Regresión Avanzada
Universidad Austral

PhD. Débora Chan

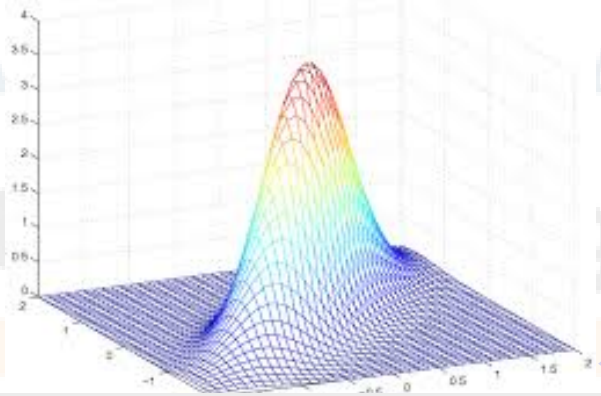
Junio-Julio de 2023

Comparación de medias en el caso multivariado

Distribución Normal Multivariada

La generalización a varias dimensiones de la densidad Normal univariada juega un papel fundamental en el análisis multivariado. Muchos de los fenómenos naturales del mundo real pueden ser estudiados por medio de la distribución Normal multivariada.

Incluso, a pesar de que el fenómeno estudiado no siga este modelo de distribución, las distribuciones de muchos de los estadísticos utilizados es aproximadamente Normal multivariada.



Distribución Normal Univariada

Expresión y Propiedades

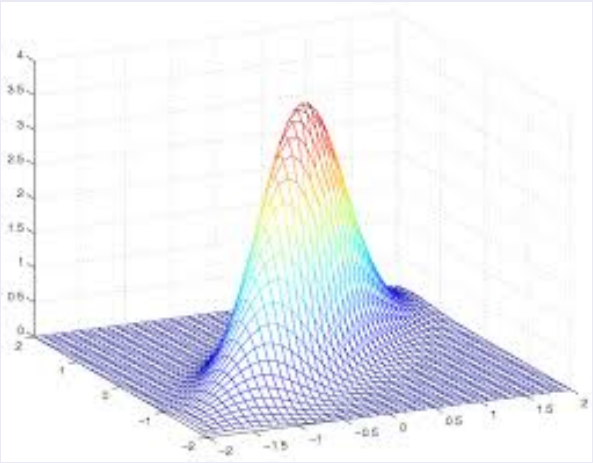
$X \sim N(\mu, \sigma^2)$, cuando su función de densidad de probabilidad es

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \forall x \in \mathbb{R}.$$

- ☆ Su gráfica es simétrica respecto de $x = \mu$.
- ☆ Su gráfica es asintótica respecto del eje de abscisas.
- ☆ Presenta un máximo en $(\mu; \frac{1}{\sqrt{2\pi}\sigma})$.
- ☆ Presenta dos puntos de inflexión, en $x = \mu - \sigma$ y en $x = \mu + \sigma$.
- ☆ La combinación lineal de v.a. Normales es otra v.a. Normal.
- ☆ El área bajo la curva dentro del intervalo $(\mu + k\sigma; \mu + t\sigma)$ no depende de μ ni de σ sino de los valores reales que tomen k y t.

Distribución Normal Univariada

Distintas Medias y Varianzas



Distribución Normal multivariada

Expresión

Un vector aleatorio continuo $X = (X_1, \cdots, X_n)^t$ tiene distribución Normal multivariada con vector de medias $\mu = (\mu_1, \cdots, \mu_n)^t$ y matriz de covarianzas Σ, simbólicamente $X \sim N_n(\mu, \Sigma)$, cuando su función de densidad de probabilidad está dada por

$$f(X, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right),$$

donde $-\infty < x_i < +\infty$ para todo $i = 1, \cdots, n$.

Cuando $n = 1$ se tienen a la distribución normal univariada como caso particular de la distribución normal multivariada.

Distribución Normal Multivariada

Propiedades

Cabe destacar las siguientes:

- ☆ Es una generalización del caso univariado.
- ☆ Tiene propiedades matemáticas que la hacen muy manejable.
- ☆ Depende de un número relativamente reducido de parámetros: n para el vector de medias y para la matriz de covarianzas $\frac{n(n+1)}{2}$.
- ☆ Ausencia de correlación equivale a independencia.
- ☆ Los datos disponibles rara vez siguen con exactitud una distribución Normal, ella distribución suele ser una aproximación útil.
- ☆ Al igual que en el caso univariado, esta distribución es el límite de la suma de vectores aleatorios independientes y con la misma distribución (Teorema Central del Límite).

Combinación Lineal de v.a. Normales Multivariadas

Coordenadas

Si $X \sim N_p(\mu, \Sigma)$, entonces cualquier combinación lineal de sus coordenadas también tiene distribución Normal; es decir, si c es un vector de constantes conocidas, entonces

$$Y = cX \sim N(c\mu, c\Sigma c^t).$$

Distribuciones Marginales

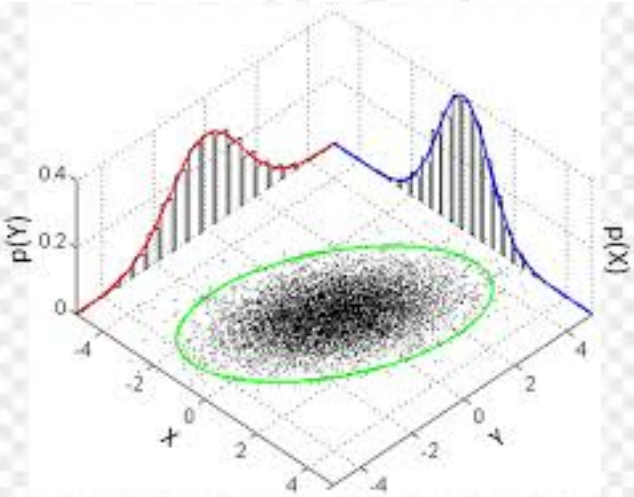
En particular para el caso de los vectores canónicos e_i, donde e_i tiene un 1 en la i-ésima coordenada y 0 enlas restantes.
Es decir que $e_i X$ corresponde a la coordenada X_i del vector aleatorio X.
Por lo tanto, cada una de las coordenadas de una distribución Normal multivariada tiene distribución Normal univariada. Equivalentemente, se puede afirmar que **las distribuciones marginales son normales**.

Distribución Normal Multivariada

Proyecciones

A partir del vector aleatorio $X = (X_1, \cdots, X_q, X_{q+1}, \cdots, X_n)^t$, la característica anterior permite definir las siguientes variables Normales

$$Y_1 = (X_1, \cdots, X_q)^t \qquad \text{e} \qquad Y_2 = (X_{q+1}, \cdots, X_n)^t.$$



La recíproca de esta propiedad no es cierta!!

Distribución Normal Multivariada

Ejemplo Transformación

Consideramos el vector aleatorio $X \sim N_3(\mu, \Sigma)$ donde $\mu = (1, 2, 3)^t$ y Σ es la matriz identidad de tamaño 3×3.
Sean las combinaciones lineales

$$\begin{cases} Y_1 &= 2X_1 + 3X_2, \\ Y_2 &= -1X_1 + 2X_3, \end{cases}$$

que pueden expresarse en notación matricial como

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}.$$

Distribución Normal Multivariada

Ejemplo Transformación

Entonces, la distribución conjunta de la variable $Y = (Y_1, Y_2)^t$ es Normal multivariada con parámetros μ_Y y Σ_Y dados por:

$$\mu_Y = \begin{pmatrix} 2 & 3 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 8 \\ 5 \end{pmatrix}$$

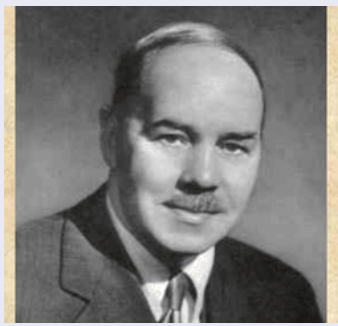
y

$$\Sigma_Y = \begin{pmatrix} 2 & 3 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 3 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 13 & -2 \\ -2 & 5 \end{pmatrix}.$$

Distribución Normal Multivariada

Estos estimadores verifican las siguientes propiedades:

- $\overline{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$.
- $nS \sim W$ donde $W_{p,n-1}(\Sigma)$ indica la distribución de Wishart con $n-1$ grados de libertad.
- La distribución de Wishart es una extensión al caso multivariado de la distribución χ^2 (ampliaremos este concepto en la próxima sección).
- $\widehat{\mu}$ y $\widehat{\Sigma}$ son independientes.



Jhon Wishart (1898-1956) ayudante de Pearson y de Fisher en Cambridge, consultor de la FAO.

Distribución Normal Multivariada

Comportamiento Asintótico de los Estimadores de Máxima Verosimimlitud

Ley débil de los grandes números para el caso multivariado: Si X es una variable multivariada con esperanza $E(X) = \mu$ y matriz de covarianzas $V(X) = \Sigma$, entonces vales las siguientes convergencias en probabilidad

$$\bar{X} \xrightarrow{P} \mu \qquad \text{y} \qquad S \xrightarrow{P} \Sigma.$$

Teorema central del límite para el caso multivariado: Si X es una variable aleatoria p-variada con esperanza $E(X) = \mu$ y matriz de covarianzas $V(X) = \Sigma$, entonces se tiene la siguiente convergencia en distribución

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N_p(0, \Sigma).$$

Distribución de Wishart

Definición

Se dice que una matriz W cuadrada de orden p; es decir, tiene p filas y p columnas, sigue una **distribución de Wishart**, $W \sim \mathcal{W}(\Sigma, p, n)$, si y sólo si W puede escribirse como

$$W = \sum_{i=1}^{n} X_i X_i^t,$$

donde X_i son vectores aleatorios independientes e idénticamente distribuidos con $X_i \sim N_p(0, \Sigma)$. Es decir que, la distribución de Wishart obedece a la suma de los productos entre distribuciones Normales multivariadas independientes de media 0 y varianza común Σ, y sus respectivas traspuestas.

Distribución Wishart

Generalización de χ^2

La distribucion de Wishart generaliza la distribución χ^2_n. De hecho, es fácil ver que si $p = 1$, entonces $\mathcal{W}(\sigma^2, 1, n) = \sigma^2 \chi^2_n$.

block title

La suma de variables aleatorias independientes con distribución Wishart con iguales parámetros Σ y p, es una nueva variable aleatoria con distribución Wishart conservando estos mismos parámetros.
Es decir, si $W_1 \sim \mathcal{W}(\Sigma, p, n_1)$ y $W_2 \sim \mathcal{W}(\Sigma, p, n_2)$ son independientes, entonces

$$W_1 + W_2 \sim \mathcal{W}(\Sigma, p, n_1 + n_2)$$

.

Distribuciones muestrales de la media y la varianza

Caso Univariado de muestra normal con media μ y varianza σ^2

$\overline{X}$ y varianza muestral S^2 satisfacen

$$\overline{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right) \qquad \text{y} \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

La última expresión equivale a decir

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Distribuciones Muestrales de Media y Varianza

¿Cómo se generaliza para el caso de un vector Normal multivariado?

Sean $X_1, X_2, \cdots, X_n$ vectores aleatorios independientes idénticamente distribuidos tales que $X_i \sim N_p(\mu, \Sigma)$ para $i = 1, \cdots, n$. Sean $\overline{X}$ el vector de medias y V la matriz de varianzas-covarianzas muestrales, desconocida la media poblacional, vale decir centrada en la media muestral, tenemos respectivamente

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{y} \qquad V = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^t.$$

Entonces

$$\overline{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right) \qquad \text{y} \qquad (n-1)V \sim W(\Sigma, p, n-1).$$

Distribución del Vector de Medias muestral

Consideremos una muestra aleatoria simple de tamaño n de un vector aleatorio de p componentes con distribución $N_p(\mu, \Sigma)$, digamos $X = (X_1, X_2, \cdots, X_n)^t$. Como $X_i \sim N_p(\mu, \Sigma)$, vale que $X_i - \mu \sim N_p(0, \Sigma)$ para todo $i = 1, \cdots, n$.

Sea U la matriz definida como la suma de los productos entre las desviaciones de cada X_i respecto de la media poblacional y su traspuesta

$$U = X^t X = \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^t.$$

Vale decir la matriz de las observaciones centradas en las medias de la población. La matriz U tiene una distribución de Wishart $U \sim \mathcal{W}(\Sigma, p, n)$.

Distribuciones Muestrales

Caso particular

Para el caso particular en el que $p = 1$, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ resulta ser una estimador insesgado de σ^2, la varianza de la población cuando la media poblacional μ es conocida. Del mismo modo,

$$D = \frac{1}{n}U = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)^t$$

es un estimador insesgado de Σ cuando el vector de medias poblacional μ es conocido. La matriz D es semi-definida positiva y resulta definida positiva cuando Σ es inversible.

Distribución de Hotelling

Definición

La distribución de Hotelling es una generalización de la distribución t-Student. Recordemos que la variable aleatoria t de Student se define como el cociente entre una variable aleatoria Normal estándar y la raíz cuadrada de una variable aleatoria Chi cuadrado, independiente de la variable Normal del numerador, divida por sus grados de libertad. Es decir, si $Z \sim N(0,1)$ y $U \sim \chi^2_n$, entonces $T = \dfrac{Z}{\sqrt{U/n}} \sim t_n$. Como ya hemos visto, en esta variable se basa el estadístico de contraste para la media poblacional de una distribución Normal cuando la varianza poblacional es desconocida, obteniendo que

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}} \sim t_{n-1}.$$

Distribución de Hotelling

Relación con la Distribución F

Consideremos el vector aleatorio $X \sim N_p(0, I_p)$, donde I_p indica la matriz identidad de tamaño $p \times p$.
Sea además la matriz $V \sim \mathcal{W}(I_p, p, n)$ independiente de X, entonces $nX^t V^{-1} X$ sigue una distribución de Hotelling de parámetro n denotada $T^2_{p,n}$. Simbólicamente,

$$nX^t V^{-1} X \sim T^2_{p,n}.$$

Existe una relación entre la distribución de Hotelling y la de Fisher-Snedecor. Si $Q \sim T^2_{p,n}$ es de Hotelling, entonces

$$\frac{n-p+1}{np} Q \sim F_{p,n-p+1},$$

donde $F_{p,n-p+1}$ denota la distribución de F con p y $n-p+1$ g.l.

Estadístico del Test de Hotelling

Por otro lado, si $X \sim N_p(\mu, \Sigma)$ y $U \sim \mathcal{W}(\Sigma, p, n)$, siendo $\overline{X}$ y U independientes, entonces

$$n(\bar{X} - \mu)^t U^{-1}(\bar{X} - \mu) \sim T^2_{p,n}.$$

En particular, si $X_1, X_2, \cdots, X_n$ es una muestra aleatoria, con $X_i \sim N_p(\mu, \Sigma)$ $(i = 1, \cdots, n)$ y V es la matriz de varianzas-covarianzas muestral, se verifica que

$$(n-1)(\bar{X} - \mu)^t V^{-1}(\bar{X} - \mu) \sim T^2_{p,n-1}.$$

Test para comparar medias de dos poblaciones

El Modelo

Sea X una variable aleatoria observada en dos poblaciones. Dadas dos muestras multivariadas independientes de dos poblaciones normales con la misma matriz de varianzas-covarianzas, nos interesa comparar sus vectores medios.

