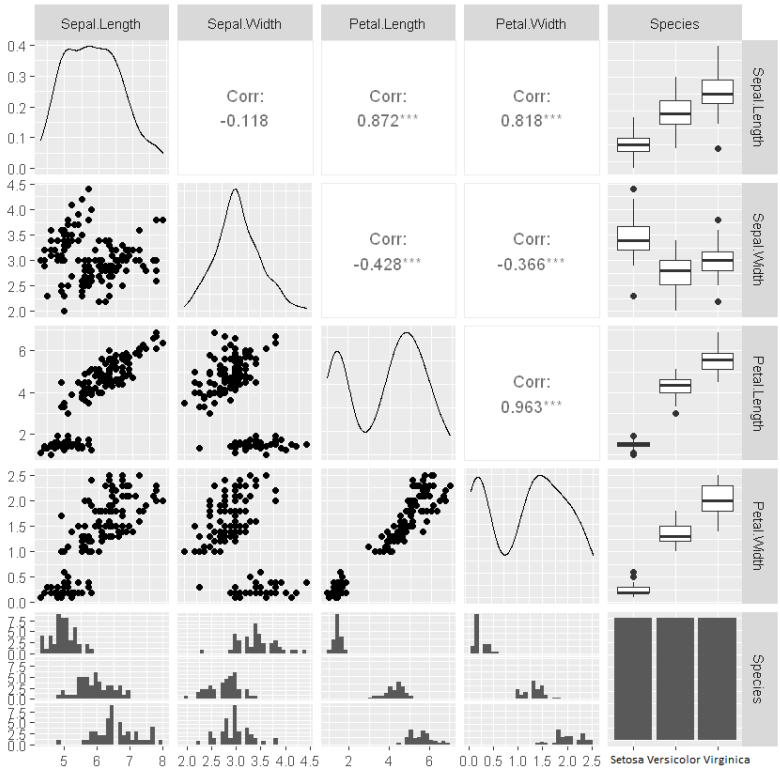

Figura 4: Scatter plots, histogramas y boxplots

Algunas observaciones que se desprenden de los gráficos anteriores son:

- Los boxplots y los histogramas (en la última columna y en la última fila) que provienen de los features longitud de pétalo y ancho de pétalo muestran que son variables que separan las clases de una buena manera. Principalmente la setosa parece ser una clase con longitud y ancho de pétalo mucho menor que las otras y además con baja variabilidad entre 0 y 2 cm para la longitud del pétalo y entre 0 y 0,5 cm para el ancho del pétalo. Es posible que si se tiene en cuenta alguno de estos features al realizar clasificaciones por LDA y QDA esta clase tenga una muy buena precision y recall.

- De manera contraria, los boxplots y los histogramas que provienen de los features longitud de sépalo y ancho de sépalo no son tan buenos separando las instancias de clases diferentes. En particular el largo del sépalo es muy similar entre las flores del tipo versicolor y las de virginica.

- En general estos features parecen tomar valores similares para las instancias de las clases versicolor y virginica, esto puede hacer que estas dos clases sean las mas difíciles de diferenciar

pues es probable que no existe una recta o función cuadrática que separe perfectamente a las instancias de estas clases.

Para finalizar el análisis inicial y complementar lo mencionado en el párrafo anterior, se muestran algunos de los scatter plots anteriores coloreando con distinto color a las instancias de diferentes clases:

