Análisis de Componentes y Discriminantes
PCA, LDA, CCA y otras A's

FaMAF

2019

Análisis de componentes y discriminantes

- Un acercamiento al problema de exceso de dimensionalidad es reducir dimensión mediante combinaciones de características.
- Las combinaciones lineas son muy atractivas pues son simples de calcular y tratables de forma analítica.
- Dos formas clásicas de definir transformaciones lineales son
  - Principal Component Analysis or PCA busca la proyeccion que mejor representa los datos en el sentido de mínimos cuadrados.
  - Multiple Discriminant Analysis or MDA busca la proyección que mejor separa los datos en el sentido de mínimos cuadrados.

Análisis de componentes principales

- Consideremos el problema de representar la muestra $\{x_1, \ldots, x_n\}$ por un único vector x_0.
- Esto es, encontrar x_0 tal que la suma de las distancias entre x_0 y x_k es tan chica como sea posible. Si J_0 es el criterio del cuadrado del error

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{x}_k\|^2 \tag{1}$$

buscamos x_0 que minimize J_0. La solución es la media muestral

$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \tag{2}$$

Análisis de componentes principales: x_0

Esto se verifica de la siguiente forma

$$
\begin{aligned}
J_0(\mathbf{x}_0) &= \sum_{k=1}^{n} \|(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_k - \mathbf{m})\|^2 \\
&= \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{m}\|^2 - 2\sum_{k=1}^{n} (\mathbf{x}_0 - \mathbf{m})^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{m}\|^2 - 2(\mathbf{x}_0 - \mathbf{m})^t \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{m}\|^2 + \underbrace{\sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2}_{\text{independent of } \mathbf{x}_0}
\end{aligned}
\tag{3}
$$

Como la segunda suma es independiente de x_0, $J_0(x_0)$ se minimiza en $x_0 = m$.

Análisis de componentes principales: $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

- La media muestral es un representación de dimensión cero del dataset. Es simple pero no revela nada sobre la variabilidad de los datos.
- Proyectando los datos sobre una linea que pase por la media se puede obtener una representación de dimensión uno mas interesante
- Sea $\mathbf{e}$ el vector unitario en la dirección de la linea

$$\mathbf{x} = \mathbf{m} + a\mathbf{e} \tag{4}$$

donde el escalar a corresponde a la distancia del punto $\mathbf{x}$ a la media m.

Análisis de componentes principales: $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

Si representamos $\mathbf{x_k}$ con $\mathbf{m} + a_k\mathbf{e}$ podemos encontrar un conjunto de coeficientes óptimos a_k minimizando el el criterio del cuadrado del error en este caso

$$J_1(a_1, \ldots, a_n, \mathbf{e}) = \sum_{k=1}^{n} \|(\mathbf{m} + a_k\mathbf{e}) - \mathbf{x}_k\|^2 = \sum_{k=1}^{n} \|a_k\mathbf{e} - (\mathbf{x}_k - \mathbf{m})\|^2$$
$$= \sum_{k=1}^{n} a_k^2\|\mathbf{e}\|^2 - 2\sum_{k=1}^{n} a_k\mathbf{e}^t(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \quad (5)$$

Derivando e igualando a cero se obtienen las soluciones

$$a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m}) \quad (6)$$

Análisis de componentes principales: $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

Si minimizamos J_1 con respecto a $\mathbf{e}$, encontramos la mejor dirección y conjunto de coeficientes que representan los datos.

$$\begin{aligned} J_1(\mathbf{e}) &= \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\ &= -\sum_{k=1}^{n} \left[\mathbf{e}^I (\mathbf{x}_k - \mathbf{m})\right]^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\ &= -\sum_{k=1}^{n} \mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) (\mathbf{x}_k - \mathbf{m})^t \mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\ &= -\mathbf{e}^t \mathbf{S} \mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \end{aligned} \tag{7}$$

donde

$$\mathbf{S} = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m}) (\mathbf{x}_k - \mathbf{m})^t \tag{8}$$

Análisis de componentes principales: $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

- El vector $\mathbf{e}$ que minimiza J_1 maximiza $\mathbf{e}^t\mathbf{S}\mathbf{e}$.
- Usemos el método de los multiplicadores de Lagrange para maximizar $\mathbf{e}^t\mathbf{S}\mathbf{e}$ sujeto a la restricción $\|\mathbf{e}\| = 1$,
- Si λ es el multiplicador, derivando

$$u = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda\left(\mathbf{e}^t\mathbf{e} - 1\right) \tag{9}$$

con respecto a $\mathbf{e}$ queda

$$\frac{\partial u}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} \tag{10}$$

Análisis de componentes y discriminantes

- Igualando a cero queda que $\mathbf{e}$ debe ser un autovector de la matriz de dispersión

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e} \tag{11}$$

- En particular, como $\mathbf{e}^t\mathbf{S}\mathbf{e} = \lambda\mathbf{e}^t\mathbf{e} = \lambda$, se sigue que para maximizar $\mathbf{e}^t\mathbf{S}\mathbf{e}$ hay que seleccionar el autovector correspondiente al mayor autovalor de la matriz de dispersión,
- Esto significa que para encontrar la mejor proyección unidimensional de los datos (mejor en el sentido del menor error cuadrático medio) proyectamos los datos en la línea que pasa por la media muestral en dirección del autovector de la matriz de dispersión correspondiente al mayor autovalor.

Análisis de componentes principales:
$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$

Este resultado puede extenderse a proyecciones d- dimensionales .
Si queremos representar ahora

$$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i \tag{12}$$

donde $d' < d$. No es difícil ver que la función criterio

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2 \tag{13}$$

se minimiza cuando los vectores $\mathbf{e}_1, \cdots, \mathbf{e}_{d'}$ son los autovectores correspondientes a los autovalores de la matriz de dispersión.

Análisis de componentes principales

- Como la matriz de dispersión es real y simétrica, los autovectores son ortogonales y forman una base del espacio de las características $\mathbf{x}$.
- Los coeficientes a_{ki} son las i-esima componentes de $\mathbf{x_k}$ en esta base y son llamadas componentes principales.
- Un método iterativo de cálculo de las componentes es pensar en elegir de a una las direcciones que maximicen la varianza de la proyección con la restricción de que las direcciones sean decorrelacionadas entre si. Por lo cual la varianza de las componentes son los autovalores de la matriz de dispersión.
- La transformación del espacio de datos mediante decorrelación es llamada transformación de Kahunen Loeve.

Análisis de componentes principales

- Las suma de las varianzas de los datos, esto es, la traza de S, es igual a la suma de los autovalores de S. Esto permite hablar del porcentaje de varianza total que recoge un componente principal:

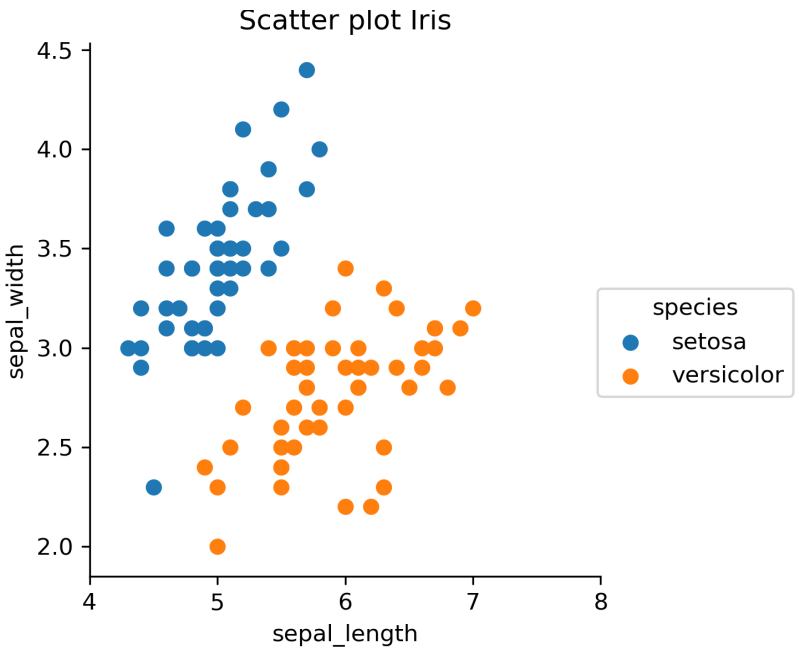
$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{\lambda_i}{\sum_{i=1}^{p} \mathrm{Var}\,(x_i)} \tag{14}$$