Vamos a contrastar las siguientes hipótesis

$$\begin{cases} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 \neq \mu_2. \end{cases}$$

Test para comparar medias de dos poblaciones

El estadístico de contraste

Sean $\bar{X}_1$ y $\bar{X}_2$ los vectores de medias muestrales, y V_1 y V_2 las matrices de varianzas-covarianzas muestrales. Un estimador insesgado de la diferencia $\mu_1 - \mu_2$ es $\bar{X}_1 - \bar{X}_2$. Como $\bar{X}_1 \sim N_p\left(\mu_1, \frac{1}{n_1}\Sigma\right)$ y $\bar{X}_2 \sim N_p\left(\mu_2, \frac{1}{n_2}\Sigma\right)$, tenemos que

$$\bar{X}_1 - \bar{X}_2 \sim N_p\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right).$$

Distribución

Asumiendo la hipótesis nula como verdadera, $\mu_1 - \mu_2 = 0$ y entonces

$$\bar{X}_1 - \bar{X}_2 \sim N_p\left(0, \left(\frac{n_1 + n_2}{n_1 n_2}\right)\Sigma\right).$$

Estimación de la matriz de Covarianzas

El estimador insesgado para la matriz de covarianzas de las dos poblaciones, Σ, es

$$S = \frac{(n_1 - 1)V_1 + (n_2 - 1)V_2}{n_1 + n_2 - 2},$$

y

$$(n_1 + n_2 - 2)S \sim \mathcal{W}(\Sigma, p, n_1 + n_2 - 2).$$

Relación con la Distancia de Mahalanobis

Como $\bar{X}_1$ es indep de $\bar{X}_2$ y ambas son independientes de S, bajo H_0 vale

$$D^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\bar{X}_1 - \bar{X}_2\right)^t S^{-1} \left(\bar{X}_1 - \bar{X}_2\right) \sim T^2_{p, n_1+n_2-2},$$

donde D resulta la **distancia de Mahalanobis** entre los vectores de medias muestrales.

Distribución del Estadístico

Por la equivalencia con la distribución F

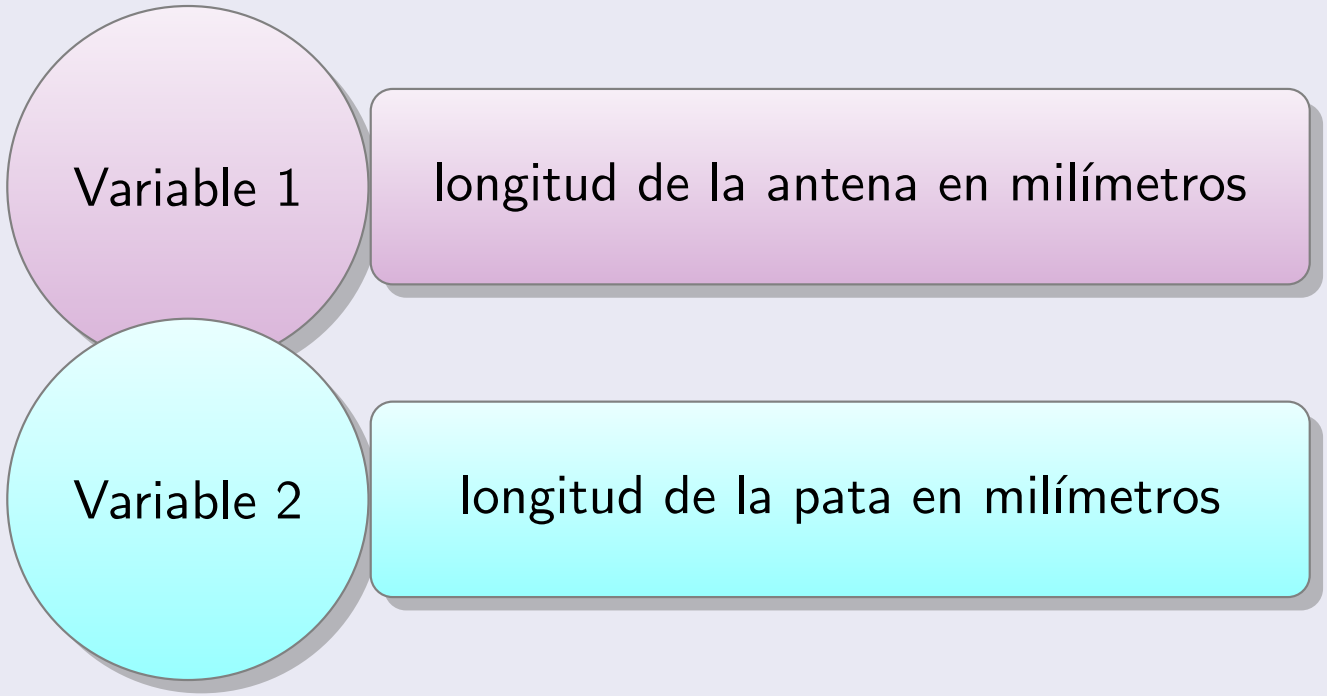
$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} D^2 \sim F_{p, n_1+n_2-p-1}.$$

Ejemplo: Avispas

Se desea comparar dos especies de avispas, conocidas como *chaqueta amarilla* y *negra pequeña*.

Ejemplo: Avispas

Para ello, se consideran las siguientes variables



Para dos muestras independientes de tamaños $n_1 = 9$ y $n_2 = 6$, se han obtenido los datos que figuran en la Tabla

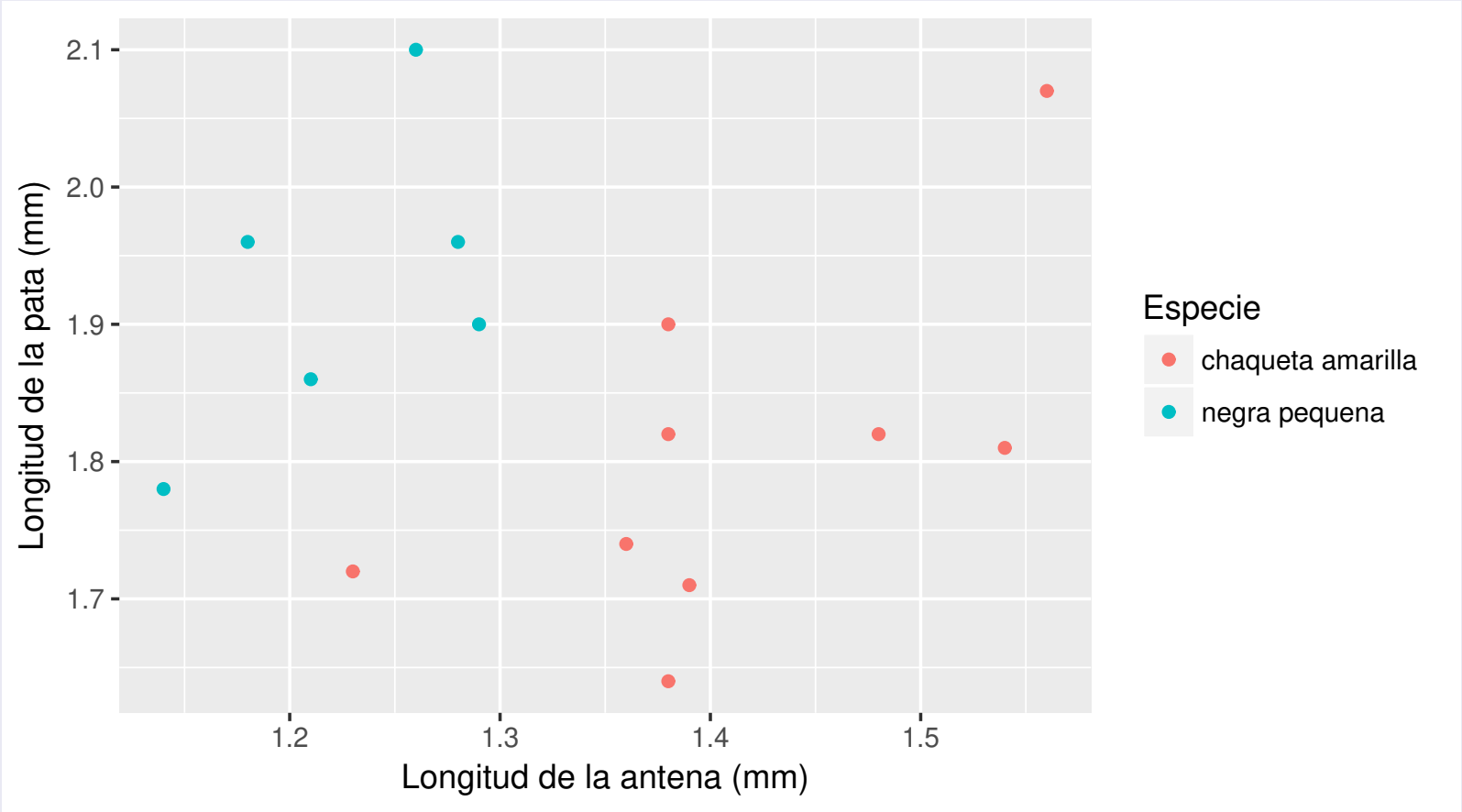
Ejemplo: Avispas

Los Datos

Antena	Pata	Especie
1.38	1.64	chaqueta amarilla
1.39	1.71	chaqueta amarilla
1.23	1.72	chaqueta amarilla
1.36	1.74	chaqueta amarilla
1.38	1.82	chaqueta amarilla
1.48	1.82	chaqueta amarilla
1.54	1.81	chaqueta amarilla
1.38	1.90	chaqueta amarilla
1.56	2.07	chaqueta amarilla
1.14	1.78	negra pequeña
1.21	1.86	negra pequeña
1.18	1.96	negra pequeña

Ejemplo: Avispas

Diagrama de Dispersión por Grupo



Ejemplo: Avispas

Visualicemos las características de cada grupo de Avispas

Los vectores de medias $\bar{X}_1$ y $\bar{X}_2$ correspondientes a las especies 'chaqueta amarilla' y 'negra pequeña' respectivamente están dados por

$$\bar{X}_1 = (1.41, 1.80), \qquad \bar{X}_2 = (1.23, 1.93).$$

Las matrices de varianzas-covarianzas S_1 y S_2 correspondientes a las especies 'chaqueta amarilla' y 'negra pequeña' respectivamente son

$$S_1 = \begin{pmatrix} 0.0103 & 0.0079 \\ 0.0079 & 0.0159 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 0.0036 & 0.0036 \\ 0.0036 & 0.0118 \end{pmatrix}.$$

Ejemplo: Avispas

Visualicemos las características de cada grupo de Avispas

La estimación insesgada de la matriz de varianzas-covarianzas común

$$S = \frac{1}{13}(8S_1 + 5S_2) = \begin{pmatrix} 0.0077 & 0.0063 \\ 0.0063 & 0.0143 \end{pmatrix}.$$

La distancia de Mahalanobis entre las dos medias muestrales,

$$D^2 = (\bar{X}_1 - \bar{X}_2)^t S^{-1} (\bar{X}_1 - \bar{X}_2) = 12.54575.$$

$$T^2_{obs} = \frac{9 \cdot 6}{9+6} D^2_{obs} = 45.1647 \qquad F_{obs} = \frac{9+6-2-1}{2(9+6-2)} T^2_{obs} = 20.8452.$$

Recordemos que el último estimador sigue una distribución $F_{2,12}$.

Análisis de Perfiles

Decisión

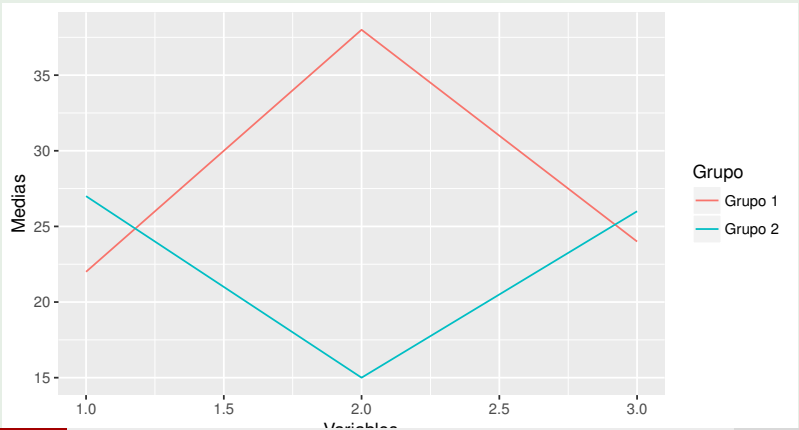
Considerando un nivel de significación del 5%, tenemos que $F_{obs} = 20.84 > F_{2,12,0.95} = 3.885$. Por lo tanto, rechazamos la hipótesis de nulidad, vale decir que no puede sostenerse la hipótesis de igualdad de los vectores medios.

Cuándo se realiza análisis de perfiles?

- El objetivo es comparar el comportamiento promedio de individuos de una o varias poblaciones y se dispone de mediciones repetidas sobre un conjunto de variables relacionadas.
- Las componentes del vector normal de interés no corresponden a diferentes variables sino a una misma variable repetida, por ejemplo en el tiempo o el espacio.
- Para comparar transversalmente y longitudinalmente medias de dos

Análisis de Perfiles

Consideremos dos grupos uno con n_1 individuos y el otro con n_2 individuos. A cada grupo se le aplica un tratamiento distinto y se mide el resultado del tratamiento en p instantes diferentes.
En las abscisas se representan los instantes o repeticiones, mientras que la ordenada indica el valor de la media en este instante. El perfil se construye uniendo las medias observadas en cada uno de los grupos, en este caso en tres instantes distintos. Las tres observaciones son realizadas sobre un mismo individuo o grupo, por lo cual no son independientes.



Análisis de Perfiles

Pretendemos dar respuesta a preguntas del estilo:

- ¿Los grupos se comportan de manera similar durante todo el proceso?
- Gráficamente, ¿las curvas que los definen son paralelas?
- ¿Los grupos tienen un nivel parecido? Es decir, ¿las curvas son del mismo nivel promedio?
- ¿No hay cambio a lo largo del tiempo? Desde el dibujo, ¿la curva promedio es horizontal?

Caso de dos perfiles

Sean μ_1 y μ_2 los vectores de medias poblacionales correspondientes a las dos poblaciones consideradas. Con valores:

$$\mu_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix} \quad \text{y} \quad \mu_2 = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix}.$$

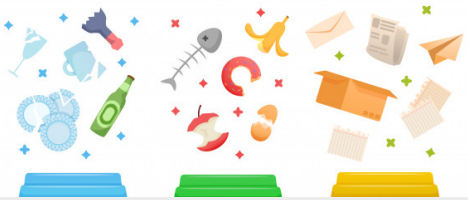
Estamos interesados en saber si los perfiles son idénticos para las dos poblaciones se debe realizar el test que plantea

$$\begin{cases} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 \neq \mu_2. \end{cases}$$

Clasificación

Vectores Medios diferentes...y??