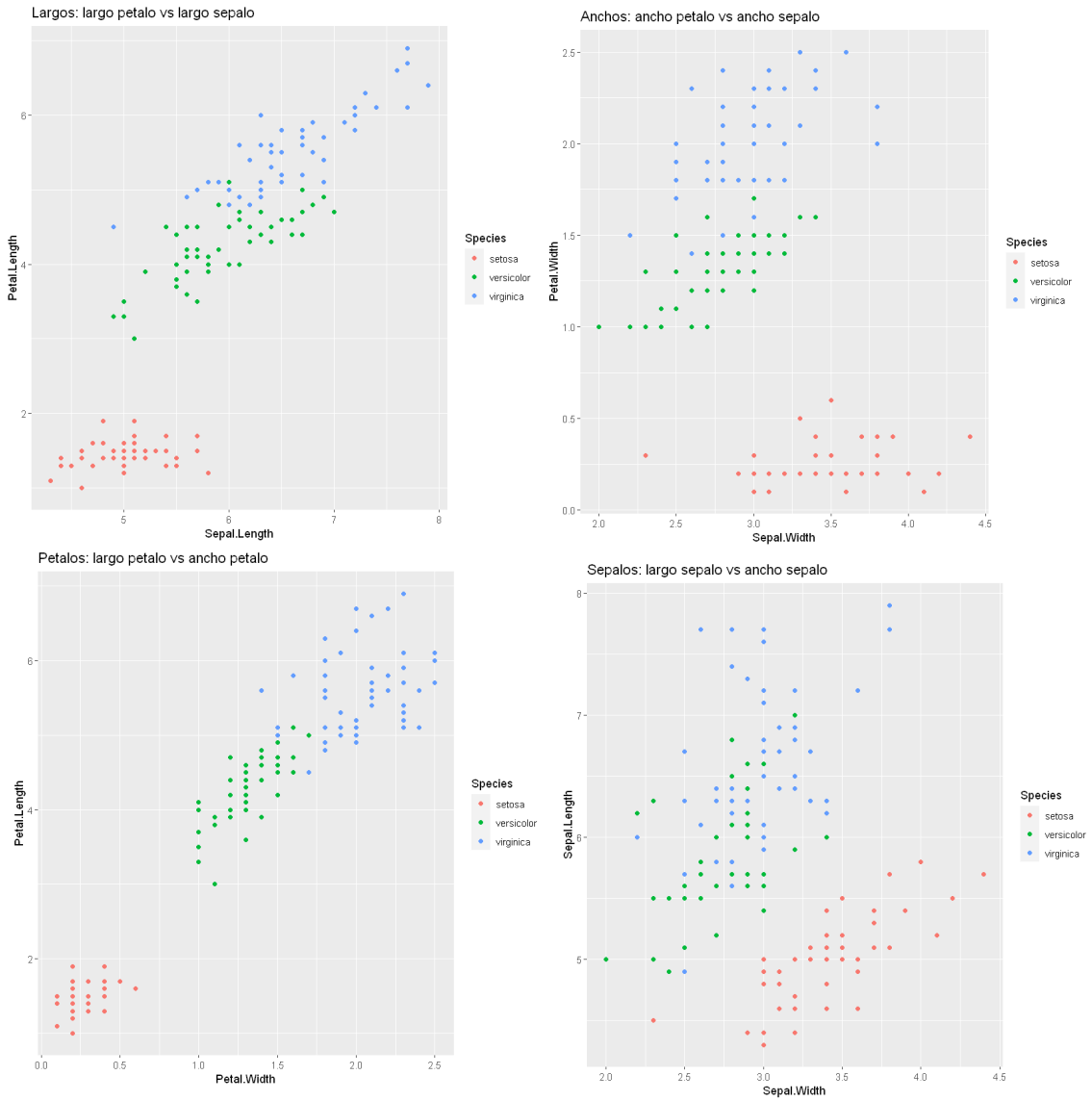

Figura 5: Scatter plots

Se observa que en el scatter plot de pétalos la separación en clases es buena y que además las instancias de una misma clase están cercanas entre sí. Por otro lado, en el scatter plot de sépalos la instancias de versicolor y virginica están entremezcladas. Por otro lado, en todos los gráficos, las instancias de la clase setosa son las mas alejadas y poco dispersas, particularmente cuando se considera el ancho del pétalo.

Al considerar solo dos features, se puede observar que las muestras dentro de cada clase, parecen venir de una distribución normal bivariada pues se encuentran distribuidas de forma elipsoidal.