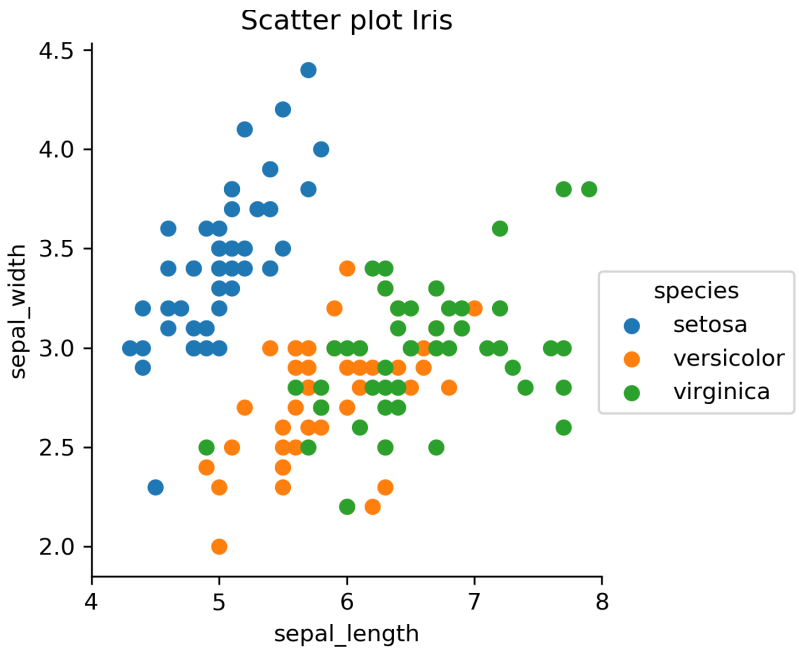
- Así, también se podrá expresar el porcentaje de variabilidad recogido por los primeros m componentes:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \mathrm{Var}\,(x_i)} \tag{15}$$

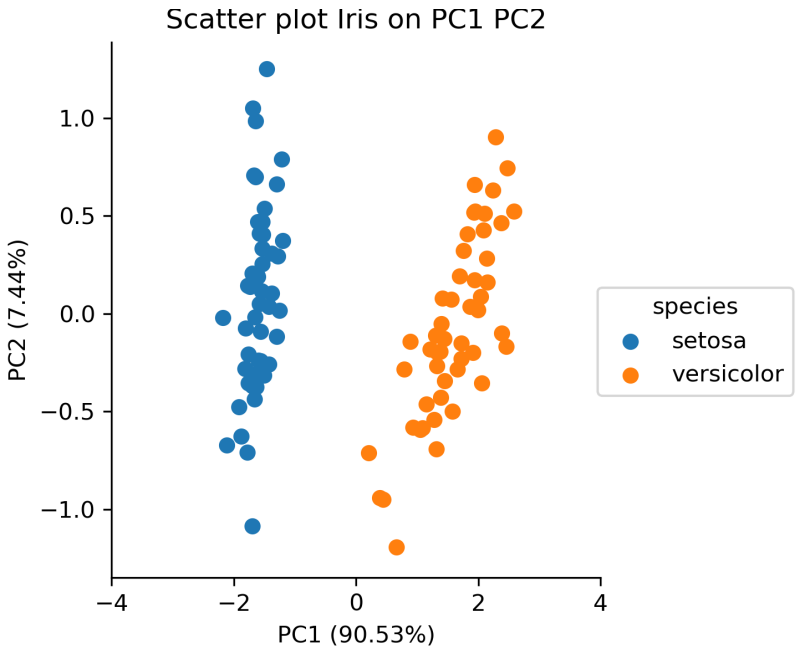
donde $m < d$.

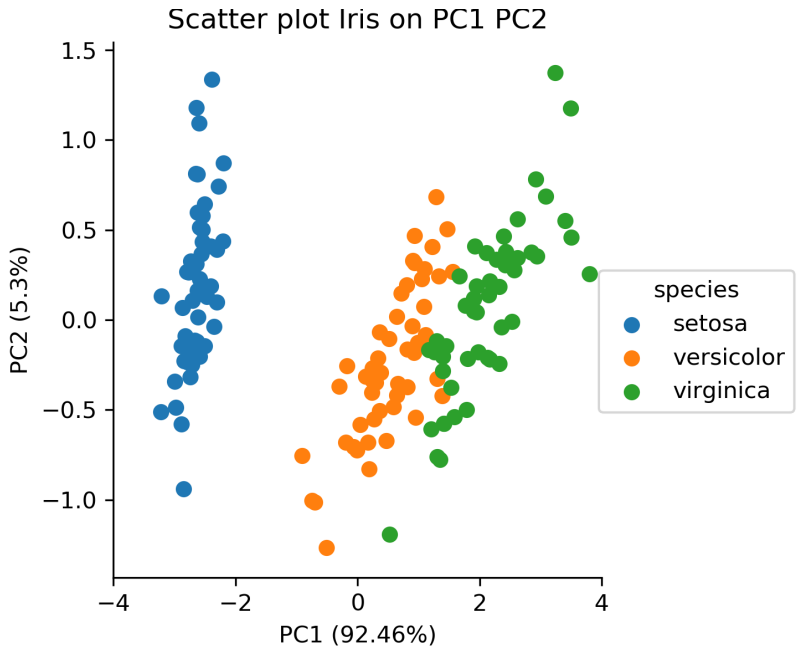
Iris sobre dos variables: ancho y largo de los sépalos





Iris graficado sobre todas las variables combinadas en las direcciones de mayor variabilidad

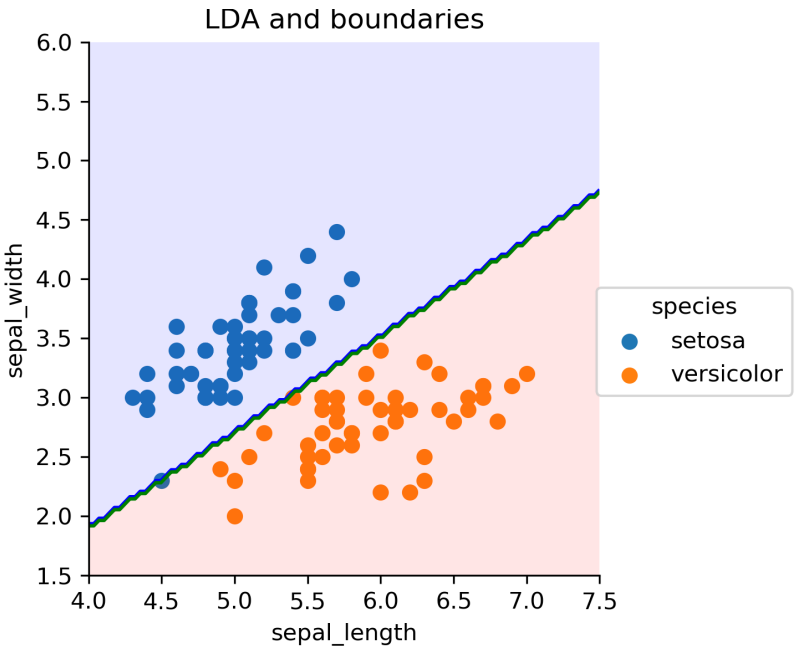


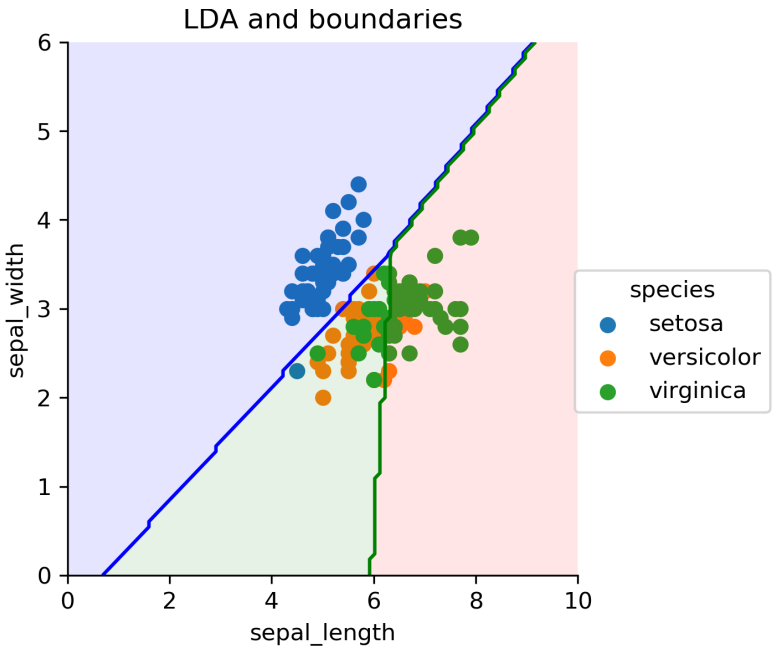


Discriminantes lineales y cuadráticos de Iris.