Si los vectores medios de los dos grupos de estudio son diferentes, es probable que el investigador esté interesado en clasificar a un nuevo individuo dentro de alguno de estos grupos. Por ejemplo, si disponemos de un vector de información sobre pacientes que han respondido bien a cierto tratamiento y otros que no han tenido los mismos resultados, podría resultar de interés decidir si a un nuevo paciente le conviene realizar este tratamiento o no. Del mismo modo, podría ocurrir aplicando este análisis para clientes que han cumplido con sus obligaciones y otros que no lo han hecho, con el objeto de decidir si resulta conveniente otorgar a un nuevo cliente una financiación o no.



Clasificación Supervisada: El problema

Se dispone de un conjunto de observaciones

Que se utilizan para entrenar un modelo y se denominan **conjunto de entrenamiento**. Para este conjunto se conocen simultáneamente los valores que asumen las variables de interés y el grupo al cual pertenece cada uno de los individuos. Por este motivo, estas técnicas de clasificación se denominan **supervisadas**.

Organización

Propósito

El análisis discriminante (DA) propuesto por Ronald Fisher

Se propone encontrar una función que depende de un conjunto de variables que denominaremos de aquí en adelante **variables discriminantes**. Esta función al ser aplicada a un nuevo individuo, devuelve un valor que permite asignar a este individuo en alguno de los grupos definidos previamente.

Veremos luego que existen algoritmos de clasificación no supervisada, como el **análisis de conglomerados** o **cluster** (del inglés), donde se desconocen tanto la cantidad de grupos como la pertenencia de los individuos del conjunto de entrenamiento a cada uno de ellos.

LDA

Tabla de Datos disponible

Donde se registraron los valores de p variables observadas sobre N individuos y el grupo de pertenencia. De este modo, la tabla es de tamaño $N \times (p+1)$.
Este análisis señala para cada una de las variables consideradas, su poder clasificatorio que está asociado con su peso en la función discriminante.

Consideremos en primera instancia el caso más sencillo, que es el compuesto por dos muestras independientes que provienen de dos poblaciones normales multivariadas:

- Población 1: $I \in P_1 \rightarrow X_I \sim N_p(\mu_1, \Sigma_1)$,
- Población 2: $I \in P_2 \rightarrow X_I \sim N_p(\mu_2, \Sigma_2)$.

LDA

Cuál es el objetivo?

Se supone conocido que un nuevo individuo I, con vector de observaciones X_I, proviene de alguna de estas dos poblaciones con probabilidades que denominaremos π_1 y π_2 respectivamente. Es decir:

$$P(I \in P_1) = \pi_1 \qquad \text{y} \qquad P(I \in P_2) = \pi_2.$$

Buscamos entonces una regla para predecir a cuál de estas dos poblaciones es más probable que pertenezca un nuevo individuo.

Se han considerado diferentes enfoques para dar respuesta a este problema y algunos a continuación.

Primer enfoque

Basado en la verosimilitud

Esta opción se define en función de la verosimilitud de X_I en cada población. Por lo que, se asigna al sujeto a la Población 1 si $L(X_I, \mu_1, \Sigma_1) > L(X_I, \mu_2, \Sigma_2)$, siendo L la función de verosimilitud que consiste en la función de probabilidad considerada como función del vector de parámetros. Esto significa que se asigna el sujeto a la Población 1 si su función de probabilidad toma un valor superior para el vector de parámetros de la Población 1 que para el vector de parámetros de la Población 2.

Es decir, se decide que $I \in P_1$ cuando

$$\frac{1}{\sqrt{(2\pi)^p}}|\Sigma_1|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X_I-\mu_1)^t\Sigma_1^{-1}(X_I-\mu_1)\right) > \frac{1}{\sqrt{(2\pi)^p}}|\Sigma_2|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X_I-\mu_2)^t\Sigma_2^{-1}(X_I-\mu_2)\right).$$

Caso Particular: $\Sigma_1 = \Sigma_2 = \Sigma$

Esta regla se simplifica decidiendo que $I \in P_1$ cuando

$$\exp\left(-\frac{1}{2}(X_I - \mu_1)^t \Sigma^{-1}(X_I - \mu_1)\right) > \exp\left(-\frac{1}{2}(X_I - \mu_2)^t \Sigma^{-1}(X_I - \mu_2)\right).$$

Realizando cálculos algebraicos:

$$(\mu_1 - \mu_2)^t \Sigma^{-1} X_I > (\mu_1 - \mu_2)^t \Sigma^{-1}\left(\frac{\mu_1 + \mu_2}{2}\right).$$

Dicho de otra manera, si $b = (\mu_1 - \mu_2)^t \Sigma^{-1}$ y $k = (\mu_1 - \mu_2)^t \Sigma^{-1}\left(\frac{\mu_1 + \mu_2}{2}\right)$, asignamos I a la primera población cuando X_I verifica que $bX_I > k$.

Cabe observar que, en general, desconocemos el valor de μ_i para $i = 1, 2$, por lo que se estima con $\overline{X}_i$. Este razonamiento puede extenderse a varios grupos.

Segundo enfoque

Basado en la distancia de Mahalanobis

Se asigna el sujeto x a la Población i si el cuadrado de la distancia de Mahalanobis al vector medio del i-ésimo grupo es menor que el cuadrado de la distancia de Mahalanobis a los restantes grupos. La distancia de Mahalanobis al centro del i-ésimo grupo está dada por

$$D_i^2 = (x - \mu_i)^t \Sigma^{-1} (x - \mu_i).$$

El valor de D_i es una medida de la proximidad de la observación x al vector de medias del i-ésimo grupo (μ_i), considerando la matriz de varianzas-covarianzas de la i-ésima población (Σ_i).

Tercer enfoque

Basado en la Probabilidad a Posteriori

Se asigna el sujeto a la Población i cuando

$$P(I \in P_i/\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}) > P(I \in P_j/\{X_1 = x_1, X_2 = x_2, \cdots, X_n$$

para todo $j \neq i$.

La principal ventaja de la regla de la probabilidad a posteriori es que cuantifica la bondad de la decisión tomada.

Cuidado!!, si bien se le dice 'probabilidad a posteriori', en realidad no es una probabilidad. El hecho es que la observación pertenece a una población o a otra, mientras que la incertidumbre proviene de la capacidad de la regla creada por el investigador para elegir la población correcta.

Tercer enfoque

Ejemplo

Por ejemplo, si contrastamos

$$\begin{cases} P(I \in P_i / \{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}) & = 0.53, \\ P(I \in P_j / \{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}) & = 0.47; \end{cases}$$

no estamos tan seguros de haber clasificado correctamente. Sin embargo, por el contrario si el contraste es

$$\begin{cases} P(I \in P_i / \{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}) & = 0.93, \\ P(I \in P_j / \{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}) & = 0.07; \end{cases}$$

estamos mucho más seguros de haber clasificado correctamente.

Reglas basadas en estimaciones de los parámetros

Desconocemos el verdadero valor de los parámetros poblacionales

Trabajamos con sus estimaciones puntuales.
Para las medias poblacionales utilizamos como estimadores a los vectores de medias muestrales

$$\hat{\mu}_1 = \bar{X}_1 \qquad \text{y} \qquad \hat{\mu}_2 = \bar{X}_2.$$

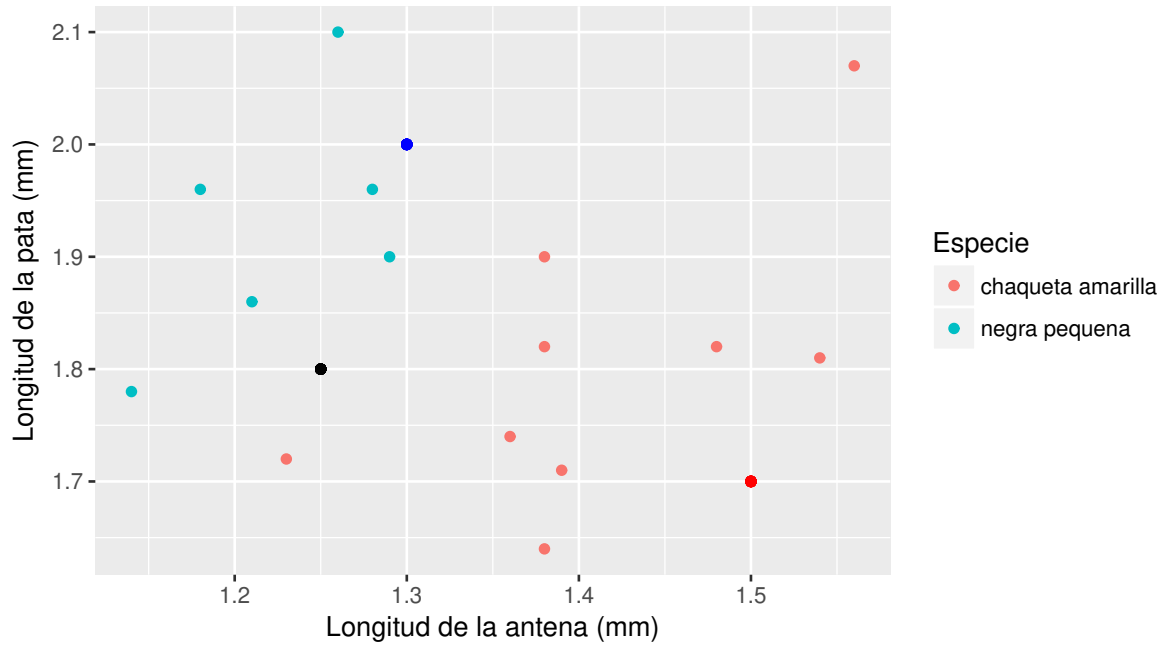
Para la matriz de covarianzas, utilizamos la matriz de varianzas-covarianzas muestral amalgamada, si pueden suponerse iguales, cuya expresión está dada por

$$V = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

S_1 y S_2 son las matrices de covarianzas muestrales de las Poblaciones 1 y

Ejemplo Avispas

En el ejemplo de las dos especies de avispas se ha probado aplicando el test de Hotelling, que las medias de ambas poblaciones son significativamente distintas. Supongamos ahora que tenemos una nueva observación y queremos clasificarla.



Ejemplo Avispas

Nueva Observación

Consideremos una nueva observación con coordenadas $x = (1.25, 1.8)$, que está representada en la Figura en color negro. Si la nueva observación fuera la del punto azul o rojo, no tendríamos muchas dudas respecto de a qué especie asignar la nueva avispa. Sin embargo, no es tan claro a qué grupo debería asignarse la observación del punto negro debido a que se encuentra en una zona fronteriza entre ambos grupos.

Ejemplo Avispas

Destaquemos

que para este ejemplo las observaciones se indican en un color distinto para cada una de las poblaciones, lo cual resulta posible y sencillo dado que sólo hemos observado dos variables. Si hubiéramos observado más variables, esta visualización podría resultar más compleja o incluso, imposible.
En la Figura resulta evidente que las dos variables, dadas por la longitud de la antena y de la pata, permiten discriminar entre estas dos poblaciones de avispas.

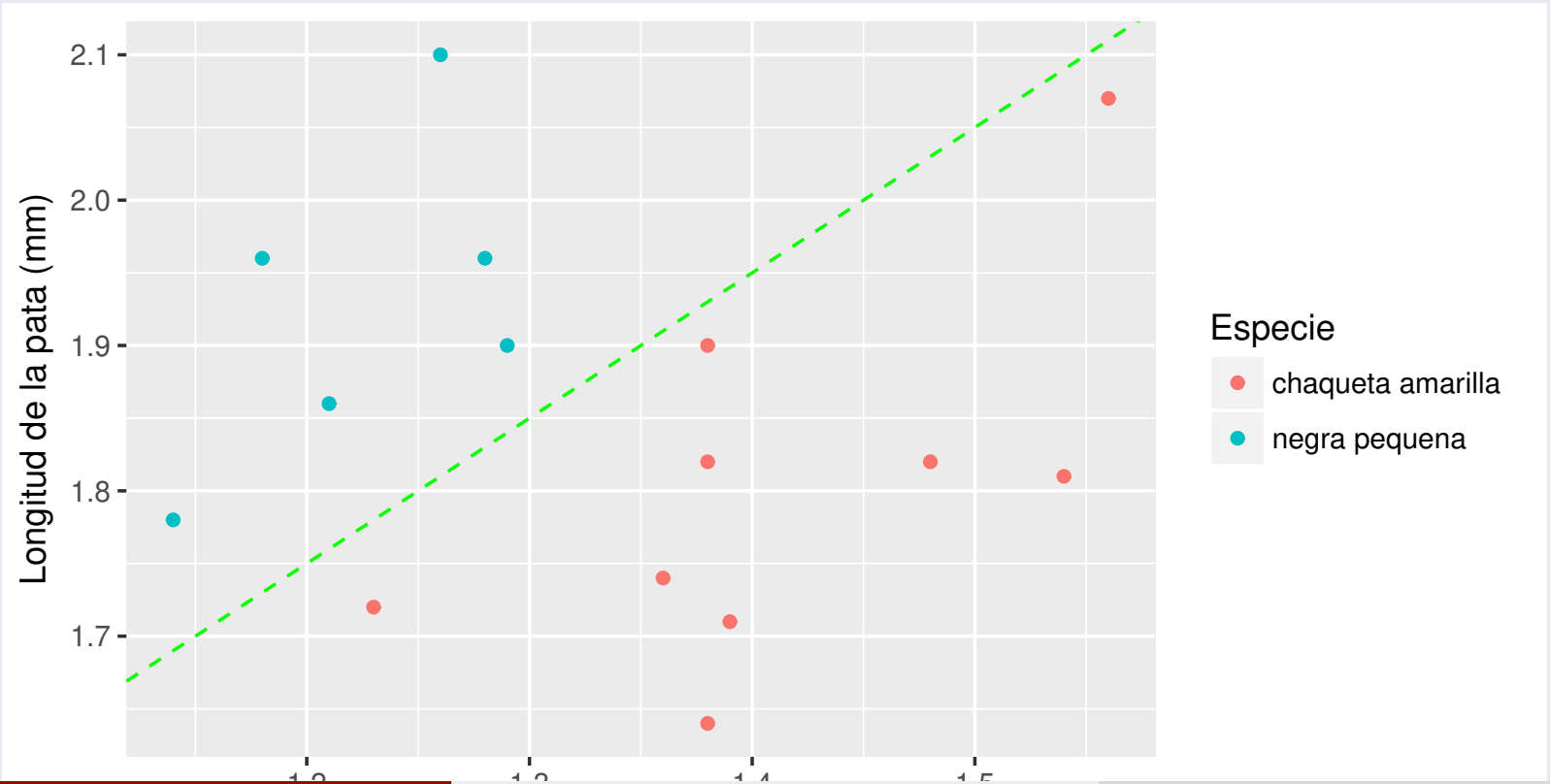
La pregunta que deberíamos hacernos es

¿Cómo trazar una línea que discrimine las poblaciones de la mejor manera posible? O bien, *¿En qué dirección proyectar las observaciones, de modo tal que las proyecciones aparezcan tan separadas como sea posible?*

Ejemplo Avispas

Discriminación Lineal

En la Figura se muestra la línea que mejor discrimina las dos especies de avispas consideradas en el ejemplo.



Variabilidad Total

Descomposición de la SCT

Hemos visto en ANOVA, la suma de cuadrados totales puede descomponerse en la suma de cuadrados entre y dentro de los grupos en el caso univariado.
Extendamos esta idea para las matrices de varianzas-covarianzas.
La matriz de covarianzas total T es la suma de las matrices de covarianzas dentro de los grupos W y entre grupos B.
La expresión matricial es:

$$T = W + B.$$

donde B es la estimación de la variabilidad **between** (dentro) de los grupos y
W es la estimación de la variabilidad **within** (dentro) de los grupos.