Es importante el hecho de que LDA y QDA utilizan como supuesto el hecho de que los feature se distribuyan normal solo al condicionarla a una clase, pues por ejemplo, el qqplot del features ancho de sépalo (sin condicionar a ninguna clase) muestra que no sigue una distribución normal:

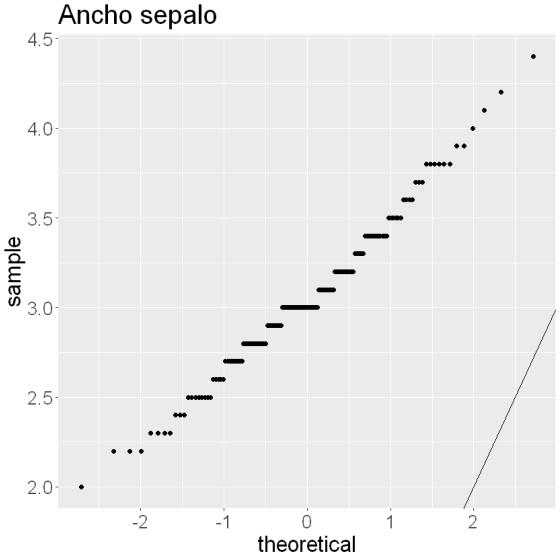

Figura 6: Qqplot ancho sépalo

4. Clasificación

En esta sección se buscarán clasificadores de LDA y QDA utilizando primero sólo 2 de los features: ancho del sépalo y ancho del pétalo y luego se buscarán clasificadores que utilizan todos los features. Se comprarán también las métricas que se obtienen a partir de las predicciones que se realizan con estos clasificadores.

4.1. Clasificación utilizando LDA y 2 features

Como se mencionó en el párrafo anterior, en esta subsección se construye un clasificador de LDA utilizando los features ancho de sépalo y ancho de pétalo y luego se los hace clasificar. En esta subsección y en las siguientes llamaremos X_1 y X_2 al ancho del sépalo y al ancho del pétalo respectivamente, $X = (X_1, X_2)^T$ y las clases 1,2 y 3 se corresponderán con las clases setosa, virginica y versicolor respectivamente.

Un estimador para la matriz de covarianza de las clases (que en el caso de LDA se supone que es la misma para todas las clases) es:

$$\hat{\Sigma} = \frac{1}{n-K}(\hat{\Sigma}_1(n_1 - 1) + \hat{\Sigma}_2(n_2 - 1) + \hat{\Sigma}_3(n_3 - 1))$$

Donde:

- n_1 n_2 y n_3 son la cantidad de observaciones de la clase 1,2 y 3 respectivamente.
- $n = n_1 + n_2 + n_3$ la cantidad de observaciones totales.
- $\hat{\Sigma}_i$ es la estimación insesgada de la matriz de covarianza para la clase i
- K es la cantidad de clases, en este caso 3

Es decir, $\hat{\Sigma}$ es un promedio ponderado de las matrices de covarianza de cada clase.
Además se obtiene $\hat{\mu}_i$ como:

$$\hat{\mu}_i = \overline{X|G_{=i}}$$

La estimación de estos parámetros resultó:

$$\hat{\Sigma}_1 = \begin{bmatrix} 0{,}143 & 0{,}009 \\ 0{,}009 & 0{,}011 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 0{,}104 & 0{,}047 \\ 0{,}047 & 0{,}075 \end{bmatrix} \qquad \hat{\Sigma}_3 = \begin{bmatrix} 0{,}098 & 0{,}041 \\ 0{,}041 & 0{,}039 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 0{,}115 & 0{,}0327 \\ 0{,}032 & 0{,}041 \end{bmatrix} \qquad \hat{\mu}_1 = \begin{bmatrix} 3{,}428 \\ 0{,}246 \end{bmatrix} \qquad \hat{\mu}_2 = \begin{bmatrix} 2{,}974 \\ 2{,}026 \end{bmatrix} \qquad \hat{\mu}_3 = \begin{bmatrix} 2{,}770 \\ 1{,}326 \end{bmatrix}$$

Y recordando del análisis inicial que $\hat{\pi}_i = \frac{1}{3}$, es posible armar las fronteras de decisión calculando $L_i(x) - L_j(x) = 0$. Esto resulta en las siguientes ecuaciones lineales en x:

- $L_1(x) - L_2(x) = 0 \iff 20{,}52x_1 - 58{,}53x_2 = -0{,}78$
- $L_2(x) - L_3(x) = 0 \iff 3{,}8x_1 - 19x_2 = -22$
- $L_3(x) - L_1(x) = 0 \iff -16{,}7x_1 + 38{,}8x_2 = -21{,}26$

Graficando estas rectas en un scatter plot de x_1 vs x_2 se obtiene:

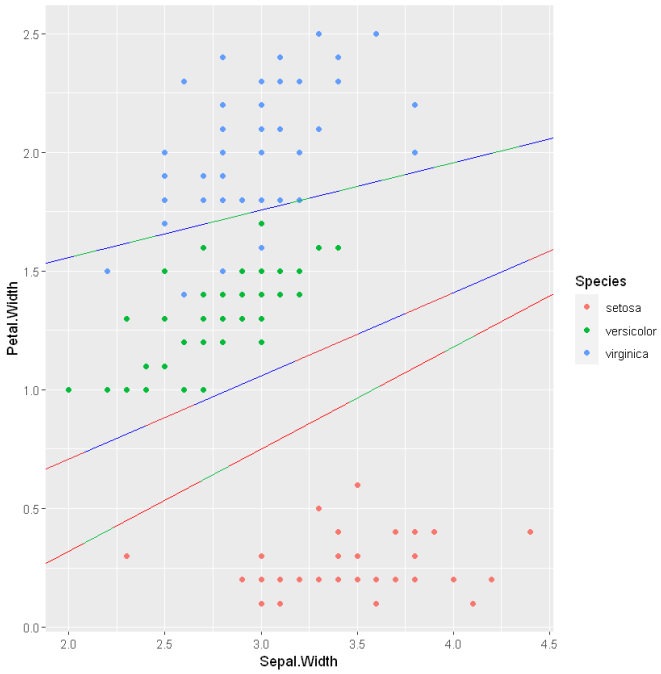

Figura 7: Fronteras de decisión LDA ancho sepalo y ancho petalo

En el gráfico anterior las rectas de frontera de decisión están coloreadas de manera que indiquen cuales clases separan. Por ejemplo, la recta azul y verde separa a las clases azul y verde (virginica y versicolor respectivamente). La regla de clasificación para una observación será entonces:

- Setosa si se encuentra por debajo de la recta verde-roja.
- Virginica si se encuentra sobre la recta verde-azul.
- Versicolor si se encuentra entre la recte verde-roja y la recta verde-azul

De esta manera las regiones que quedan definidas por estas fronteras son las que se ven en el siguiente gráfico:

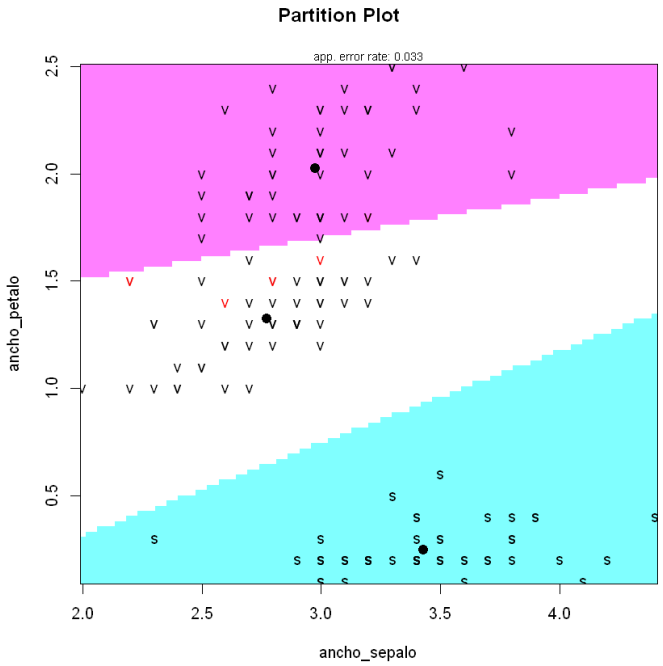

Figura 8: LDA regiones de clasificación

Según esta regla de clasificación, la nueva observación $x_0 = (3.5\ ,\ 1.75)^T$ sería clasificada como versicolor.