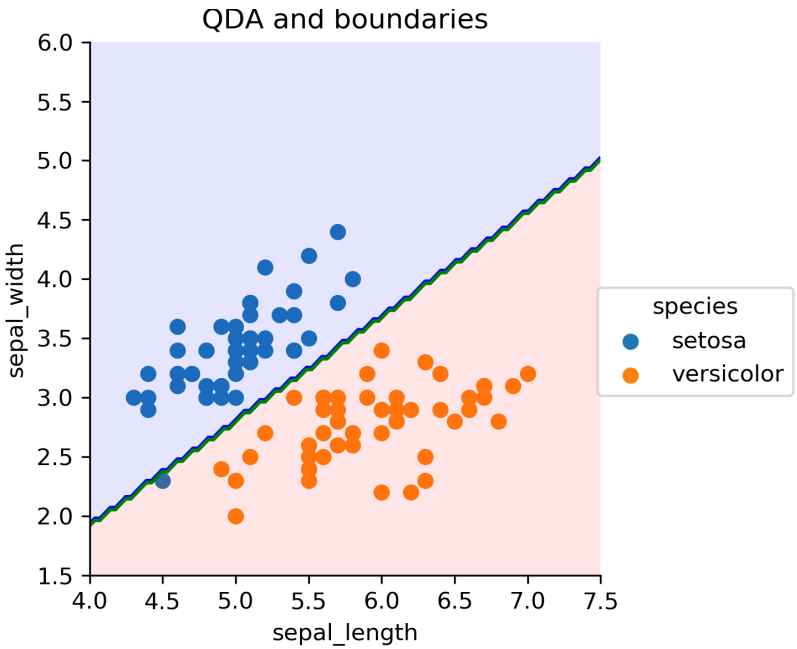
- PCA encuentra componentes que son útiles para representar datos, pero no hay razones para pensar que esas componentes son útiles para discriminar poblaciones.
- Mejora mucho Iris sobre las componentes porque algunas variables agregan información. Aún así, en vez de buscar las direcciones que separen mas los datos.
- Si suponemos Iris con distribuciones Gausianas y calculamos los discriminantes lineales y cuadráticos sobre los datos (en particular para estas variables), son mucho peores que sobre las componentes.

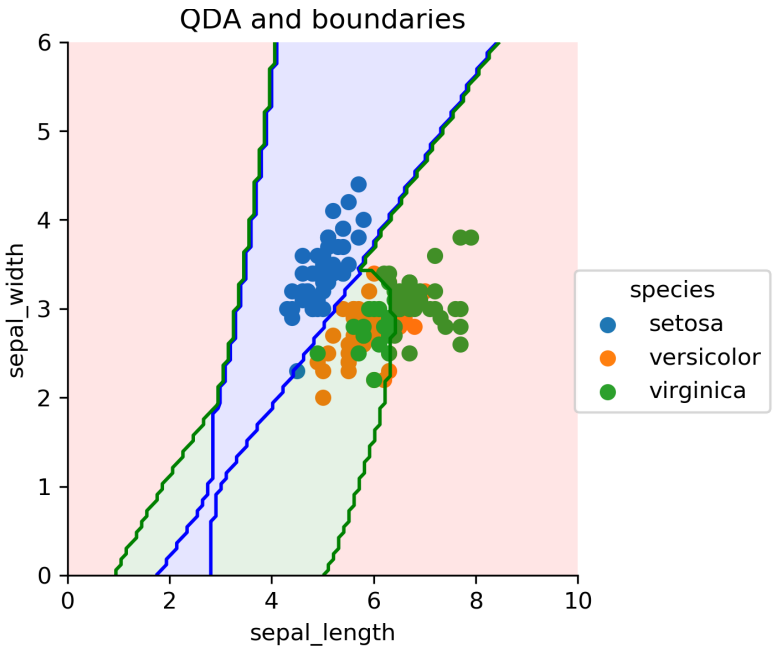
Discriminantes lineales de Iris sobre dos variables: ancho y largo de los sépalos



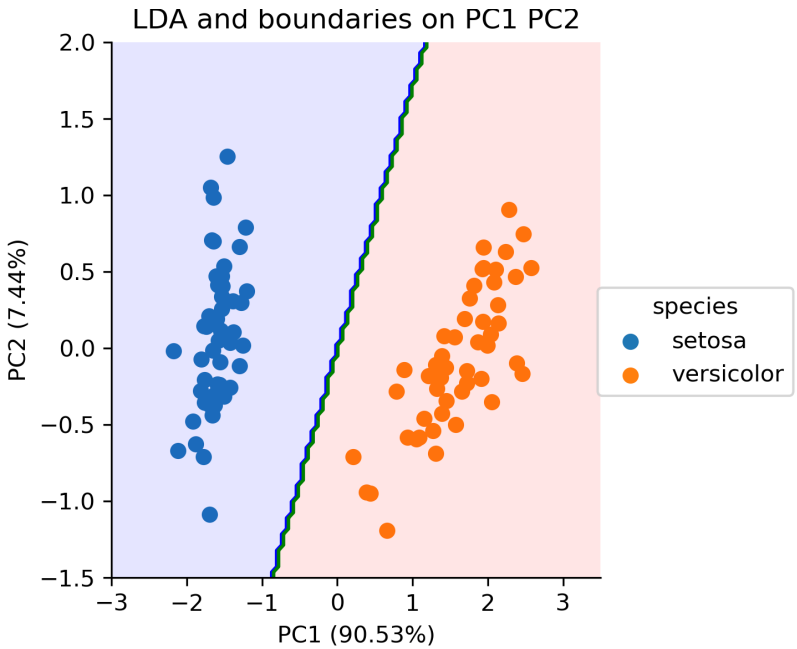


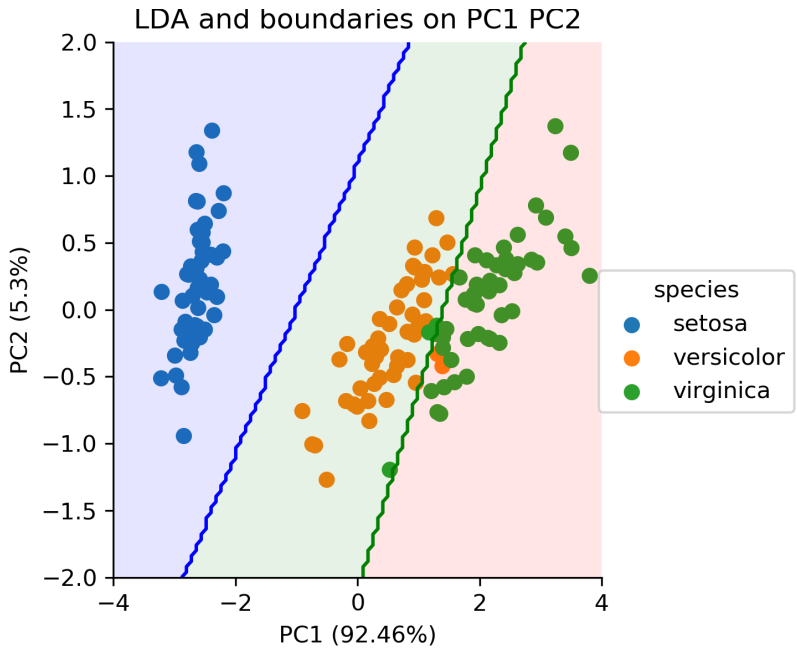
Discriminantes cuadráticos de Iris sobre dos variables: ancho y largo de los sépalos