Objetivo

Buscamos proyectar las observaciones p-variadas $X = (X_1, X_2, \cdots, X_p)$ sobre una dirección que maximice la separación entre las proyecciones de los distintos grupos de interés.

Es decir que la función discriminante para el caso de estudio, definirá las componentes discriminantes dadas por

$$Y_k = a_k^t X,$$

donde a_k es un vector de coeficientes reales. De este modo,

$$Var(Y) = Var(a^t X) = a^t T a = a^t W a + a^t B a.$$

Qué maximizamos?

¿Cuál es la expresión para estas matrices?

Maximizar la variabilidad entre los grupos con el objetivo de discriminarlos mejor, implica maximizar la varianza entre grupos B, en relación con el total de la varianza T o bien con la varianza entre los grupos W; es decir,

$$\max_a \left\{ \frac{a^t B a}{a^t W a} \right\},$$

lo que equivale a maximizar

$$\max_a \{ a^t W^{-1} B a \}.$$

Este procedimiento coincide con el esquema de maximización que hemos visto para el ACP. Luego, la primera dirección corresponde al autovector asociado al mayor autovalor de la matriz $W^{-1}B$.

Coordenadas Discriminantes

Expresión

Las siguientes coordenadas discriminantes se corresponden con los restantes autovectores de esta matriz y, naturalmente, serán independientes de la primera.
Además, en forma análoga, se puede estimar la proporción de esta separación que logra explicar cada una de las coordenadas discriminantes, como el cociente entre su autovalor y la traza de la matriz $W^{-1}B$.

Se tiene que

$$B = \sum_{i=1}^{g} (\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})^t \qquad \text{y} \qquad W = \sum_{i=1}^{g} (n_i - 1)S_i^2,$$

donde g indica la cantidad de poblaciones de estudio.

Ejemplo Avispas

Estimación de los parámetros

La estimación insesgada de la matriz de varianzas-covarianzas común, es la matriz de varianza amalgamada dada por

$$\begin{pmatrix} 0.0077 & 0.0063 \\ 0.0063 & 0.0143 \end{pmatrix}.$$

En la tabla se muestran los vectores medios total y por grupo.

	Antena	Pata
Media general	1.3373	1.8527
Media chaqueta amarilla	1.4111	1.8033
Media negra pequeña	1.2267	1.9267

Ejemplo Avispas

Estimación de las Matrices de Covarianza

Las matrices de covarianzas dentro de los grupos y entre grupos son respectivamente

$$W = \begin{pmatrix} 0.1002 & 0.0817 \\ 0.0817 & 0.1863 \end{pmatrix} \quad \text{y} \quad B = \begin{pmatrix} 0.0177 & -0.0118 \\ -0.0118 & 0.0079 \end{pmatrix}.$$

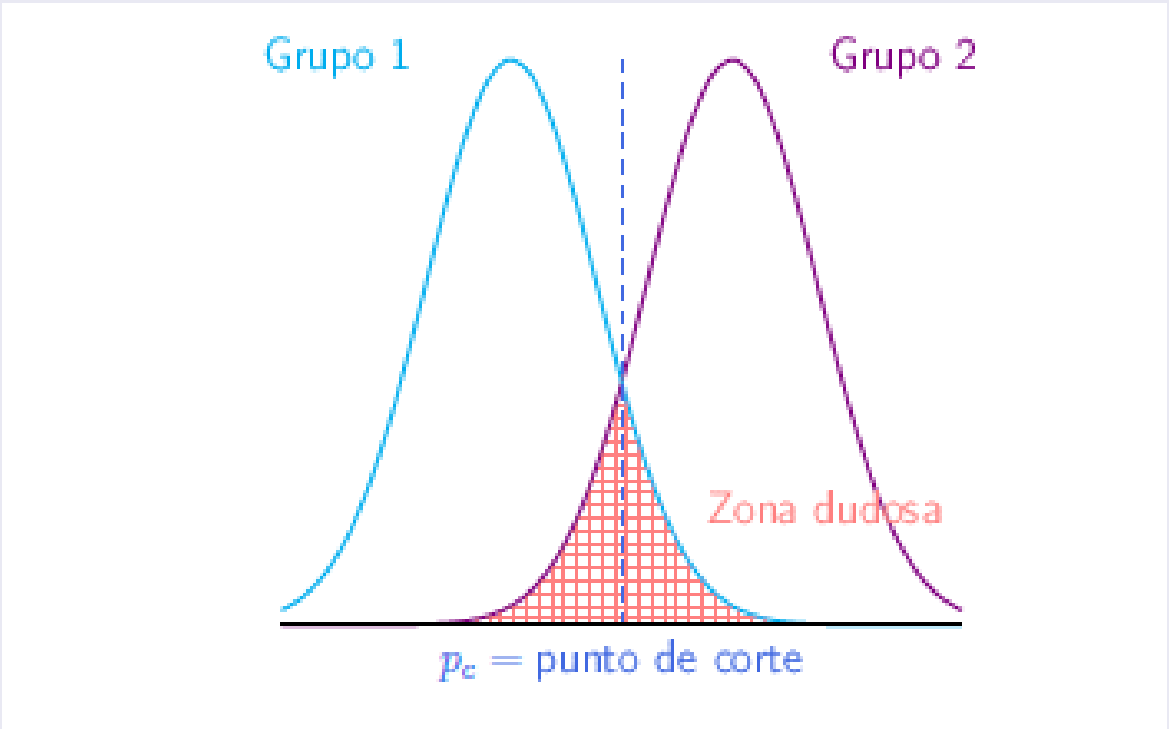
De este modo, la matriz discriminante es

$$W^{-1}B = \begin{pmatrix} 0.3552 & -0.2375 \\ -0.2192 & 0.1466 \end{pmatrix}.$$

Ejemplo Avispas

Errores en la Clasificación

En realidad estamos clasificando las observaciones en función de su proyección sobre un eje. Como se puede apreciar en la Figur, podemos equivocarnos en ambos sentidos.



Ejemplo Avispas

Validación de Supuestos

El análisis discriminante lineal (ADL) sólo es válido si se satisfacen los supuestos distribucionales de normalidad multivariada, independencia de las observaciones y homocedasticidad.
Veamos si se cumplen los supuestos para nuestro ejemplo.

- Normalidad: test de Shapiro-Wilk multivariado
  `W = 0.95885, p-value = 0.6725`
- Homocedasticidad : test M de Box.
  ```
  Box's M-test for Homogeneity of Covariance Matrices
  Chi-Sq (approx.)  = 1.3654, df = 3, p-value = 0.7137
  ```
- Independencia: se cumple por diseño.

No se rechaza el supuesto de normalidad ni el de homocedasticidad. Por lo que estamos en condiciones entonces de calcular la función discriminante lineal

Ejemplo Avispas

La salida del análisis discriminante lineal en R es:

```
Prior probabilities of groups:
 chaqueta amarilla  negra pequeña
               0.6             0.4
Group means:
                    avispas$Antena  avispas$Pata
 chaqueta amarilla         1.411111      1.803333
 negra pequeña             1.226667      1.926667
Coefficients of linear discriminants:
                           LD1
 avispas$Antena  -14.601952
 avispas$Pata      9.011903
```

Ejemplo Avispas

La expresión de la función discriminante lineal es:

$$f(X_1, X_2) = -14.6X_1 + 9.01X_2,$$

donde X_1 y X_2 son las variables de interés dadas respectivamente por la longitud de la antena y de la pata de la avispa.
R agrega una constante a esta función discriminante, en este caso el valor de dicha constante es 2.83. Es decir que la función que usa para calcular las coordenadas discriminantes es

$$f_R(X_1, X_2) = 2.83 - 14.6X_1 + 9.01X_2.$$

Dado que la constante es la misma para todos las observaciones, la clasificación no cambia. En los cálculos que siguen omitimos esta constante.

Ejemplo Avispas

Expresión de la Función Discriminante

El método que estamos aplicando, proyecta los puntos de las observaciones sobre la dirección que mejor discrimina entre los grupos. Hallemos entonces la proyección de los dos centroides con la dirección de proyección mostrada.
Indicamos con

$$\text{proy}_a \qquad \text{y} \qquad \text{proy}_n$$

a las proyecciones de las especies chaqueta amarilla y negra pequeña respectivamente:

$$\text{proy}_a = -14.6 \cdot 1.4111 + 9.01 \cdot 1.8033 = -4.3545,$$

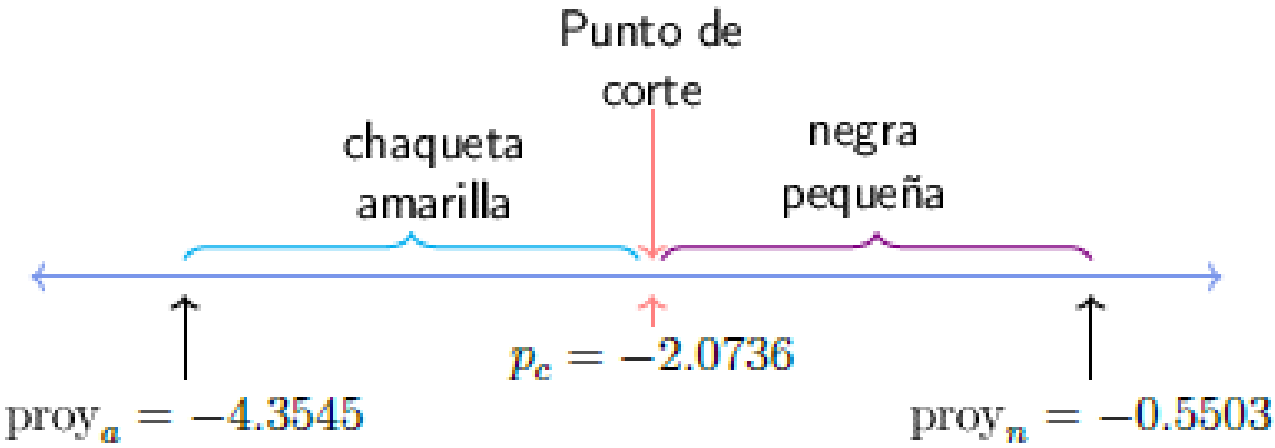
$$\text{proy}_n = -14.6 \cdot 1.2267 + 9.01 \cdot 1.9267 = -0.5503.$$

Ejemplo Avispas

Cómo decidimos a qué grupo pertenece la nueva observación?

Ahora tenemos que encontrar un punto de corte que notamos p_c, para lo cual medimos la distancia entre las dos proyecciones anteriores, teniendo en cuenta la proporción de cada grupo en el conjunto de observaciones. Esto nos permite establecer el valor a partir del cual se separan los grupos:

$$p_c = \text{proy}_a + 0.6(\text{proy}_n - \text{proy}_a) = -2.0736.$$



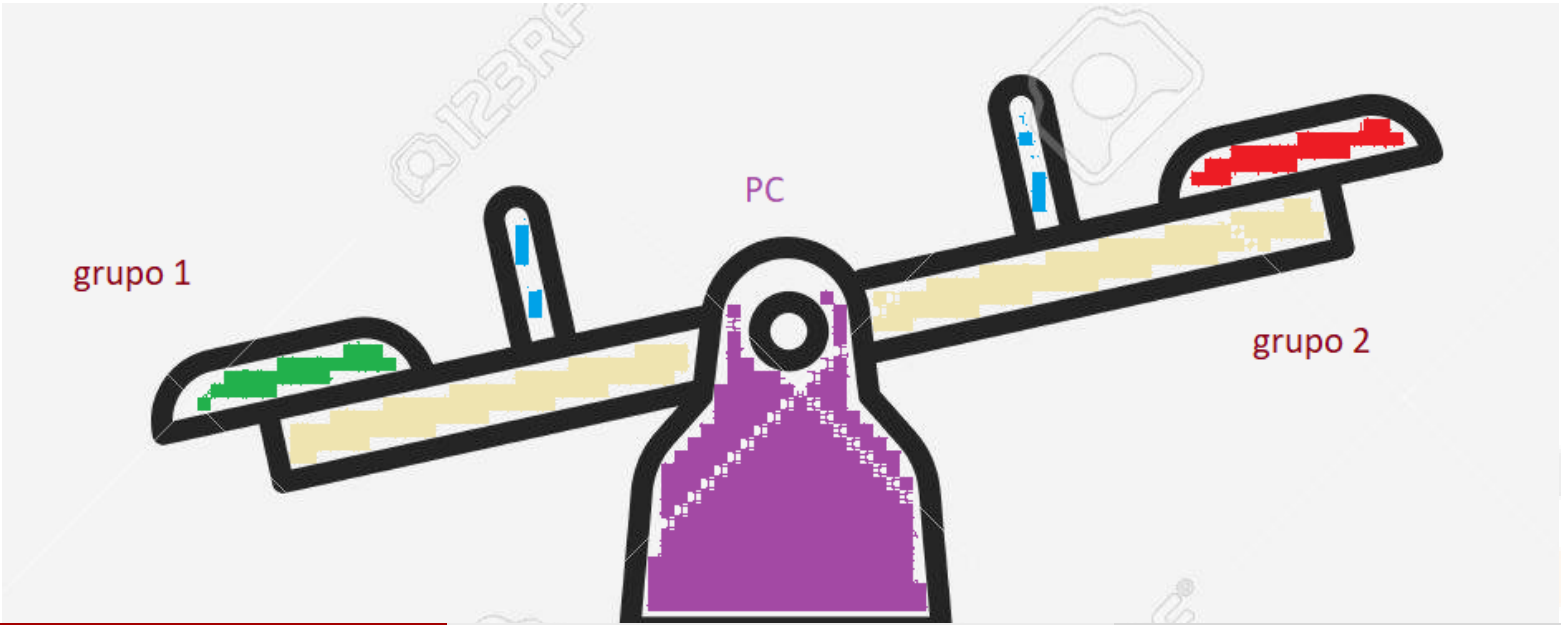
Interpretación de la Función Discriminante

Esto significa que si la proyección de una observación resulta:

a) superior a -2.0736, se clasificará al individuo como 'negra pequeña'

b) mientras que en caso contrario se lo clasificará como 'chaqueta amarilla'

Luego, con la regla establecida del punto de corte en $p_c = -2.0736$ se hizo la asignación a uno de los dos grupos.



Coordenadas Discriminantes

Antena	Pata	Proyección	Clasificación
1.38	1.64	-5.3716	chaqueta amarilla
1.39	1.71	-4.8869	chaqueta amarilla
1.23	1.72	-2.4608	chaqueta amarilla
1.36	1.74	-4.1786	chaqueta amarilla
1.38	1.82	-3.7498	chaqueta amarilla
1.48	1.82	-5.2098	chaqueta amarilla
1.54	1.81	-6.1759	chaqueta amarilla
1.38	1.90	-3.0290	chaqueta amarilla
1.56	2.07	-4.1253	chaqueta amarilla
1.14	1.78	-0.6062	negra pequeña
1.21	1.86	-0.9074	negra pequeña
1.18	1.96	0.4316	negra pequeña
1.28	1.96	-1.0284	negra pequeña
1.26	2.10	0.5250	negra pequeña

Ejemplo Avispas

Nueva Observación: la clasificamos

En la Tabla, se muestra la proyección para cada observación utilizando la función discriminante hallada.
Observar que la clasificación coincide con la especie observada originalmente. Para la nueva observación $(1.25, 1.8)$, calculamos

$$\text{proy}(1.25, 1.8) = -14.6 \cdot 1.25 + 9.01 \cdot 1.8 = -2.032 > p_c.$$

Luego, clasificamos al nuevo individuo dentro del grupo 'negra pequeña'.