Además los valores de $L_i(x_0)$ son los siguientes:

- $L_1(x_0) = 27{,}007$
- $L_2(x_0) = 56{,}809$
- $L_3(x_0) = 57{,}747$

Por lo que se corrobora que la clasificación de esta observación sería versicolor.

Si en vez de clasificar una nueva observación con esta regla, clasificáramos las observaciones con la que se construyo el modelo obtendríamos los siguientes resultados:

	clases_predichas_lda_anchos		
	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	49	1
virginica	0	4	46

Figura 9: Predicciones sobre el set de entrenamiento

De donde se desprenden las siguientes observaciones:

- El accuracy sobre el set de entrenamiento es de $\frac{50+49+46}{150} = 0{,}96$
- Existen 4 instancias de virginica que son clasificadas como versicolor y 1 instancia de versicolor que es clasificada como virginica. Contrastantemente, todas las instancias de setosa son clasificadas correctamente. Esto se puede ver claramente en el scatter plot anterior.

- Para la clase setosa:

  La precision es de $\frac{50}{50+0+0} = 1$ pues de las 50 flores que fueron predichas como setosa, todas efectivamente eran setosa.

  El recall es de $\frac{50}{50+0+0} = 1$ pues de las 50 flores que eran setosa, todas fueron clasificadas como setosa.

- Para la clase versicolor:

  La precision es de $\frac{46}{0+1+46} = 0{,}97$ pues de las 47 flores que fueron predichas como versicolor, 46 efectivamente eran versicolor.

  El recall es de $\frac{46}{0+4+46} = 0{,}92$ pues de las 50 flores que eran versicolor, 46 fueron clasificadas como versicolor correctamente.

- Para la clase virginica:

  La precision es de $\frac{49}{0+49+4} = 0{,}92$ pues de las 53 flores que fueron predichas como virginica, 49 efectivamente eran virginica.

  El recall es de $\frac{49}{0+49+1} = 0{,}98$ pues de las 50 flores que eran virginica, 49 fueron clasificadas como virginica correctamente.

Las métricas de precision y recall son útiles y pueden servir en casos diferentes.

Por ejemplo, si por algún motivo, se necesita que todas las observaciones que pertenezcan a la clase virginica sean clasificadas correctamente querremos tener un buen recall, sin importar tener un bajo precision, como sucede en este caso. Es decir, se estaría priorizando tener pocos falsos negativos (instancias de la clase virginica mal clasificadas) por sobre tener pocos falsos positivos (instancias clasificadas como virginica cuando en realidad no lo son).

Por otro lado, también se podría necesitar, por ejemplo, que de las instancias clasificadas como virginica, la mayor cantidad posible esté bien clasificada. En este caso se prioriza tener pocos falsos positivos por sobre tener pocos falsos negativos, es decir un buen precision sobre un buen recall.

En general, se necesita construir el modelo clasificador para clasificar nuevas observaciones y por lo tanto, las métricas de precision y recall calculadas anteriormente pueden no ser representativas de como se comportarían en esta situación. Para solucionar este problema se pueden hacer predicciones de cross-validation leave one out. Éstas se obtienen entrenando el clasificador con todos los datos excepto con uno y luego clasificando esta última observación con el modelo entrenado. La estimación del error de predicción obtenido de esta manera tiene la ventaja de ser insesgada.

Los resultados obtenidos fueron los siguientes:

	clases_predichas_lda_anchos_LOU		
	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	4	46

Figura 10: Predicciones sobre el set de entrenamiento utilizando leave one out cross-validation

	Setosa	Versicolor	Virginica
Precision	1	0.92	0.95
Recall	1	0.96	0.92

Accuracy: 0.96

Se observa que son ligeramente inferiores a las obtenidas anteriormente, pero mas representativas de como se comportarían al clasificar datos que no se observaron para construir el modelo.