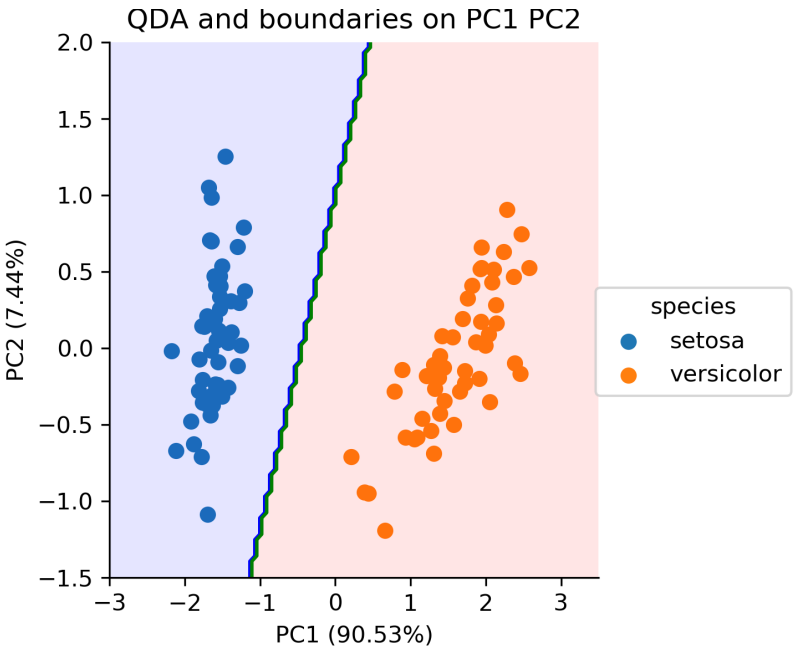


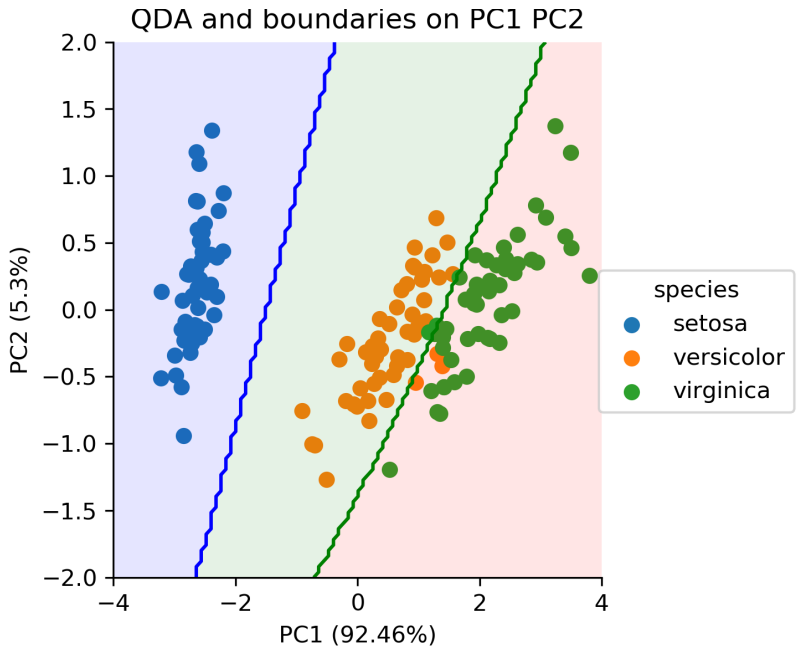
Discriminantes lineales de Iris sobre las primeras dos componentes principales





Discriminantes cuadráticos de Iris sobre las primeras dos componentes principales





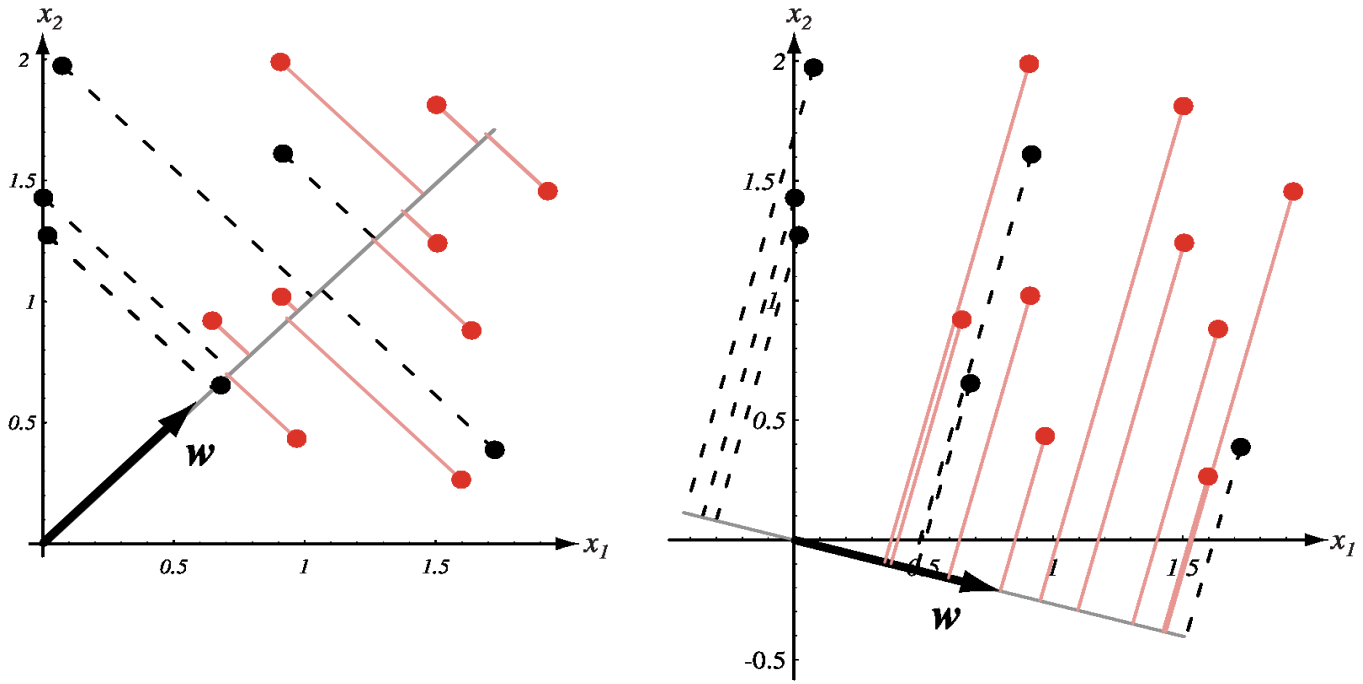
LDA: linear discriminant analysis

- Consideremos el problema de proyectar datos de d dimensiones sobre una linea.
- Si la linea es elegida al azar, aun grupos bien separados pueden tener sus proyecciones mezcladas, generando una problema de discriminación difícil.
- Si se busca cuidadosamente, en muchos problemas se puede encontrar una linea cuyas proyecciones tengan una separación mayor que en el espacio original.
- Ese es el objetivo de LDA.

LDA: linear discriminant analysis

- Supongamos que tenemos un conjunto de n muestras d-dimensionales $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de las cuales n_1 pertenecen al conjunto D_1 etiquetado como ω_1 y n_2 pertenecen al conjunto D_2 etiquetado como ω_2.
- La proyección $y = \mathbf{w}^t \mathbf{x}$ con $\|\mathbf{w}\| = 1$ produce $y_1, \ldots, y_n$ divididos en conjuntos $\mathcal{Y}_1$ y $\mathcal{Y}_2$.
- D_1 y D_2 son las muestras de entrenamiento.

LDA: linear discriminant analysis



Podemos ver que aun la mejor de las proyecciones no separa totalmente los datos.