La solución aplicada al ejemplo, tal como la hemos presentado, corresponde al **Análisis Discriminante Lineal de Fisher**.
Se llaman **puntuaciones discriminantes** a los valores que se obtienen al evaluar la función discriminante f para cualquier individuo.

Ejemplo Avispas

Forma Matricial de la Función Discriminante

La función discriminante es de la forma

$$D = f(X_1, X_2, \cdots, X_p) = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p.$$

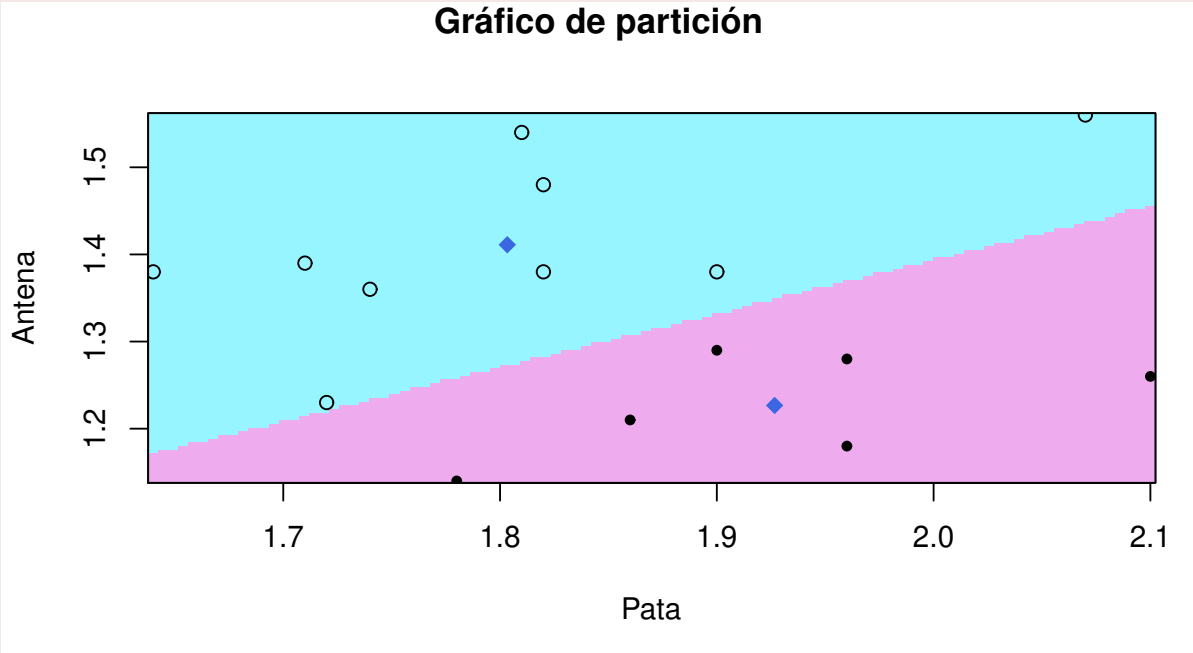
Pudiendo expresar las coordenadas discriminantes de manera matricial

$$\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{pn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}.$$

Ejemplo Avispas

Gráfico de la Clasificación

En este gráfico, los puntos llenos indican las observaciones correspondientes a la especie negra pequeña, los puntos vacíos las observaciones correspondientes a la especie chaqueta amarilla y los rombos azules las medias de cada especie.



Bondad de Clasificación

Matriz de Confusión

Después de construir la regla de discriminante, interesa conocer su capacidad para discriminar.
En caso de no tener regla alguna, se "tira la moneda" y se asigna los sujetos a un grupo mediante el azar, con lo cual la capacidad de discriminar correctamente sería del 50%.
Resulta obvio la preferencia de una regla que no se equivoque en ningún caso, o bien, que clasifique correctamente en un 95% de los casos.
Lamentablemente, esto no siempre es posible.
La **matriz de confusión** da una idea de la tasa de clasificaciones incorrectas por grupo y global.

Bondad de la Clasificación

Matriz de Confusión Avispas

En el ejemplo de las avispas todas las observaciones fueron bien clasificadas, situación no habitual. La matriz de confusión en este caso es diagonal y se exhibe en la Tabla.

	chaqueta amarilla	**negra pequeña**
chaqueta amarilla	9	0
negra pequeña	0	6

En este ejemplo en particular, al conocerse el grupo de pertenencia de cada individuo, se puede comprobar la efectividad del método de clasificación observando el porcentaje de casos bien clasificados.

Bondad de la Clasificación

Para estimar la probabilidad de clasificación correcta disponemos de tres alternativas

- **Clasificación ingenua:** se utilizan los mismos datos para construir la regla y para estimar la probabilidad de clasificación correcta.
- **Muestra de entrenamiento y de validación:** se subdividen los datos en dos submuestras al azar con, aproximadamente, 2/3 para construir o entrenar la regla y 1/3 restante para validarla. La mayor de las submuestras, **training sample** o **muestra de entrenamiento** y la menor de las submuestras, denominada **muestra de validación**.
- **Validación cruzada, cross validation, o leave one out:** se eliminan de a una las observaciones y se construye la regla sin ella y se la clasifica a esta observación con dicha regla. Finalmente, se estima la probabilidad de buena clasificación considerando la proporción de observaciones bien clasificadas de esta manera.

Bondad de la Clasificación

Observaciones

- Este análisis sólo tiene sentido cuando las medias de ambos grupos difieren significativamente.
- La ausencia de normalidad multivariante o la presencia de *outliers* conlleva a problemas en la estimación de los parámetros.
- Las matrices de varianzas-covarianzas distintas requieren el uso de técnicas de clasificación cuadráticas. Estas técnicas son conocidas como Análisis Discriminante Cuadrático de Fisher.
- En la validación cruzada es importante entender que para cada elemento que se elimina, se construye una regla distinta y por lo tanto, los coeficientes de la misma pueden variar de un caso a otro.
- La multicolinealidad genera problemas en la interpretación de los coeficientes de las variables.

Cuando hay más de dos grupos

Alternativas

Cuando el número de grupos es mayor que 2 y se sostienen los supuestos de homocedasticidad y normalidad multivariada son válidas las siguientes dos propuestas:

- **Primera propuesta:** se calcula la distancia de Mahalanobis al centroide (media) de cada grupo y un nuevo individuo se clasifica en el i-ésimo grupo si el valor de la distancia de Mahalanobis a ese grupo es la menor de todas.
- **Segunda propuesta:** se calcula la probabilidad a posteriori de que una nueva observación pertenezca a cada uno de los grupos. Se clasifica la observación en el grupo que maximiza dicha probabilidad o la función de verosimilitud.

Análisis Discriminante Lineal

Supuestos

Como ya hemos mencionado en varias oportunidades, el análisis discriminante lineal sólo es válido para el caso en que

- las variables originales tienen **distribución Normal Multivariada** .
- se puede suponer que las **matrices de varianzas-covarianzas son iguales para todos los grupos**

Cuando la distribución conjunta es Normal Multivariada

1. cualquier C.L. de sus componentes se distribuye normalmente.
2. esto se verifica para c/u de las coord. del vector de observaciones.
3. si alguna de las variables originales no se distribuye de manera normal, entonces la distribución conjunta no es normal multivariada.
4. Luego, bastará con que una componente no tenga distribución normal para asegurar que la distribución conjunta no es normal multivariada.

Validación de los Supuestos

Cómo se prueba la normalidad multivariada?

Sin embargo, si todas las componentes tienen distribución normal univariada, cabe preguntarse si este hecho es suficiente para poder sostener que la distribución conjunta es normal multivariada.

La respuesta es no.

Para testear el supuesto de normalidad multivariada es necesario aplicar un test de bondad de ajuste. Mientras que para comprobar el supuesto de homocedasticidad se puede aplicar la prueba M de Box. La hipótesis nula de esta prueba es que las matrices de varianzas-covarianzas de los grupos son iguales.

Homocedasticidad

Comparación de Matrices de Covarianza

Para comparar las matrices de varianzas-covarianzas se utiliza el determinante de la matriz de varianzas-covarianzas de cada uno de los grupos. Así como el test de Bartlet para el caso univariado, el test M de Box en el caso multivariado es sensible a la falta de normalidad. Por otra parte, si las muestras son de tamaños grandes, este test pierde efectividad, resultando más fácil rechazar la hipótesis nula.
Una **alternativa robusta** para esta prueba es el test de Levene para datos multivariados. La prueba de Levene univariada realiza un análisis de la varianza sobre los valores absolutos de las diferencias entre los valores observados y el centro del grupo, que puede tomarse como la media o su estimación robusta en el caso de querer evitar la influencia de *outliers*.

Denotando por x_{ij} al j-ésimo punto del i-ésimo grupo, y por x_{ij_k} a la k-ésima coordenada de x_{ij}, se define

$$d_{ij}^{c} = \Delta(x_{ij}, c_i),$$

donde el **vector centroide** se define como el punto que minimiza la suma de cuadrados de las distancias a cada punto del grupo

$$c_i = \min_{c} \sum_{j=1}^{n_i} (d_{ij}^{c})^2,$$

siendo n_i la cantidad de elementos del i-ésimo grupo y

$$\Delta(x_{ij}, c_i) = \sqrt{\sum_{k=1}^{p} (x_{ij_k} - c_i)^2}.$$

Este vector centroide usualmente se asume como el vector de medias muestrales, definido como $\bar{x} = (\bar{x}_1, \cdots, \bar{x}_p)$. Luego, la estadística F del ANOVA que se utiliza para probar H_0 es de la forma

$$F = \frac{(N-g)\sum_{i=1}^{g} n_i (D_{i.}^c - D_{..}^c)^2}{(g-1)\sum_{i=1}^{g}\sum_{j=1}^{n_i}(d_{ij}^c - D_{i.}^c)^2}.$$

donde g es la cantidad de grupos, n_i el total de observaciones del i-ésimo grupo y $N = \sum_{i=1}^{g} n_i$ es el total de observaciones. Además, $D_{i.}^c = \sum_{j=1}^{n_i} d_{ij}^c$ es la suma de distancias de las observaciones del i-ésimo grupo a su centroide y $D_{..} = \sum_{i=1}^{g} D_{i.}^c$ es la suma de todas las distancias.

Cuál es el estadístico de la Prueba?

Cuál es la distribución del estadístico?

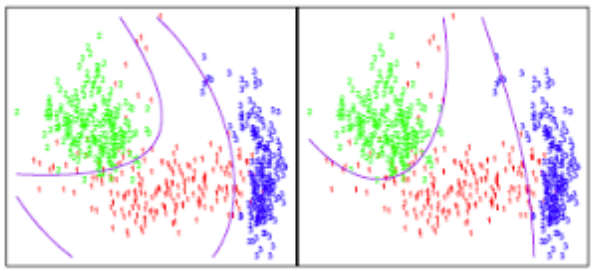
Bajo la suposición de H_0, el estadístico F de ANOVA sigue aproximadamente una distribución F de Fisher con $g-1$ y $N-g$ grados de libertad. Anderson propone el uso de medianas como una opción robusta para la prueba de Levene.

En este caso, la mediana p-dimensional no necesariamente está definida como el vector de medianas individuales para cada variable. En algunos programas estadísticos, como R, se encuentran rutinas implementadas para encontrar medianas espaciales de este tipo.

¿Qué hacer si se rechaza el supuesto de homocedasticidad?

Si se rechaza la hipótesis $H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$

Una alternativa es estandarizar por separado cada grupo con su respectiva matriz de varianzas-covarianzas estimada, de modo tal que los nuevos grupos con las observaciones transformadas tienen igual matriz de varianzas-covarianzas (identidad) y, sobre este nuevo espacio de representación, calcular el discriminante lineal. En este caso, se proyectan todos los datos y se estima la varianza de las puntuaciones discriminantes dentro de cada grupo.



Pertinencia del Análisis Discriminante

Consideraciones para la selección de variables del modelo

- Si se encuentra que una variable asume valores medios muy diferentes en los distintos grupos, es probable que resulte una buena variable para discriminar.
- En segunda instancia, con las variables para las cuales se notó diferencia, es conveniente testear la igualdad de vectores medios entre los grupos.
- Es conveniente estudiar si las matrices de varianzas-covarianzas de los distintos grupos y del grupo general son similares o no.
- Sólo en el caso de hallar diferencias significativas entre los vectores medios de los grupos y siendo que las matrices de varianzas-covarianzas resultaron similares, tendrá sentido utilizar la función discriminante lineal.

Interpretacion de los Coeficientes del Modelo

Problemas de Multicolinealidad

Los coeficientes estandarizados de la función discriminante son los que corresponden al cálculo de la función discriminante con todas las variables clasificadoras estandarizadas. Este recurso se utiliza para evitar ciertos problemas de escala que pudieran existir entre las variables.
Los coeficientes estandarizados a_{ij} pueden interpretarse como indicadores de la importancia relativa de cada una de las variables en cada función discriminante. De esta manera, si la variable x_j es importante en la función discriminante y_i, su respectivo coeficiente a_{ij} será grande en valor absoluto. Dicho de otro modo, hay una fuerte asociación entre la variable x_j y la proyección y_i.
Estos coeficientes son poco fiables si existen problemas de multicolinealidad entre las variables clasificadoras. Al estar correlacionadas las variables originales, y a veces en forma significativa, **es conveniente**

Costos de Clasificación

En algunas ocasiones es necesario ponderar los errores cometidos.

Por ejemplo, es más costoso no indicar que un paciente tiene rechazo a un órgano trasplantado, cuando efectivamente lo tiene, que indicar que sí lo tiene cuando en realidad esto no es así. En el primer caso el paciente puede agravarse, mientras que en el segundo caso es posible que se le de una medicación o se aumente la que está recibiendo en forma innecesaria. Una manera de diferenciación en la regla discriminante entre los dos tipos de errores posibles es asignar un costo a cada error. También podría utilizarse información previa, como por ejemplo si uno supiera que el rechazo a un órgano trasplantado ocurre en a lo sumo el 20% de los pacientes.

Costos de Clasificación

Formalización de la Idea

- P_1 con función de densidad $f(x, \theta_1)$ donde θ_1 es un vector de parámetros que caracteriza a la densidad de la primera población.
- P_2 con función de densidad $f(x, \theta_2)$ donde θ_2 es un vector de parámetros que caracteriza a la densidad de la segunda población.

Una regla discriminante general, partirá al espacio p-dimensional en dos regiones R_1 y R_2 de modo tal que, si una observación pertenece a R_1 será clasificada en el primer grupo y en caso contrario, será clasificada en el segundo grupo.

.