4.2. Clasificación utilizando QDA y 2 features

Para realizar esta clasificación por QDA se utilizaran los dos mismos features que en la subsección anterior: ancho del sépalo y ancho del pétalo.

Se utilizan las matrices de covarianza estimadas $\hat{\Sigma}_i$ que se utilizaron en la subsección anterior para hallar $\hat{\Sigma}$ y también las mismas probabilidades a priori estimadas $\hat{\pi}_i$ y medias estimadas $\hat{\mu}_i$

La fronteras de decisión dadas por $Q_i(x) - Q_j(x) = 0$ fueron las siguientes:

- $Q_1(x) - Q_2(x) = 0 \iff 6{,}15x_1^2 + 3{,}89x_1 - 4{,}68x_1x_2 + 6{,}43 - 76{,}59x_2^2 - 22{,}28x_2 = 0$
- $Q_2(x) - Q_3(x) = 0 \iff 4{,}6x_1^2 - 21{,}2x_1x_2 - 3{,}78x_1 + 27{,}1x_2^2 + 9{,}42x_2 - 15{,}18 = 0$
- $Q_3(x) - Q_1(x) = 0 \iff 10{,}75x_1^2 + 0{,}11x_1 - 25{,}88x_1x_2 - 49{,}49x_2^2 - 12{,}86x_2 - 8{,}75 = 0$

Se observa que son funciones cuadráticas en x y no lineales como en el caso de LDA. Se pueden ver en el siguiente gráfico:

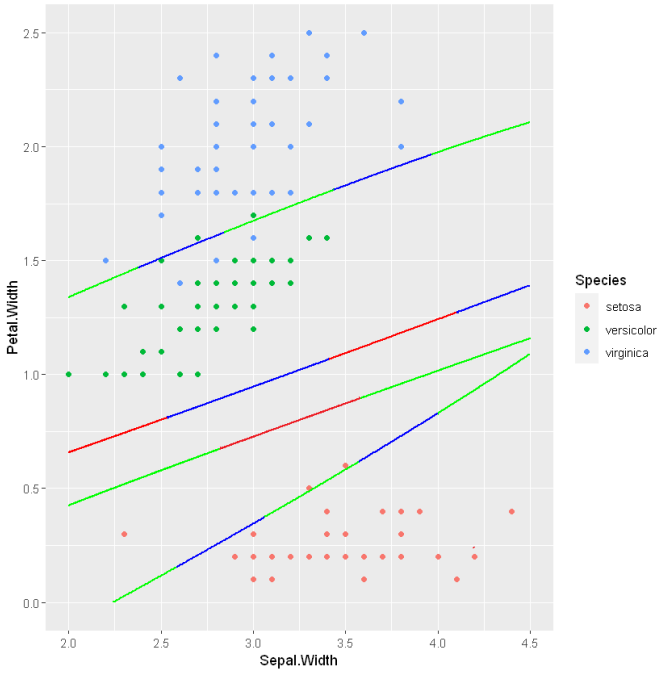

Figura 11: QDA

Se destacan los siguientes aspectos:

- Las clases virginica y versicolor están separadas por una elipse que encierra a las observaciones de la clase versicolor.
- Las curvas azul-roja y roja-verde son hipérbolas que se asemejan mucho, a las rectas azul-roja y roja-verde halladas en LDA (al menos en el rango ploteado).
- Existen instancias de versicolor que LDA clasifica correctamente y QDA no
- Existen instancias de virginica que QDA clasifica correctamente y LDA no

De esta manera las regiones que quedan definidas por estas fronteras son las que se ven en el siguiente gráfico:

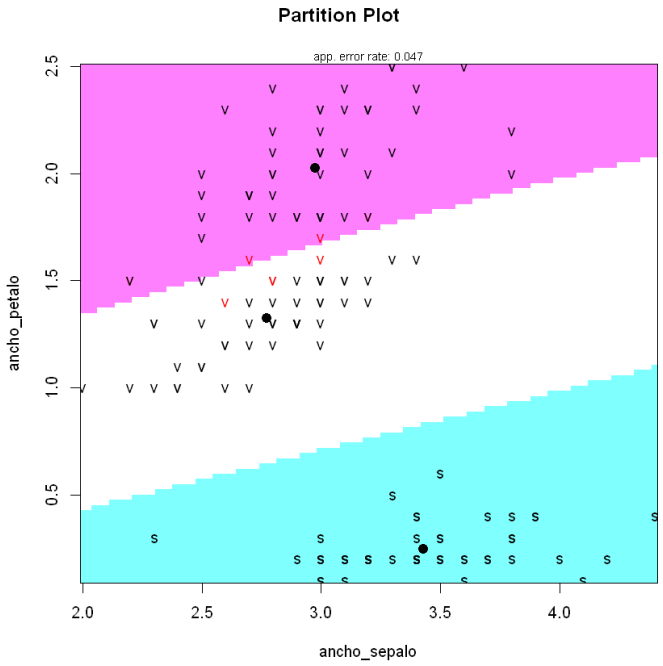

Figura 12: QDA regiones de clasificación

Utilizando este clasificador, clasificaríamos a la nueva observación $x_0 = (3.5\ ,\ 1.75)^T$ también como versicolor pues los valores de $Q_i(x_0)$ son los siguientes:

- $Q_1(x_0) = -104{,}881$
- $Q_2(x_0) = -2{,}335$
- $Q_3(x_0) = -1{,}065$

Además esto se corrobora pues es un punto dentro de la elipse.

Analizando las métricas de accuracy, precision y recall de las predicciones de CV-leave one out se obtuvieron los siguientes resultados:

```
                QDA_anchos_predicciones_LOU
                 setosa versicolor virginica
setosa               50          0         0
versicolor            0         45         5
virginica             0          3        47
```

Figura 13: Predicciones sobre el set de entrenamiento utilizando QDA y leave one out cross-validation

	Setosa	Versicolor	Virginica
Precision	1	0.93	0.90
Recall	1	0.90	0.94

Accuracy: 0.95

Se destacan las siguientes observaciones:

- La precision y el recall de setosa es el mismo en QDA y en LDA
- La precision de versicolor es mayor en QDA que en LDA
- El recall de versicolor es menor en QDA que en LDA
- La precision de virginica es menor en QDA que en LDA
- El recall de virginica es mayor en QDA que en LDA
- El accuracy es mejor en LDA que en QDA

Si bien las diferencias observadas son pocas, podrían darnos una idea de cuando utilizar LDA y cuando QDA dependiendo de que se requiera priorizar: pocos falsos positivos o pocos falsos negativos y de qué clase.

4.3. Clasificación utilizando LDA, QDA y 4 features

En esta subsección se buscarán los clasificadores de LDA y QDA que se obtienen al considerar los 4 features, es decir: largo y ancho del sépalo y largo y ancho del pétalo. Dado que la visualización en este caso es complicada, se compararán de acuerdo a sus resultados en las métricas de precision, recall y accuracy que se obtienen al predecir sobre el set de entrenamiento y utilizando leave one out cross validation.

clases_predichas_lda_completo

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

clases_predichas_lda_completo_loo

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

clases_predichas_qda_completo

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

clases_predichas_qda_completo_loo

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	3
virginica	0	1	49

Figura 14: Predicciones sobre el set de entrenamiento utilizando 4 features, LDA, QDA y leave one out cross-validation

LDA	Setosa	Versicolor	Virginica
Precision	1	0.97	0.96
Recall	1	0.96	0.98
Accuracy	0.98		

LDA LOO	Setosa	Versicolor	Virginica
Precision	1	0.97	0.96
Recall	1	0.96	0.98
Accuracy	0.98		

QDA	Setosa	Versicolor	Virginica
Precision	1	0.97	0.96
Recall	1	0.96	0.98
Accuracy	0.98		

QDA LOO	Setosa	Versicolor	Virginica
Precision	1	0.98	0.94
Recall	1	0.94	0.98
Accuracy	0.97		

A partir de los resultados anteriores se pueden realizar las siguientes observaciones:

- Tanto la estimación sesgada como la estimación insesgada (la que se obtiene utilizando leave one out CV) del error de clasificación coinciden en el caso de LDA. Esto es un indicaría que el clasificador funciona aproximadamente igual de bien al clasificar instancias con las que entreno que con instancias nuevas.