LDA: linear discriminant analysis

- Si buscamos la mejor dirección discriminatoria $\mathbf{w}$, una medida de la separación entre los puntos proyectados es la diferencia entre las medias muestrales.

$$\begin{aligned} \mathbf{m}_i &= \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x} \\ \tilde{m}_i &= \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i \end{aligned} \tag{16}$$

- La distancia entre las medias proyectadas es

$$|\tilde{m}_1 - \tilde{m}_2| = \left| \mathbf{w}^t \left(\mathbf{m}_1 - \mathbf{m}_2 \right) \right| \tag{17}$$

  y se puede hacer esta diferencia lo mas grande posible escalando $\mathbf{w}$.
- La escala mas apropiada es algún factor que considere las varianzas de las clases.

LDA: linear discriminant analysis

- Definimos la dispersión de las muestras proyectadas de la clase ω_i como

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2 \tag{18}$$

- Entonces $(1/n)\left(\tilde{s}_1^2 + \tilde{s}_2^2\right)$ es un estimador de la matriz de dispersion conjunta y $\left(\tilde{s}_1^2 + \tilde{s}_2^2\right)$ se llama la dispersion dentro de la muestra proyectada.

LDA: linear discriminant analysis

- El discriminante lineal de Fisher usa la función lineal $\mathbf{w}^t\mathbf{x}$ para la cual la funcion criterio

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{19}$$

  es máxima e independiente de $\|\mathbf{w}\|$.
- Mientras que $\mathbf{w}$ que maximiza $J(\mathbf{w})$ lleva a la mejor separación entre los dos conjuntos proyectados, se necesita también un criterio de umbral para tener un clasificador.
- Primero encontraremos el $\mathbf{w}$ óptimo y luego los umbrales.

LDA: linear discriminant analysis

- Escribamos $J(\mathbf{w})$ como una función de $\mathbf{w}$. Sean S_i y $S_{\mathbf{w}}$ las matrices de dispersión

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i) (\mathbf{x} - \mathbf{m}_i)^t \quad \mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

- Entonces podemos escribir

$$\tilde{s}_i^2 = \sum_{\mathbf{x} \in \mathcal{D}_i} \left(\mathbf{w}^t \mathbf{x} - \mathbf{w}^t \mathbf{m}_i\right)^2 = \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{w}^t (\mathbf{x} - \mathbf{m}_i) (\mathbf{x} - \mathbf{m}_i)^t \mathbf{w} = \mathbf{w}' \mathbf{S}_i \mathbf{w}$$

  Por lo cual la suma de esas dispersiones resulta
  $\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t \mathbf{S}_W \mathbf{w}$

- En forma similar, la separación entre las medias proyectadas pueden escribirse como

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \left(\mathbf{w}' \mathbf{m}_1 - \mathbf{w}^t \mathbf{m}_2\right)^2 = \mathbf{w}^t (\mathbf{m}_1 - \mathbf{m}_2) (\mathbf{m}_1 - \mathbf{m}_2)^t \mathbf{w} = \mathbf{w}' \mathbf{S}_B \mathbf{w}$$

donde $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2) (\mathbf{m}_1 - \mathbf{m}_2)^t$

LDA: linear discriminant analysis

- LLamamos a $\mathbf{S}_W$ la matriz de dispersión dentro de la clase, la cual es proporcional a matriz de covarianza muestral del grupo completo de datos d dimensionales.
- La matriz es simétrica, definida positiva y usualmente, si $n > d$, no singular.
- LLamamos a $\mathbf{S}_B$ la matriz de dispersión entre clases. Es simétrica, definida positiva y como es el producto de dos vectores, de rango 1.

LDA: linear discriminant analysis

- En términos de estas matrices,

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}} \tag{20}$$

- Esta expresión es conocida como el cociente de Rayleigh y es fácil ver que el vector $\mathbf{w}$ que maximiza $J(\mathbf{w})$ debe satisfacer

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

para alguna constante λ, el cual es un problema de autovalores generalizado.

LDA: linear discriminant analysis

- Si $\mathbf{S}_W$ es no singular se tiene un problema de autovalores regular

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}$$

- Como $\mathbf{S}_B\mathbf{w}$ tiene dirección de $\mathbf{m}_1 - \mathbf{m}_2$, la solución del $\mathbf{w}$ que maximiza $J(\mathbf{w})$ es

$$\mathbf{w} = \mathbf{S}_W^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right) \tag{21}$$

el cual es llamado discriminante de Fisher.

LDA: Clasificación

- Para diseñar el clasificador, es necesario encontrar el umbral de decisión a favor de ω_1.
- Si ambas poblaciones tienen igual varianza, y digamos que la proyección de la media de ω_1 está a la derecha de la ω_2, la regla de Fisher asigna una nueva observación x_0 a ω_1 si la proyección de x_0 sobre el discriminante es mayor al punto medio entre las proyecciones de las medias muestrales

$$m = \left[\mathbf{S}_W^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right)\right]^t \frac{1}{2}\left(\mathbf{m}_1 + \mathbf{m}_2\right)$$

- Esto es, asigno x_0 a ω_1

$$\left[\mathbf{S}_W^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right)\right]^t x_0 - m \geq 0 \tag{22}$$

LDA: Clasificación

- Cuando las densidades condicionales $p(\mathbf{x}|\omega_i)$ son normales multivariadas con igual matriz de covarianza proporcional a $\mathbf{S}_W$ se puede calcular el umbral directamente, pues el borde de decisión óptimo es

$$\mathbf{w}^t x + w_0 = 0 \qquad \mathbf{w} = \mathbf{S}_W^{-1}(\boldsymbol{m_1} - \boldsymbol{m_2}) \tag{23}$$

  y w_0 es una constante que involucra $\mathbf{w}$ y las densidades a priori.

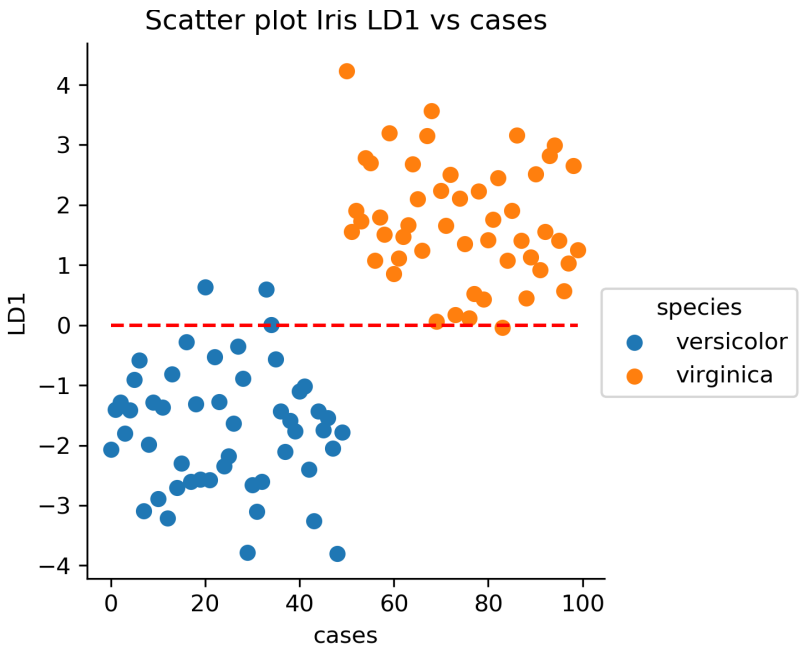
$$w_0 = \frac{1}{2}(\boldsymbol{m_1} - \boldsymbol{m_2}) + \frac{\ln\left[P(\omega_1)/P(\omega_2)\right]}{(\boldsymbol{m_1} - \boldsymbol{m_2}))^t \mathbf{S}_W^{-1}(\boldsymbol{m_1} - \boldsymbol{m_2})} \tag{24}$$