Costos y Probabilidades de Pertenencia

Llamamos $C(i/j)$ al costo de clasificar a un individuo en la i-ésima población cuando en realidad el mismo pertenece a la j-ésima población y $P(i/j)$ a la probabilidad de que esto ocurra. También designamos con p_i a la probabilidad de que un individuo pertenezca al i-ésimo grupo

¿Cuál será entonces el costo promedio de clasificación errónea de una observación seleccionada de la población general de forma aleatoria?

el costo total está dado por

$$CT = p_1 C(2/1)P(2/1) + p_2 C(1/2)P(1/2).$$

Regla: asignamos un individuo a la primera población cuando se verifica que

$$p_1 C(2/1)P(2/1) < p_2 C(1/2)P(1/2),$$

vale decir si

$$p_1 C(2/1)f(x, \theta_2) < p_2 C(1/2)f(x, \theta_1).$$

Caso Particular

Si las probabilidades de pertenencia a los grupos son iguales

$X \in P_1$ si se verifica que $C(2/1)f(x, \theta_2) < C(1/2)f(x, \theta_1)$.

Más aún, si consideramos los costos de mala clasificación para ambos grupos iguales, $X \in P_1$ si vale que

$$f(x, \theta_2) < f(x, \theta_1)$$

.

Concluimos que, en el caso de costos iguales de clasificación y probabilidades de pertenencia a cada grupo idéntica:

La regla se reduce a maximizar la función de verosimilitud.

Organización

QDA

Análisis discriminante cuadrático

Cuando el supuesto de homocedasticidad no puede sostenerse, una opción es utilizar el análisis discriminante cuadrático de Fisher que construye la regla: $x \in P_1$ cuando

$$\frac{1}{\sqrt{(2\pi)^p}}|\Sigma_1|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)\right) > \quad (1)$$

$$\frac{1}{\sqrt{(2\pi)^p}}|\Sigma_2|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2)\right) \quad (2)$$

siendo $\Sigma_1 \neq \Sigma_2$.

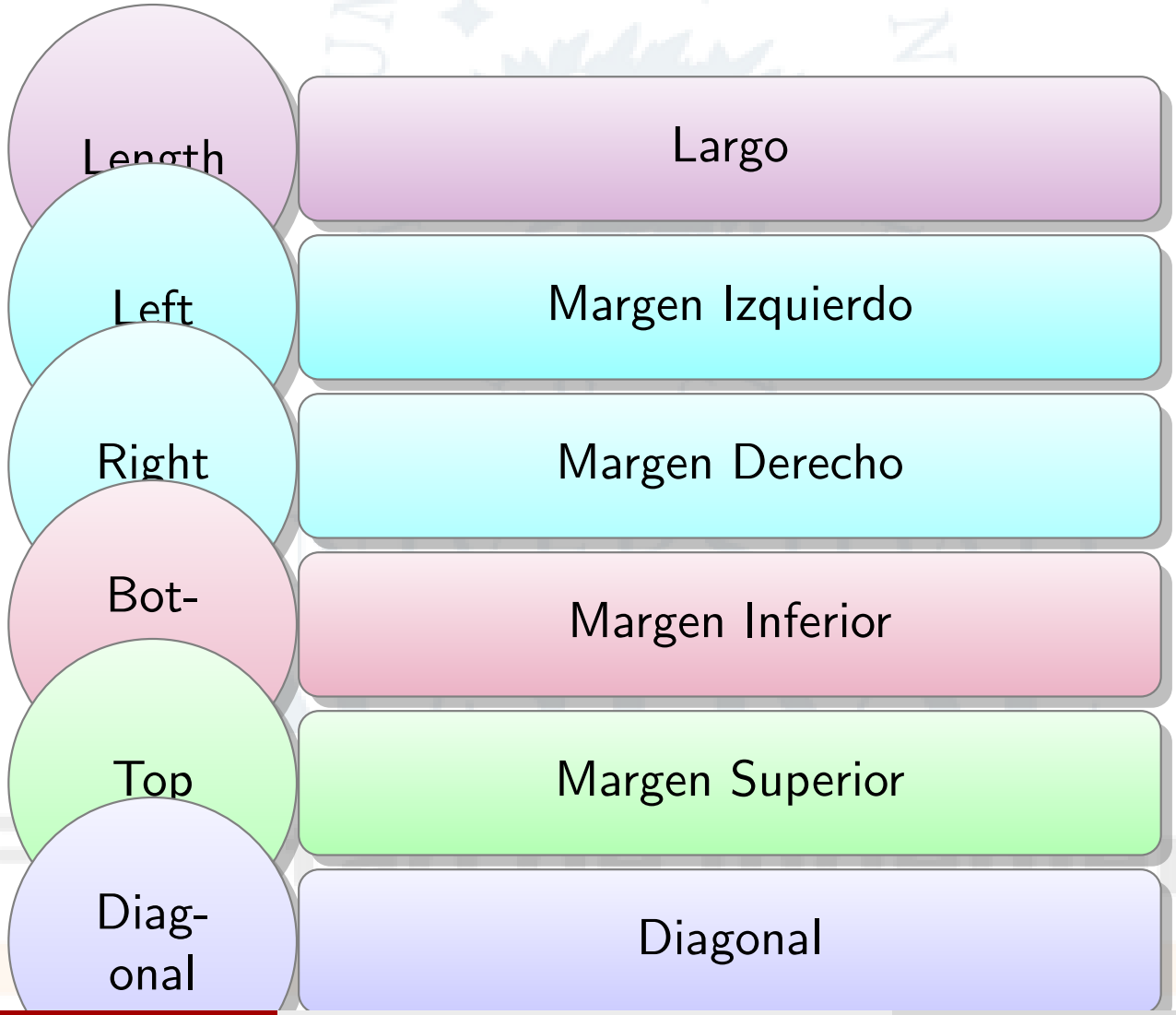
El discriminante se dice **cuadrático** porque el término de segundo orden no se cancela. Esto ocurre porque no se satisface la igualdad de las matrices de varianzas y covarianzas de los grupos QDA(*quadratic discriminant analysis*).

Ejemplo Billetes

Datos Banknote de la biblioteca `mclust` de `R`

En la misma se dispone de medidas sobre 200 billetes de los cuales 100 son legítimos y los otros 100 son falsos (apócrifos). Estos datos corresponden a imágenes de aspecto similar. Se utilizó una cámara industrial para la digitalización de las imágenes con una resolución de 660 ppp aproximadamente. Cada imagen tiene 400×400 *pixeles* y están en escala de grises. La herramienta de transformación de *Wavelet* se usó para extraer características de las imágenes.

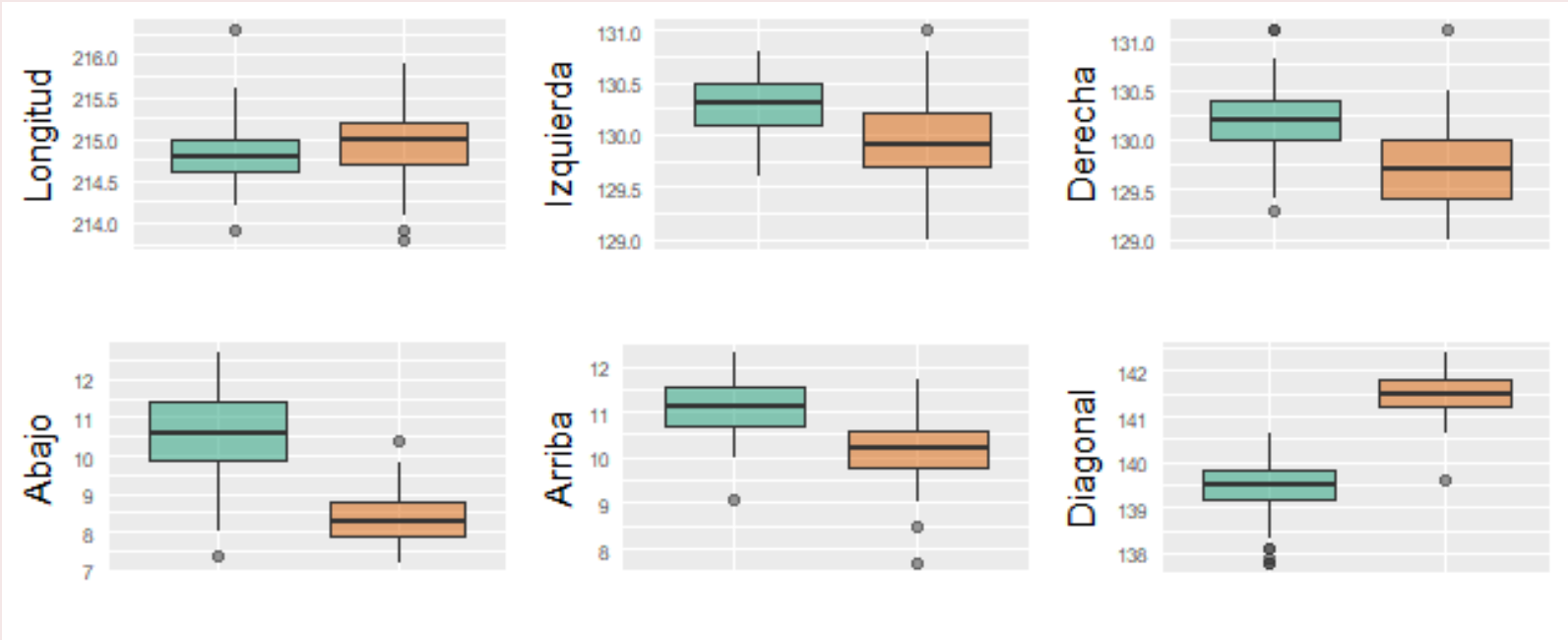
Variables Consideradas (status: legítimo o no)



Ejemplo Billetes

Análisis univariado gráfico

En la Figura se pueden apreciar los *boxplots* de cada una de estas variables en los grupos definidos por la variable categórica `Status`. En la misma, todas las variables analizadas parecen discriminar entre los billetes legítimos y los apócrifos.



Test de Hotelling

Comparación de vectores Medios

La salida correspondiente en `R` incluye los siguientes resultados:

- El estadístico de contraste: `$stats$statistic` 2412.451.
- Los grados de libertad: `$stats$df` 6 193.
- El número de observaciones del primer grupo: `$stats$nx` 100.
- El número de observaciones del segundo grupo: `$stats$ny` 100.
- El número de variables consideradas o bien la dimensión de los vectores que se comparan: `$stats$p` 6.
- El p-valor del contraste: `$pval` 0.

La decisión es por lo tanto rechazar la hipótesis de igualdad de vectores medios.

Ejemplo Billetes

Validación de Supuestos: Normalidad Multivariada

Se observa severos apartamientos del supuesto de normalidad univariada lo que indica que no se cumple la normalidad multivariada.
Sin embargo, vamos a aplicar el test de normalidad multivariada de Shapiro-Wilk para analizar este supuesto.
La salida de esta prueba en R es:

```
Shapiro-Wilk normality test
data:  Z
W = 0.95953, p-value = 1.758e-05
```

Por lo tanto, se rechaza el supuesto de normalidad multivariada.

Ejemplo Billetes

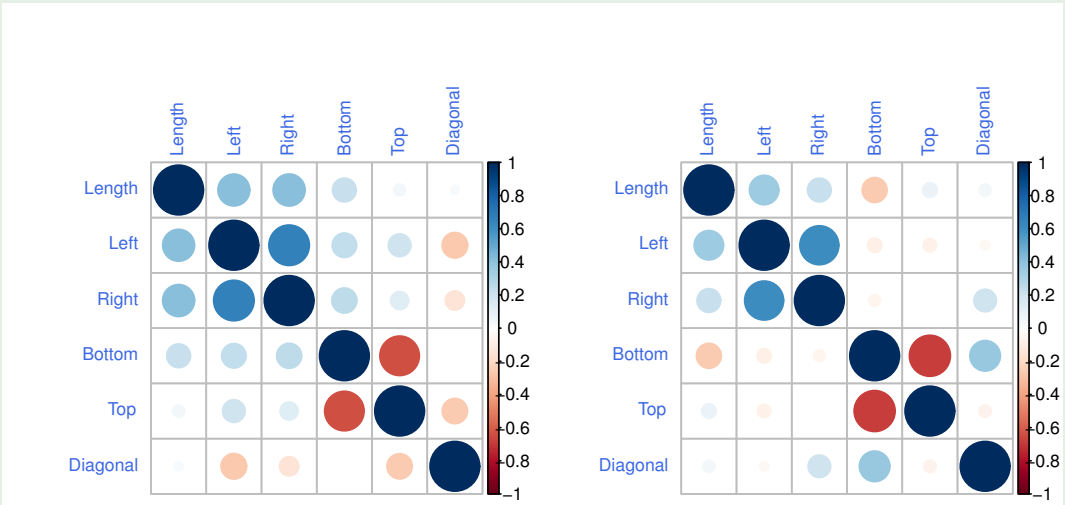
Homocedasticidad

Al efectuar el análisis del supuesto de homocedasticidad, la salida correspondiente en R es:

```
Box's M-test for Homogeneity of Covariance Matrices
Chi-Sq (approx.)  = 121.9, df = 21, p-value = 3.198e-16
```

Con lo cual también **se rechaza el supuesto de homocedasticidad**.

Es similar la estructura de la matriz de correlación de los grupos?



Ejemplo Billetes

Aplicamos el análisis discriminante cuadrático

A pesar de saber que no se satisface el supuesto de normalidad. Aún en este caso nos interesa ver cómo clasifica el método.

Las medidas resumen por grupo resultan:

```
Group means:
               Length   Left  Right  Bottom   Top  Diagonal
counterfeit     214.8  130.3  130.2    10.5  11.1     139.5
genuine         214.9  129.9  129.7     8.3  10.2     141.5
```

Qué se aprecia y qué no se puede apreciar en esta salida?

Ejemplo Billetes

Matriz de Confusión correspondiente a la clasificación ingenua

	Clase predicha	
Clase real	**Apócrifo**	**Genuino**
Apócrifo	100	0
Genuino	1	99

La tasa de error estimada en forma ingenua es: 0.5%.