- Todas las métricas obtenidas en estos clasificadores son significativamente mejores que las que se obtuvieron cuando se utilizaron tan solo dos features. Esto puede deberse a la influencia del feature largo de pétalo, que como se vio en los histogramas del análisis inicial demostraba tener una varianza entre grupos alta.
- QDA tiene una mayor precision de versicolor y un mayor recall de virginica que LDA, sin embargo tiene un peor desempeño en la precision de virginica y el recall de versicolor que LDA
- Nuevamente setosa es la clase cuya clasificación tiene mejor desempeño.

5. Conclusiones

A lo largo del trabajo práctico se han logrado construir clasificadores de LDA y de QDA que clasifican a flores a partir de sus medidas de largo y ancho de sépalo y de pétalo. A partir del análisis inicial y de los desempeños de los clasificadores según las métricas de precision, recall y accuracy se pueden obtener las siguientes conclusiones:

- Condicionando a cada clase, los features parecen seguir una distribución normal, esto se puede ver en los qqplots realizados en el análisis inicial pues los cuantiles muestrales normalizados se asemejaban a los de una normal estándar. Además al graficar scatter plots de a pares de features es posible observar que parecen venir de una distribución normal bivariada pues las observaciones están dispersas de forma elipsoidal. Esto implica que el supuesto de normalidad de los features condicionado a las clases, se cumple.
- Los features largo y ancho de pétalo son buenos separando las instancias de clases diferentes y agrupando las instancias de una misma clase. Esto se puede ver en los histogramas del análisis inicial.
- La clase setosa es la clase mas sencilla de clasificar. Esto se debe a que sus valores de ancho y largo de pétalo son mucho menores que para las clases virginica y versicolor. Por lo tanto un clasificador de QDA y LDA que utilice estos features aprovechará esta varianza que existe entre las distintas clases.
- Las clases virginica y versicolor no son siempre sencillas de distinguir. Para algunos valores de features estas clases se solapan y por lo tanto la separación (lineal o cuadrática) no podrá ser perfecta. Por ejemplo, si bien en general las flores de clase virginica tienen un mayor ancho de pétalo que las de la clase versicolor, existen instancias de clase versicolor con mayor ancho de pétalo que instancias de clase virginica.
- Al estimar el error de predicción, es útil utilizar cross-validation. Al entrenar el modelo con todas las observaciones menos con una y luego predecir sobre la observación no utilizada, la estimación nos da una idea de como se comportará el modelo sobre datos no vistos. Además previene problemas de overfitting: un alto accuracy sobre el set de entrenamiento podría ser simplemente resultado de sobreajustar los datos de entrenamiento, lo que podría provocar un mal desempeño sobre nuevas observaciones.
- Las métricas de precision y recall son muy útiles para comparar los modelos encontrados y para decidir cual de ellos utilizar dependiendo de las necesidades y prioridades que se requieran. Si se buscan tener pocos falsos positivos sobre una clase, entonces interesará que el modelo tenga una buena precision sobre la clase. Por otro lado si se requiere que haya pocos falsos negativos sobre una clase, entonces se buscará que el modelo tenga un buen recall sobre la clase.
- Los modelos de QDA y LDA encontrados utilizando los 4 features tuvieron un mejor desempeño que los encontrados utilizando 2 features. Es posible que se deba a la utilización del feature largo de pétalo que como se analizó inicialmente lograba una buena separación de las clases.

6. Bibliografía

1 Flury B., A First Course in Multivariate Statistics

2 Seber G., Multivariate Observations

3 Rencher, A. (n.d.). Methods of Multivariate Analysis (2nd ed.). Brigham Young University: John Wiley Sons, Inc.

4 Boente Graciela, Yohai Victor, Apunte teórico