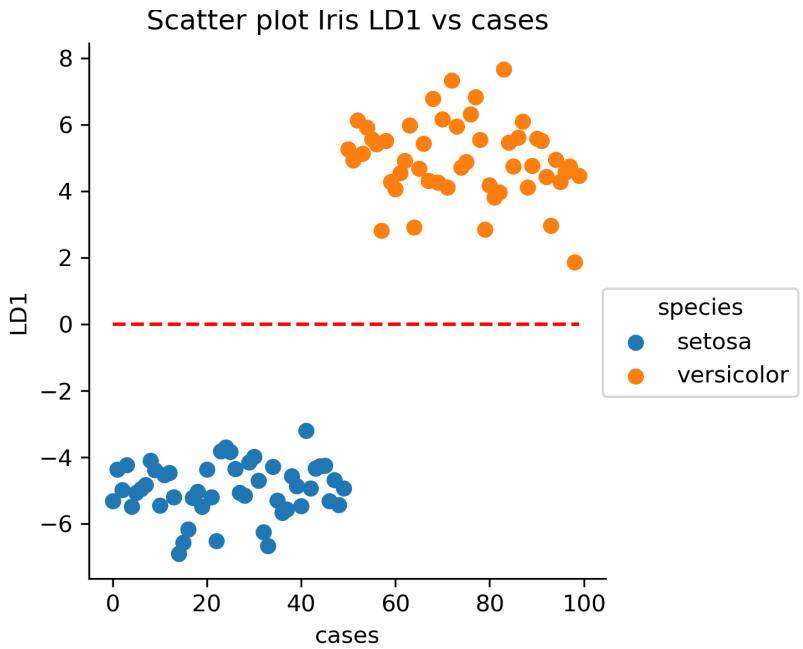
- Observemos que si las densidades a priori son iguales, la regla resulta igual a la regla de Fisher.

LDA: Clasificación

- La regla plug-in reemplaza las medias y matriz de varianza covarianza por las versiones muestrales obtenidas por máxima verosimilitud.
- En forma general, si se suavizan los datos proyectados o se ajusta una normal unidimensional, se elige w_0 de tal forma que las densidades a posteriori correspondiente sean iguales como pide la regla de Bayes.

Proyeccion de Iris sobre el primer discriminante





MDA: Multiple Discriminant Analysis

- Para el problema de c clases, la generalización natural del discriminante de Fisher consta de $c-1$ funciones discriminantes. Por lo cual la proyecccion es de un espacio d-dimensional a un espacio $c-1$ dimensional.
- La matriz de dispersión dentro de las clases es

$$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i \tag{25}$$

con

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \tag{26}$$

y

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x} \tag{27}$$

MDA: Multiple Discriminant Analysis

- La generalización de la matriz de dispersión entre clases no es tan obvia.
- Si la media total $\mathbf{m}$ y la matriz de dispersión total $\mathbf{S}_T$ por

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i \qquad \mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t \quad (28)$$

MDA: Multiple Discriminant Analysis

- Sigue que

$$\begin{aligned}\mathbf{S}_T &= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^t \\ &= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t + \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t \\ &= \mathbf{S}_W + \sum_{i=1}^{c} n_i (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t \end{aligned} \tag{29}$$

por lo cual se define como la matriz de dispersion entre clases como

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t \tag{30}$$

y queda $\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$

MDA: Multiple Discriminant Analysis

- La proyeccion de un espacio d dimensional a uno $c-a$ dimensional se realiza con $c-1$ funciones discriminantes

$$y_i = \mathbf{w}_i^t \mathbf{x} \quad i = 1, \ldots, c-1 \tag{31}$$

- Si las y_i son vistas como componentes de un vector y y los vectores de pesos w_i son vistos como las columnas de $\mathbf{W}$, una matriz $d \times (c-1)$, la proyección resulta

$$\mathbf{y} = \mathbf{W}^t \mathbf{x} \tag{32}$$

- Las muestras $\mathbf{x}_1, \ldots, \mathbf{x}_n$ son proyectadas sobre $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

MDA: Multiple Discriminant Analysis

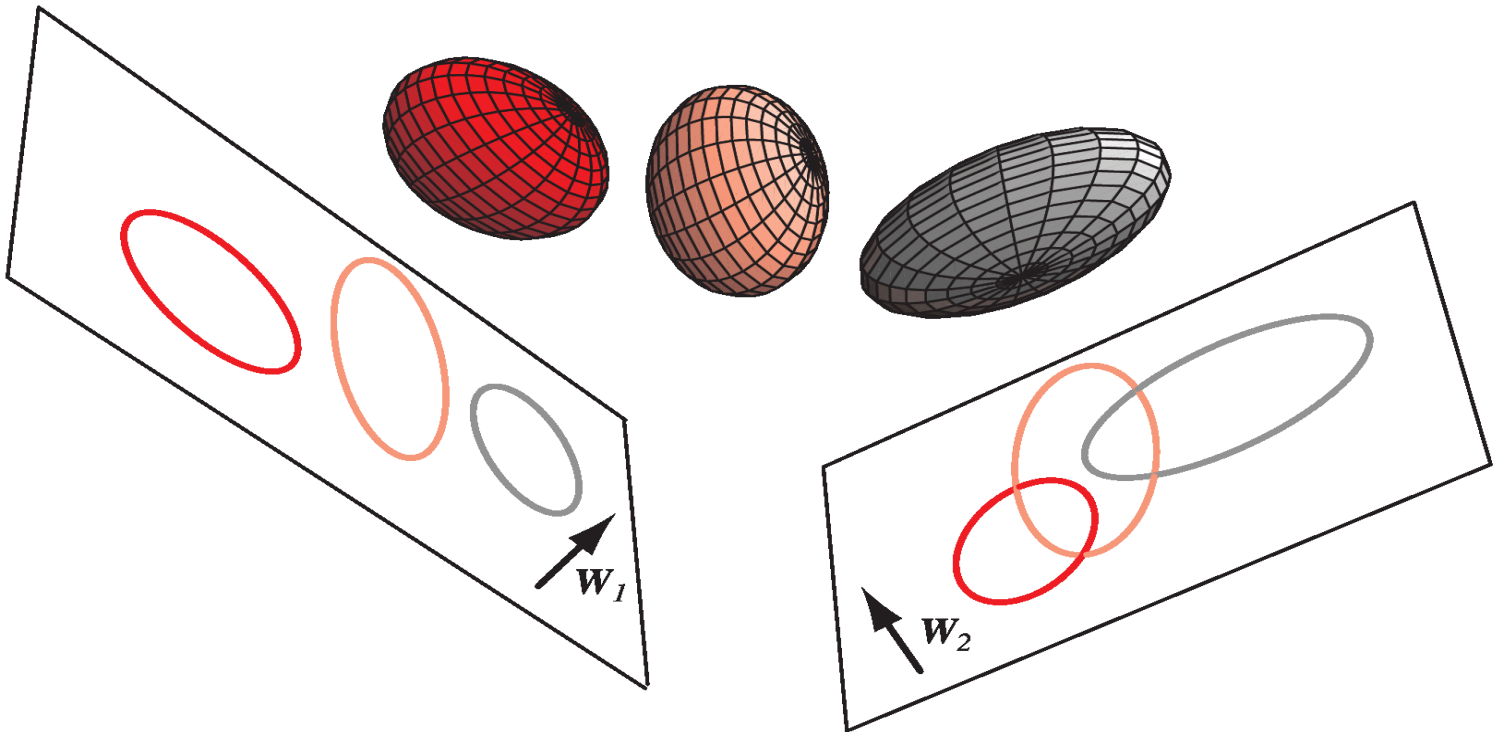
- Sus medias y matrices de dispersion se definen como

$$\begin{aligned}
\widetilde{\mathbf{m}}_i &= \frac{1}{n_i} \sum_{\mathbf{y} \in \mathcal{Y}_i} \mathbf{y} \\
\widetilde{\mathbf{m}} &= \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{\mathbf{m}}_i \\
\tilde{\mathbf{S}}_W &= \sum_{i=1}^{c} \sum_{\mathbf{y} \in \mathcal{Y}_i} (\mathbf{y} - \widetilde{\mathbf{m}}_i) (\mathbf{y} - \tilde{\mathbf{m}}_i)^t \\
\tilde{\mathbf{S}}_B &= \sum_{i=1}^{c} n_i (\widetilde{\mathbf{m}}_i - \widetilde{\mathbf{m}}) (\widetilde{\mathbf{m}}_i - \widetilde{\mathbf{m}})^t
\end{aligned} \tag{33}$$