Ejemplo Billetes

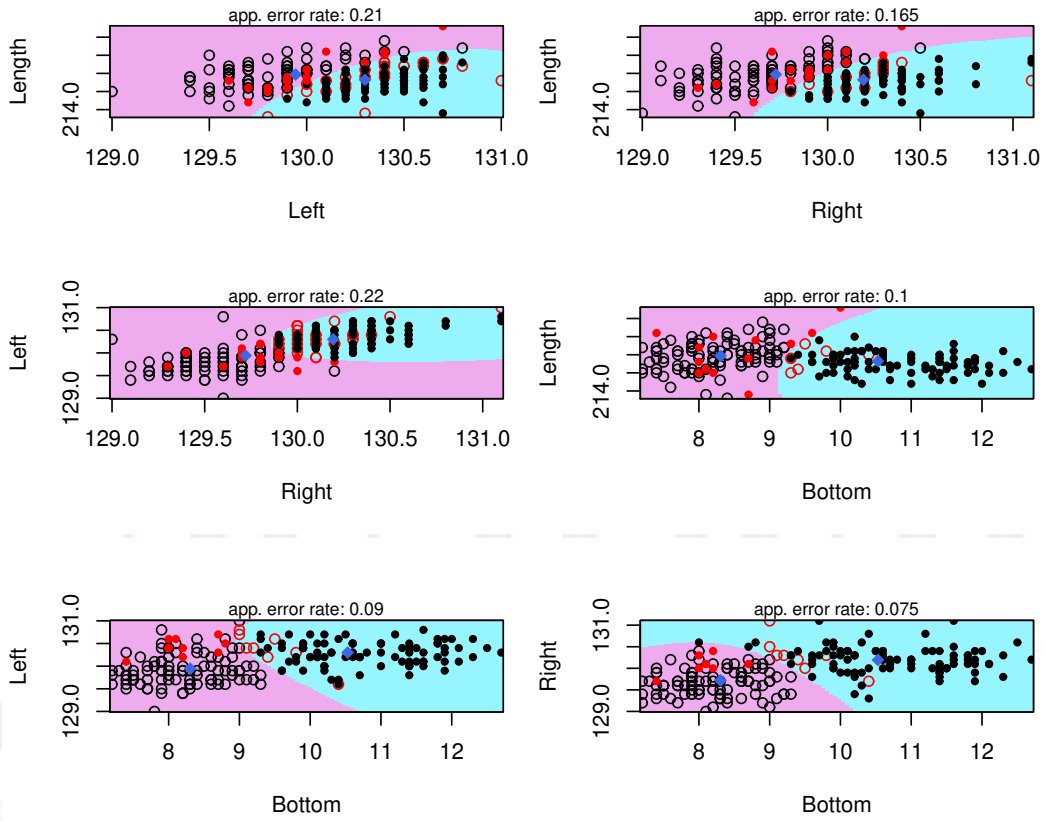
Matriz de confusión correspondiente a la muestra de validación

	Clase predicha	
Clase real	**Apócrifo**	**Genuino**
Apócrifo	32	1
Genuino	0	47

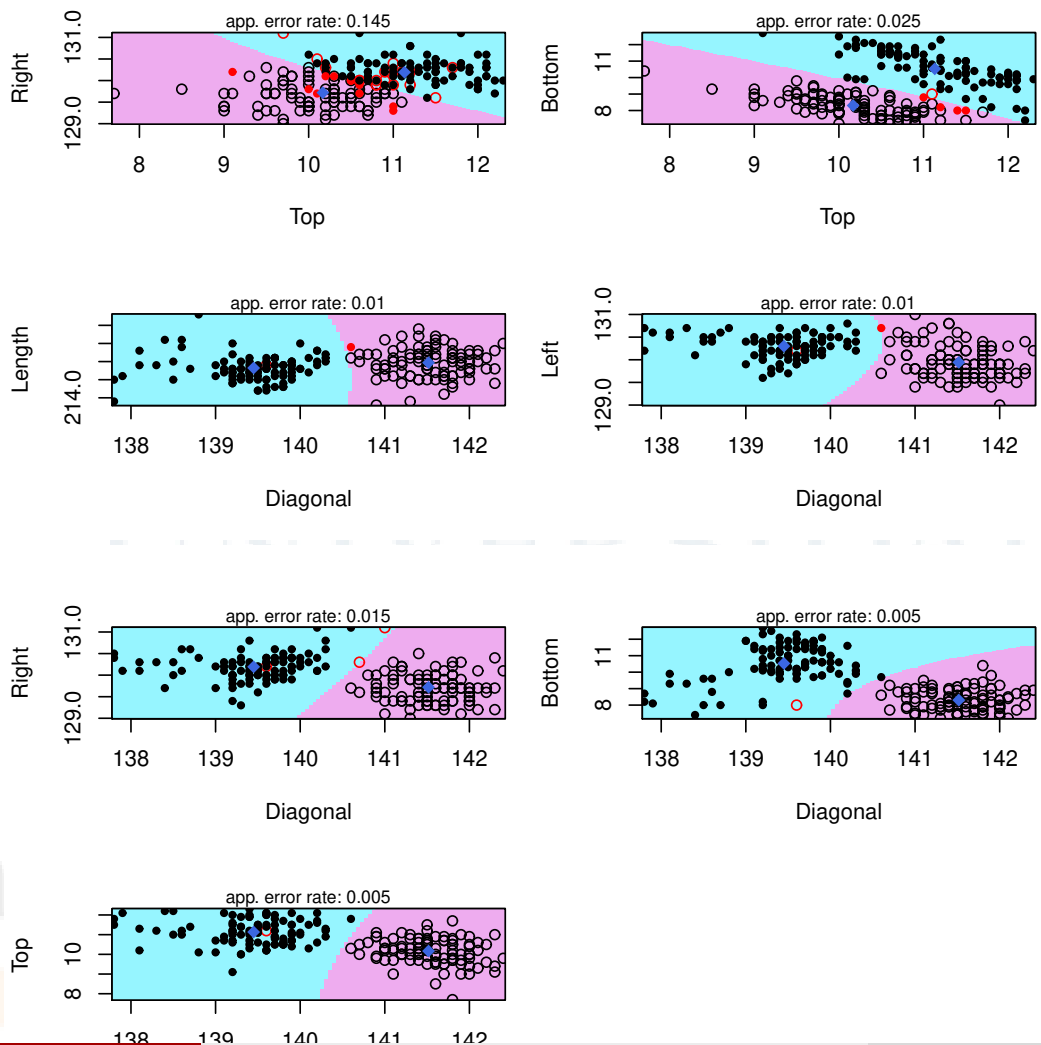
La tasa de error estimada en este caso es: 1.4%. Le parece razonable?

Ejemplo Billetes

Clasificación de los billetes, los puntos negros indican los que son apócrifos y los círculos vacíos los genuinos.



Ejemplo Billetes



Alternativas robustas

Cuando el supuesto de normalidad no se sostiene

Como ya hemos visto en secciones anteriores, la aplicación de la función discriminante requiere del cumplimiento del supuesto de normalidad multivariada.

Sin embargo, **este supuesto generalmente no se cumple** y, en algunos casos, aún cumpliéndose, la función discriminante es afectada por la presencia de observaciones atípicas o *outliers*.

Cuando el supuesto de normalidad se sostiene pero se aprecia la presencia de *outliers* es recomendable recurrir a la versión robusta del análisis discriminante cuadrático.

Sin embargo, cuando el supuesto de normalidad no puede sostenerse, el modelo robusto ya no resulta adecuado y se debe recurrir a otras alternativas de clasificación.

Análisis Discriminante Robusto

Se estima en forma robusta el centro y la escala

Se define la **distancia de Mahalanobis robusta** como:

$$RD_{ij} = \sqrt{(x_{ij} - \hat{\mu}_{j,ROB})^t \hat{\Sigma}^{-1}_{j,ROB}(x_{ij} - \hat{\mu}_{j,ROB})},$$

donde $\hat{\mu}_{j,ROB}$ y $\hat{\Sigma}_{j,ROB}$ son respectivamente los estimadores robustos de posición y de dispersión del j-ésimo grupo. Pueden obtenerse mediante MCD (*Minimun Covariance Determinant*) o MVE (*Minimum Volume Ellipsoid*).

Se propone el estimador **FAST MCD** para garantizar la eficiencia computacional.

Si l denota la cantidad de poblaciones, π_k la probabilidad de pertenecer k-ésima población P_k; la **regla de discriminante cuadrático robusto** (RQDA), utilizando MCD o FAST MCD, que permite clasificar a un nuevo individuo X, indica que:

$$X \in P_k \qquad \text{si} \qquad \widehat{d}_k^{RD}(X) < \widehat{d}_j^{RD}(X) \qquad \forall j = 1, \cdots, l, j \neq k;$$

donde $\widehat{d}_j^{RD}(X)$ es la logverosimilitud (logaritmo de la función de probabilidad, pensada como función de los parámetros) del vector de observaciones X a la j-ésima población:

$$\widehat{d}_j^{RD}(X) = -\frac{1}{2}\ln|\hat{\Sigma}_{j,MCD}| - \frac{1}{2}(X - \hat{\mu}_{j,MCD})^t \widehat{\Sigma}_{j,MCD}^{-1}(X - \widehat{\mu}_{j,MCD}) + \ln(\widehat{p}_j^R),$$

estimación robusta de la probabilidad de pertenencia a la población j-ésima, calculada excluyendo las observaciones clasificadas como atípicas, $\widehat{p}_j^R = \frac{\widetilde{n}_j}{\widetilde{n}}$ con $\widetilde{n}_j$ la cantidad de *non-outliers* en el grupo j y $\widetilde{n} = \sum_{j}^{l} \widetilde{n}_j$.

Estimaciones Robustas

Matriz de Covarianza Amalgamada

En el caso lineal, es suficiente estimar la varianza común amalgamada, para lo cual se han propuesto los siguientes tres enfoques:

- Ponderar las matrices de covarianza robustas de cada grupo.
- Ponderar las observaciones.
- Basarse en un algoritmo para estimar el determinante común.

Ejemplo Billetes

Matriz de Confusión del Discriminante Robusto

Retomamos el ejemplo de los billetes, que clasifica billetes en apócrifos o genuinos, para aplicar la alternativa robusta. En la Tabla se muestra el resultado de este análisis robusto.

	Clase predicha	
Clase real	**Apócrifo**	**Genuino**
Apócrifo	100	0
Genuino	1	99

Nota: el error de clasificación del discriminante cuadrático en este caso resulta similar a la propuesta robusta.

Organización

SVM Máquinas de Soporte Vectorial

Origen: Vapnik en los '90

Desde el inicio, sus sólidos fundamentos teóricos han hecho que fueran aceptadas. La teoría de las SVM es una nueva técnica de clasificación y ha sido aplicada a múltiples disciplinas en los últimos años. Si bien originalmente fueron diseñadas para resolver problemas de clasificación binaria, en la actualidad se aplican para resolver problemas más complejos como los de regresión, agrupamiento y multiclasificación.
Entre los campos de aplicación más difundidos podemos mencionar los siguientes:

- visión artificial,
- reconocimiento de caracteres,
- clasificación de proteínas,
- procesamiento de lenguaje natural,

SVM

Están dentro de los Clasificadores Lineales

Dado que una SVM construye un hiperplano o conjunto de hiperplanos en el espacio original cuando los conjuntos son linealmente separables o bien en el espacio transformado, denominado espacio de características, cuando los conjuntos no son linealmente separables.

En muchas de estas aplicaciones, las SVM ha probado un desempeño superior al de las máquinas de aprendizaje tradicional como las redes neuronales, y se han convertido en herramientas poderosas para dar solución a los problemas de clasificación.

Vectores de Soporte

La definición de los vectores de soporte permite formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o incuso ningún conocimiento de los datos fuera de esta frontera.

Kernel o Núcleo

Objetivo del mapeo

Los datos son mapeados por medio de alguna transformación que denominaremos **kernel** o **núcleo**, a un **espacio de características**, que es un espacio de dimensión mayor que el original, en el cual se logra una mejor separación entre las clases, por ejemplo una separación lineal.

Esta frontera en el espacio de características se corresponde con una curva o varias curvas en el espacio original, constituyendo una frontera que logra separar los datos en las distintas clases.

Las SVM surgieron originalmente como clasificadores para dos clases. Sin embargo, es posible modificar la formulación del algoritmo para que sea posible aplicarlo para realizar una clasificación multiclase.

SVM

Queremos encontrar una función lineal que separe los objetos

según su clase ubicados en un espacio bidimensional, como por ejemplo, las especies de las avispas.
Una opción posible para solucionar este problema es mapear el espacio de entrada en un espacio de dimensión mayor dentro del cual es más sencillo buscar el hiperplano óptimo.

Cada punto de entrenamiento $x \in \mathbb{R}^n$ pertenece a una de dos clases, que podrían ser etiquetadas como 1 o -1. Sea $z = \phi(x)$, notación correspondiente al mapeo en el espacio de características que llamaremos Z. Buscamos un hiperplano de la forma

$$wz + b = 0,$$

donde w es un vector en $\mathbb{R}^n$ y $b \in \mathbb{R}$, tal que separe los elementos en las dos clases definidas.

SVM

El par (w, b) determina el hiperplano

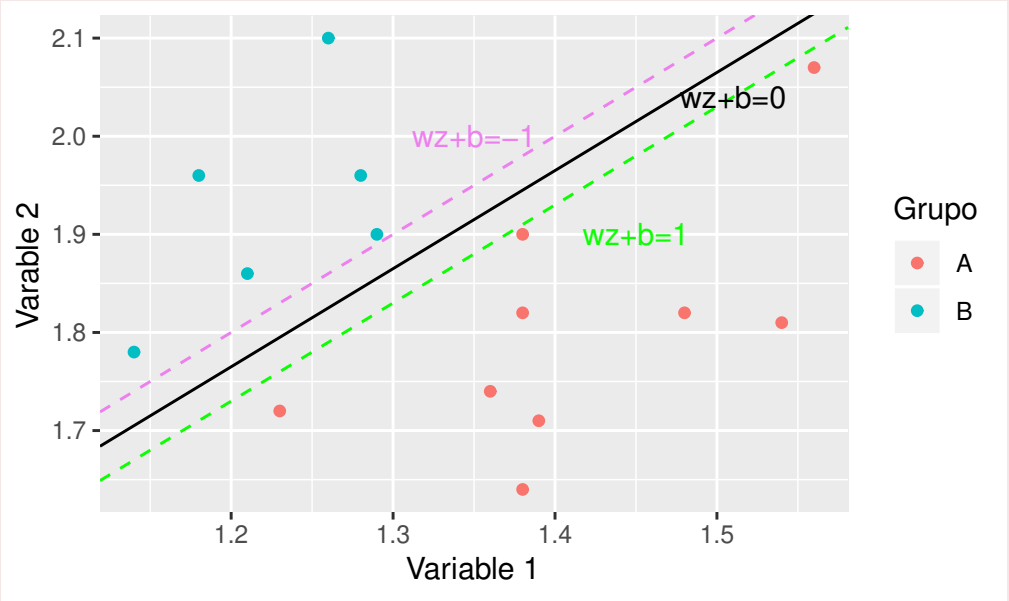
y dado un punto x_i, nos permitirá asignarlo a una de las dos clases definidas. Sea (x_i, y_i) con $x_i \in \mathbb{R}^n$ y $y_i \in \{-1, 1\}$. La función de separación tiene la siguiente forma

$$f(z_i) = y_i(w^t z_i + b) \geq 0.$$

Si $sg(w^t z_i + b) = 1$ la observación x_i se asigna al grupo A ($y_i = 1$), caso contrario se asigna al grupo B ($y_i = -1$).

Hasta acá, el planteo es similar al realizado para el del análisis discriminante lineal.

Otra manera de pensar en la separación de los conjuntos es buscar el hiperplano que maximice el margen m entre los dos conjuntos. Este problema suele resolverse con el método de multiplicadores de Lagrange.



La idea entonces es extender la capacidad de discriminar de esta nueva metodología, a conjuntos de observaciones que no estén separados linealmente, generalizando el criterio establecido.

SVM: Maximización del margen

Hay Solución única para el Problema

Desde un punto de vista algorítmico, el problema del margen geométrico se reduce a una optimización cuadrática con restricciones lineales que puede ser resuelto con programación cuadrática o mult. de Lagrange.
La propiedad de convexidad garantiza la unicidad de la solución. Los vectores de soporte permiten formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o incluso ningún conocimiento de los datos fuera de esta frontera.