- Entonces

$$\tilde{\mathbf{S}}_W = \mathbf{W}^t \mathbf{S}_W \mathbf{W} \quad \tilde{\mathbf{S}}_B = \mathbf{W}^t \mathbf{S}_B \mathbf{W} \tag{34}$$

MDA: Multiple Discriminant Analysis



Aqui

$\mathbf{W_1}$ genera el subespacio optimo que separa las clases

MDA: Multiple Discriminant Analysis

- Deseamos encontrar la matriz de transformación $\mathbf{W}$ que en algun sentido maximice el radio entre las matrices de dispersión
- Una medida escalar simple de la dispersión es el determinante de la matriz de dispersión, que es el producto de los autovalores, los cuales son una medida de la variabilidad en las direcciones principales.
- Usando esa medida se define la función de criterio

$$J(\mathbf{W}) = \frac{\left|\tilde{\mathbf{S}}_B\right|}{\left|\tilde{\mathbf{S}}_W\right|} = \frac{|\mathbf{W}^t \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_W \mathbf{W}|} \tag{35}$$

MDA: Multiple Discriminant Analysis

- Se puede ver que todo vector que maximiza $J(\mathbf{W})$ debe cumplir con la ecuación

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i \tag{36}$$

- Por lo cual las columnas de la matriz optima $\mathbf{W}$ son los autovectores generalizados que corresponden a los autovalores mas grandes de la matriz $\mathbf{S}_W^{-1}\mathbf{S}_B$.
- Como $\mathbf{S}_B$ es una suma de c matrices, de las cuales solo $c-1$ son independientes, $\mathbf{S}_B$ tiene rango a lo sumo $c-1$, y hay no mas de $c-1$ autovalores no nulos.

MDA: Multiple Discriminant Analysis

- Si la matriz $\mathbf{S}_W$ es isotrópica, los autovectores con autovalores no nulos generan el espacio de los vectores $\mathbf{m}_i - \mathbf{m}$ y puede encontrarse $\mathbf{W}$ mediante la ortogonalizacion de Gram Schmidt.
- Si $\mathbf{S}_W$ es no singular, no es necesario calcular su inversa pues no se la precisa, es mas eficiente computar los ceros del polinomio característico

$$|\mathbf{S}_B \mathbf{w}_i - \lambda_i \mathbf{S}_W \mathbf{w}_i| = 0 \tag{37}$$

y luego resolver

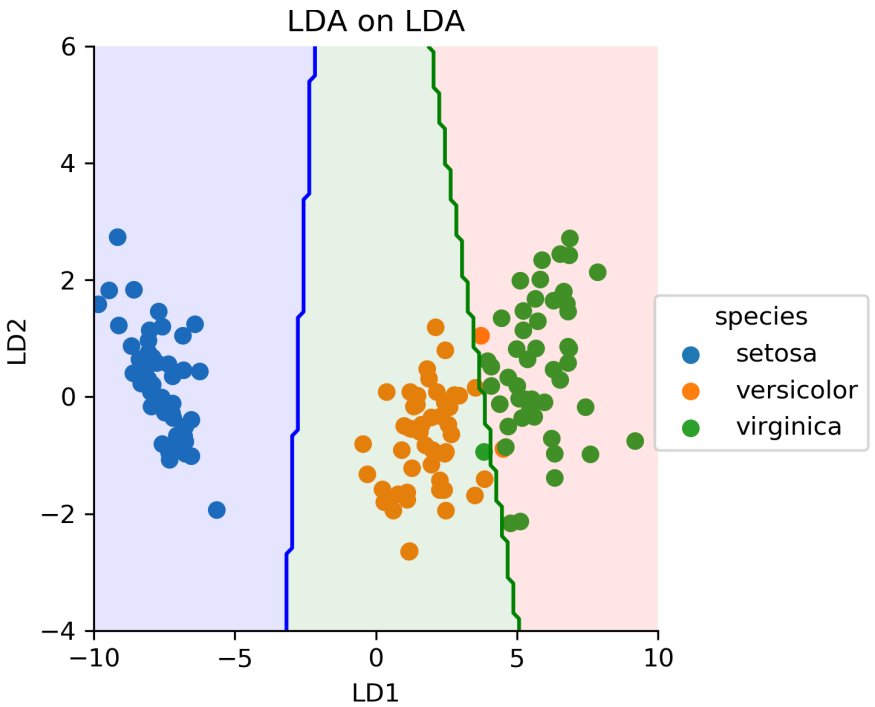
$$(\mathbf{S}_B - \lambda_i \mathbf{S}_W)\mathbf{w}_i| = 0 \tag{38}$$

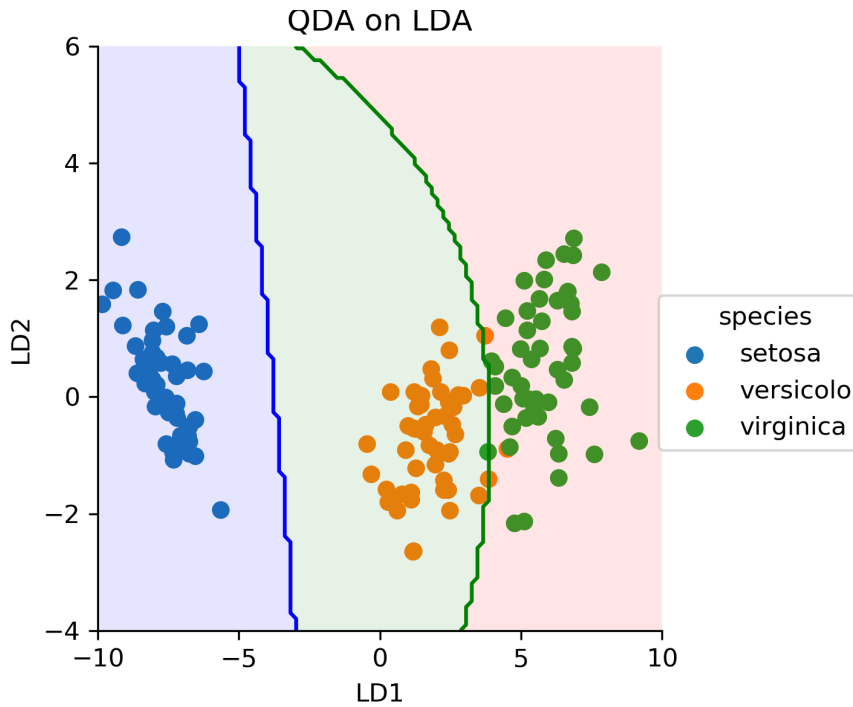
directamente para calcular $\mathbf{w}_i$.

MDA: Multiple Discriminant Analysis

- Una vez proyectados los datos se tiene un problema de clasificación en un espacio de dimensión menor.
- LDA y QDA pueden usarse para clasificar en el espacio original si las hipótesis de Normalidad e igualdad de varianzas son razonables, o puede usarse la proyección para reducir dimensión y aplicarse otro método sobre los datos transformados.

LDA and QDA en proyecciones discriminantes [LDA1 LD2]





MDA: Discriminación múltiple en el espacio original

- Por ejemplo, si se tienen datos d dimensionales y c clases si se consideran las matrices de varianza covarianza iguales Σ, y las clases son balanceadas, se definen los discriminantes de Fisher en el espacio original como

$$g_i(\mathbf{x}) = -\frac{1}{2} d(\mathbf{x}, \mathbf{m_i})_{\Sigma^{-1}} = -\frac{1}{2} (\mathbf{x} - \mathbf{m_i})^T \Sigma^{-1} (\mathbf{x} - \mathbf{m_i})$$

  y el clasificador clasifica a $\mathbf{x}$ en c_i si

$$g_i(\mathbf{x}) = \max_k g_k(\mathbf{x}) = \min_k d(\mathbf{x}, \mathbf{m_i})_{\Sigma^{-1}}$$

- Esto es, clasifica a $\mathbf{x}$ en la clase mas cercana segun la distancia de Mahalanobis al centroide de la clase.

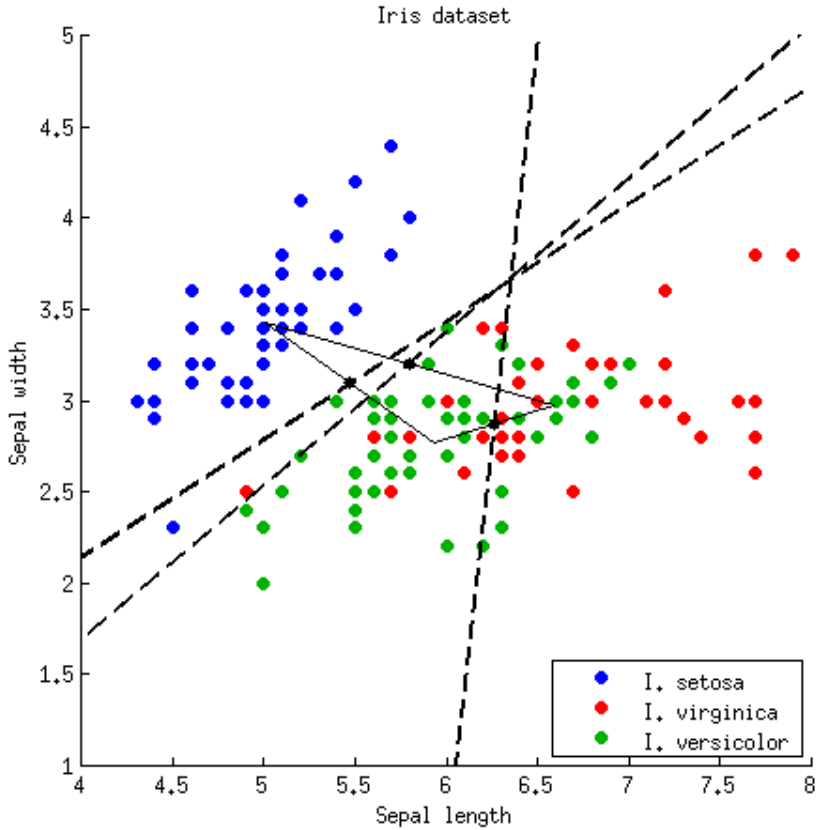
MDA: Multiple Discriminant Analysis

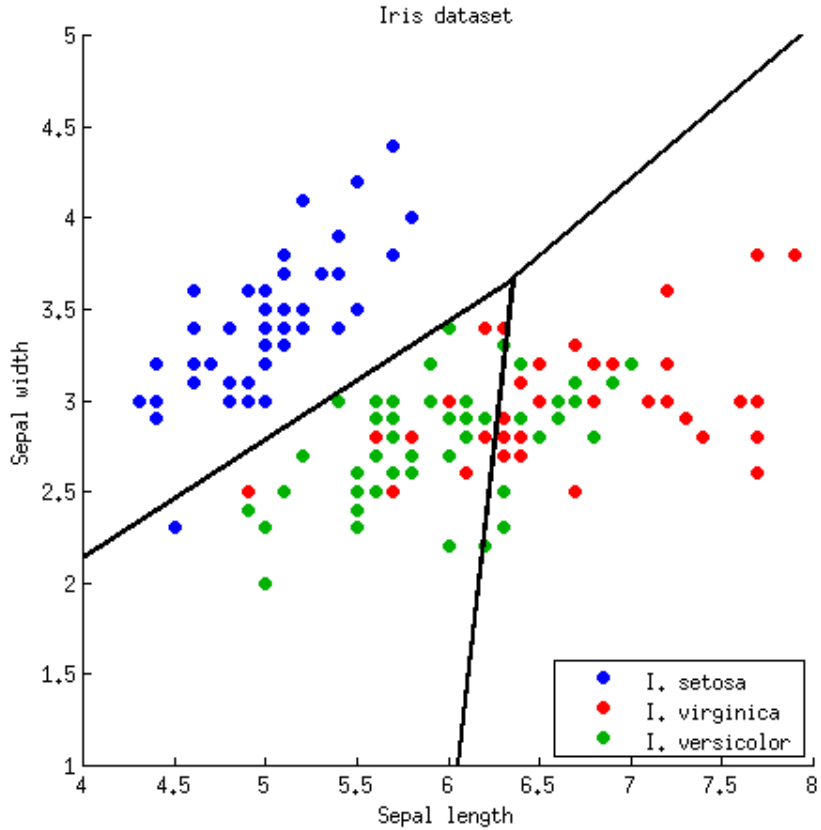
- Esto produce una Voronoi teselation, o partición del espacio por hiperplanos que se calculan igualando los discriminantes

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \qquad d(\mathbf{x}, \mathbf{m_i})_{\Sigma-1} = d(\mathbf{x}, \mathbf{m_j})_{\Sigma-1}$$

- Los bordes son ortogonales a $\Sigma^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ ya que la distancia de Mahalanobis induce una métrica d dimensional y pasan por el punto medio $(\mathbf{m}_1 + \mathbf{m}_2)/2$
- Iris tiene cuatro dimensiones, no se puede mostrar la clasificacion generada en ese espacio, por lo cual se "muestra" el resultado sobre un par de variables, o directamente se calcula sobre un par de variables.
- Este último recurso es sub-óptimo. Proyectando sobre PCA1 y PCA2 se obtiene un clasificacion mucho mejor que sobre cualquier par de variables de Iris. Lo mismo ocurre proyectando sobre LD1 yLD2.

MDA: Multiple Discriminant Analysis





MDA: Multiple Discriminant Analysis

- En el caso que las poblaciones sean normales con igual varianza, con distribución a priori igual, los discriminantes anteriores son equivalentes a las densidades a posteriori, generando la clasificación óptima.
- Si las densidades a priori son diferentes, el borde de las regiones ya no pasa por el punto medio de la linea que une las medias, sino que se acerca a la media de la clase con mas probabilidad a priori.

MDA: Multiple Discriminant Analysis

- Cuando las densidades condicionales $p(\mathbf{x}|\omega_i)$ son normales multivariadas con igual matriz de covarianza proporcional a $\mathbf{S}_W$ se puede calcular el umbral directamente, pues el borde de decisión óptimo es

$$\mathbf{w}^t x + w_0 = 0 \qquad \mathbf{w} = \mathbf{S}_W^{-1}(\boldsymbol{m_i} - \boldsymbol{m_j}) \tag{39}$$

  y w_0 es una constante que involucra $\mathbf{w}$ y las densidades a priori.

$$w_0 = \frac{1}{2}(\boldsymbol{m_i} - \boldsymbol{m_j}) + \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{m_i} - \boldsymbol{m_j}))^t \mathbf{S}_W^{-1}(\boldsymbol{m_i} - \boldsymbol{m_j})} \tag{40}$$

- Observemos que si las densidades a priori son iguales, la regla resulta igual a la regla de Fisher.

MDA: Multiple Discriminant Analysis

- Observemos que el cálculo de las variables canónicas, i.e. las direcciones discriminantes, es un método de reducción de dimensión.
- El tamaño de los autovectores permite decidir una reducción mayor que $c - 1$.
- La clasificación sobre el espacio reducido generado por las k componentes principales o las k variables canónicas se considera un problema completamente nuevo, donde nuevas distribuciones pueden ser ajustadas o métodos no paramétricos aplicados.

MDA: Multiple Discriminant Analysis

- Por ejemplo, en este nuevo espacio puede ser posible estimar matrices de covarianza separadas por clase, y usar hipótesis de normalidad que no podía usarse en el espacio original. No es lo mismo estimar una matriz 6x6 que una 2x2.
- Resulta raro volver a realizar un discriminante despues de proyectar sobre el discriminante, pero es una estrategia valida para reducir dimensión.

MDA: Multiple Discriminant Analysis

- Es importante notar que la literatura confunde ambos métodos por tener el mismo nombre, dado que el caso de 2 poblaciones que Fisher consideró se proyecta para clasificar. Y generaliza a 3 poblaciones calculando los bordes de las regiones de a dos.
- La generalización de Rao es para reducir dimensión e involucra la matriz de covarianza entre poblaciones. Es mas, usa la traza en vez del determinante, pero aun cuando es un criterio distinto produce las mismas variables canónicas.