Un conjunto S se dice **linealmente separable** cuando existe

un par $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ tal que se satisface el sistema de inecuaciones

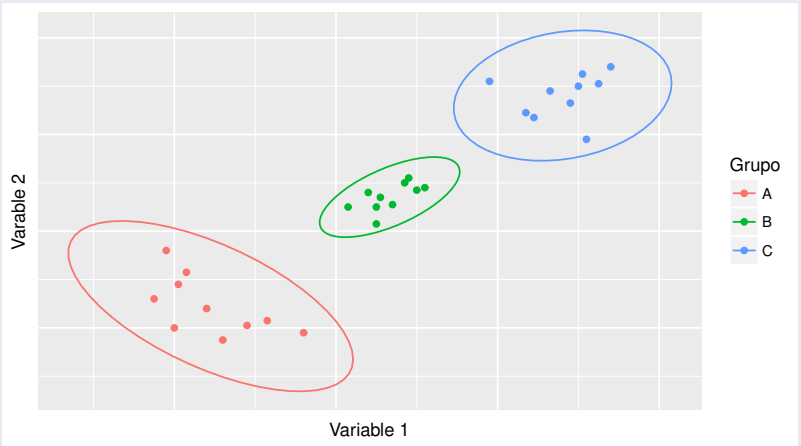
$$\begin{cases} wz_i + b \geq 1 & \text{si } y_i = 1, \\ wz_i + b \leq -1 & \text{si } y_i = -1, \end{cases}$$

Conjuntos Linealmente Separables

Envolventes Convexas

Cuando se trata de un conjunto linealmente separable, una estrategia usual para encontrar w consiste en los siguientes pasos:

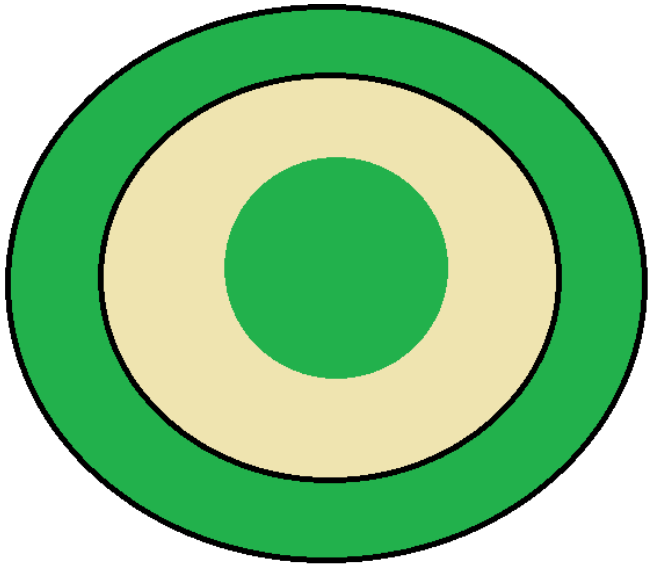
- Encontrar las envolventes convexas para los puntos de cada clase (ver Figura).
- Buscar los dos puntos más cercanos de cada envolvente.
- Encontrar el plano w que biseca a la recta que une ambos puntos.



SVM equivalencia entre enfoques

Si se realiza un planteo matemático de ambos enfoques se puede ver que en realidad son equivalentes. Sin embargo, en el uso habitual de SVM se utiliza la terminología del segundo enfoque.

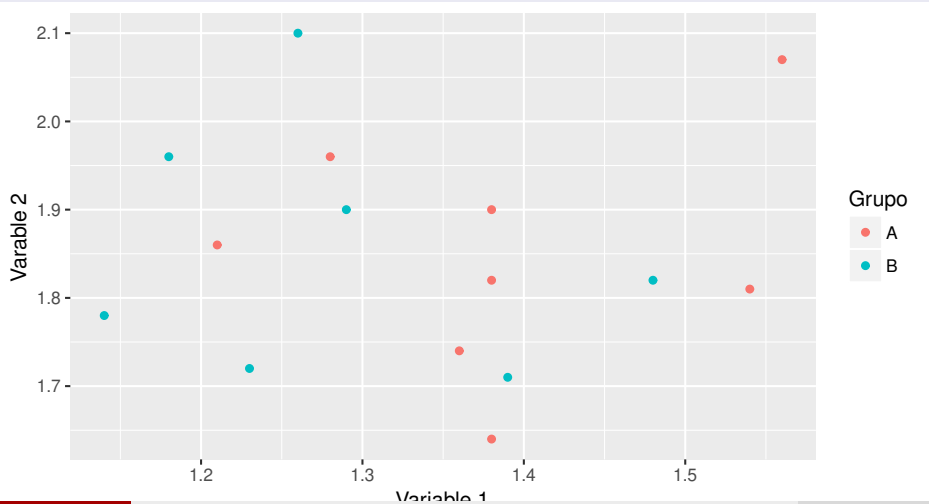
¿Cómo es el procedimiento cuando las clases no son linealmente separables?



SVM

Conjuntos no separables linealmente

Cuando no se satisface la condición para que un conjunto sea linealmente separable, se dice que el conjunto **no es linealmente separable**. Al trabajar con estos datos, no es posible separarlos linealmente. Sin embargo, pueden definirse nuevas variables, a partir de las originales de modo tal que en el nuevo espacio, los conjuntos resulten linealmente separables. Las nuevas variables podrían ser x^2, y^2, xy, entre otras.



Mapeo

Observaciones

- El proceso de pasar de un espacio de entrada con m dimensiones a otro espacio de características de $m* > m$ dimensiones, se llama **mapeo**.
- El aumento de la dimensión del problema conlleva el riesgo de sobreajuste, aunque no resulta tan grave con SVM.
- La disponibilidad inicial de contar con muchas variables conduce a un aumento exponencial de las variables en el mapeo.

SVM

Síntesis de los pasos del procedimiento

- Se mapean los puntos de entrada a un espacio de características de una dimensión mayor; por ejemplo, si los puntos de entrada están en $\mathbb{R}^2$ pueden ser mapeados a $\mathbb{R}^3$.
- Se busca en la imagen de este mapeo un hiperplano que los separe y que maximice el margen entre las clases.
- La solución del hiperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados **vectores soporte**.

Espacio de Características

Kernels y mapeo

Generalmente no se tiene ningún conocimiento sobre la función de mapeo más conveniente, que denotamos con ϕ, por ende el cálculo de la separación en el nuevo espacio parece imposible. Sin embargo, las SVM poseen una buena propiedad que no hace necesario ningún conocimiento acerca de ϕ.
Sólo es necesaria una función K que calcule el producto escalar de los puntos de entrada en el espacio de características Z; vale decir

$$K(x_i, x_j) = \phi(x_i)\phi(x_j) = z_i \cdot z_j.$$

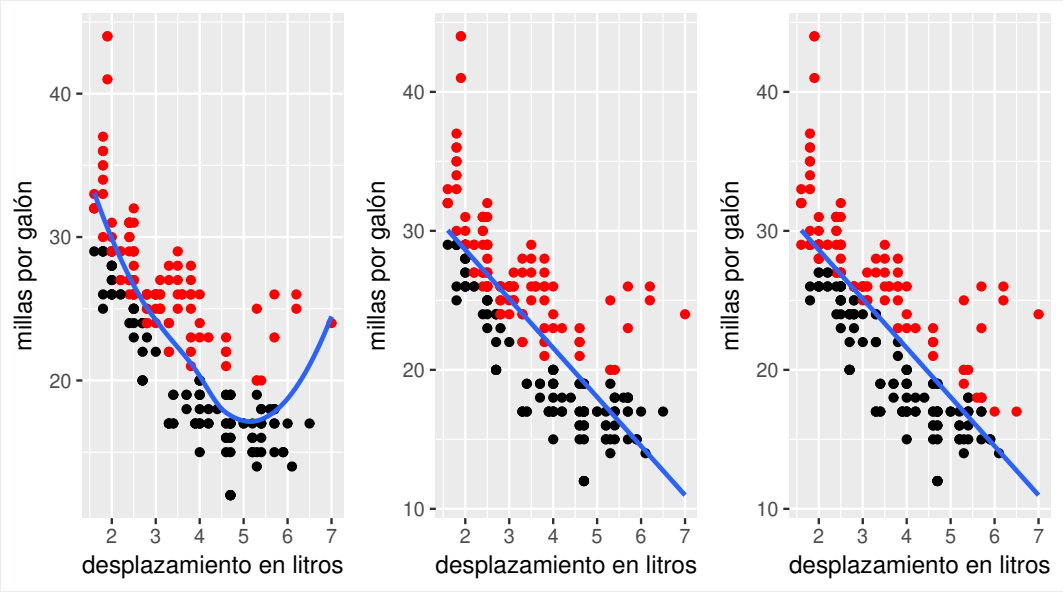
Luego, si $\phi(x)$ denota el mapeo de la variable x, los **kernels** son objetos matemáticos que satisfacen

$$K(x_i, x_i) = \phi(x_i)\phi(x_i).$$

Maximización de la distancia de separación entre grupos

Efecto del Mapeo

El efecto de la aplicación del mapeo y los *kernels* se puede apreciar en la siguiente Figura aplicada a la base de datos `mpg` de `R` relacionada con la economía de combustible según la marca de automóvil.



Núcleos o Kernels conocidos

Los *kernels* más conocidos

- **Kernel Lineal:** $K(x_i, x_j) = \langle x_i, x_j \rangle$.
- **Kernel polinomial de grado** h**:** $K(x_i, x_j) = (\langle x_i, x_j \rangle + \tau)^h$.
- **Kernel sigmoideo:** $K(x_i, x_j) = \tanh(\langle x_i, x_j \rangle + \tau)$.
- **Kernel gaussiano:** $K(x_i, x_j) = \exp(\gamma |x_i - x_j|^2)$.

Si se va a tratar con datos que no son linealmente separables, el análisis previo puede ser generalizado introduciendo algunas variables no negativas $\zeta_i \geq 0$, de tal modo que el sistema

$$\begin{cases} wz_i + b \geq 1 & \text{si } y_i = 1, \\ wz_i + b \leq -1 & \text{si } y_i = -1; \end{cases}$$

es modificado a

$$\left\{ y_i(wz_i + b) \geq 1 - \zeta_i \quad \text{para } 1 \leq i \leq n. \right.$$

Los valores de ζ_i corresponden a los puntos que no satisfacen el sistema de inecuaciones original; es decir, el de la definición de conjunto separable linealmente. De este modo, $\sum_{i=1}^{n} \zeta_i$ se convierte en una medida de bondad de la clasificación.

Solución dual

Conversión del Problema

Hemos convertido el problema de hallar el hiperplano, en el problema de minimizar la expresión

$$\min \left\{ \frac{1}{2} w.w + C \sum_{i=1}^{n} \zeta_i \right\}.$$

La constante C es un parámetro de regularización y puede ser ajustada durante la formulación de las SVM.
Este problema de optimización cuadrática puede resolverse mediante su dual, que conduce a la ecuación

$$\max W(\alpha) = \max \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j z_i z_j \right\},$$

Ventajas del Método

Dentro de las ventajas de esta técnica podemos mencionar las siguientes.

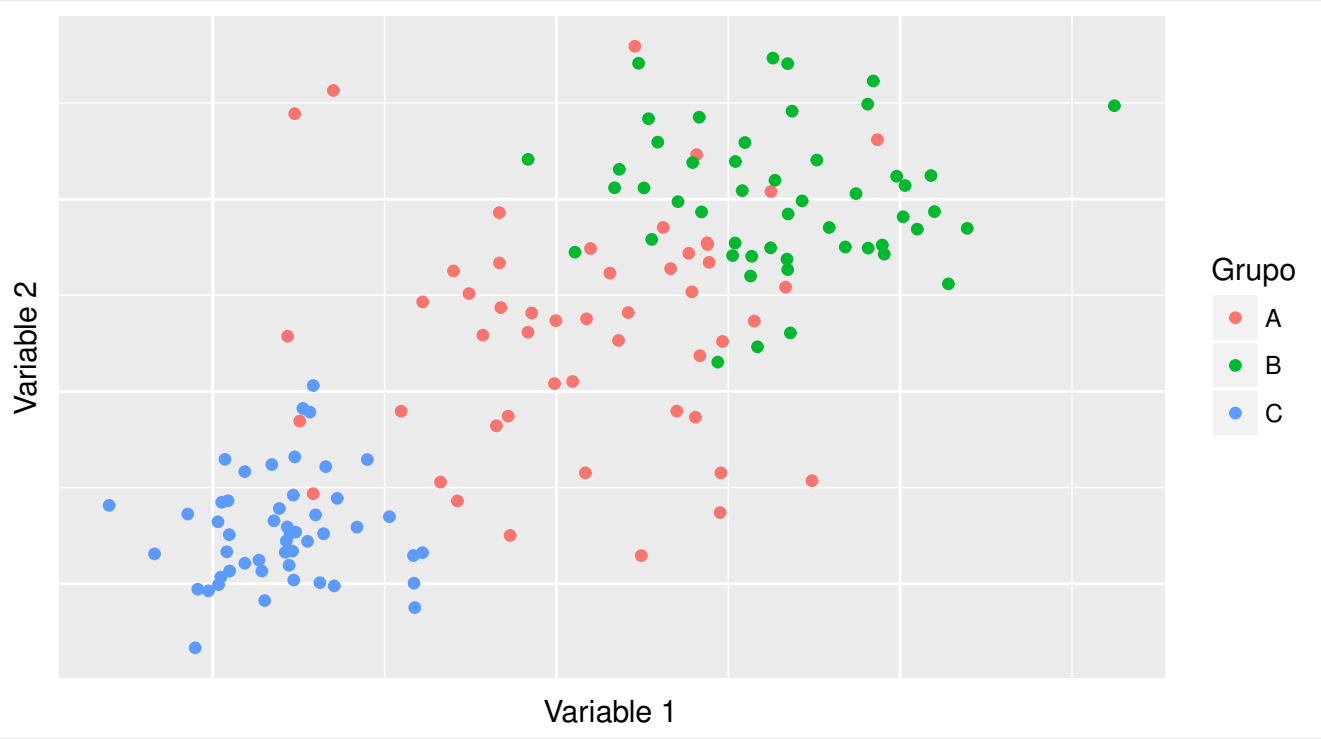
- El entrenamiento es relativamente sencillo.
- No existe un óptimo local.
- Se escala relativamente bien para datos en espacios de alta dimensión.
- El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente.
- Datos no tradicionales, como cadenas de caracteres o árboles, pueden ser ingresados como entrada de las SVM.

Por el contrario, la mayor debilidad radica en que es necesaria una buena función *kernel*; es decir, se necesitan metodologías eficientes para definir los parámetros de inicialización de las SVM.

Una observación interesante es que una lección aprendida en las SVM se

Veamos cómo utilizar una máquina de soporte vectorial en R

En la Figura podemos ver la representación de los datos simulados, donde se aprecia que si bien los grupos están bastante separados, no son linealmente separables.



La Tabla muestra la matriz de confusión del modelo

A partir de esta salida se sabe que en el grupo A, hay 34 individuos de los cuales 21 resultaron bien clasificados, hay un 38% de error. En el grupo B hay 34 individuos de los cuales 31 resultaron adecuadamente clasificados, el porcentaje de error es 9%. Finalmente, en el grupo C hay 31 individuos que fueron todos clasificados correctamente. La tasa de error global resulta de aproximadamente el 16%.

	Clasificación por modelo		
Grupo original	**A**	**B**	**C**
A	21	10	3
B	3	31	0
C	0	0	31

Visualización de la Clasificación

Bondad de la Clasificación

En la Figura se muestra la representación de la clasificación por SVM donde se pueden identificar la cantidad de datos mal clasificados de cada uno de los grupos mediante esta regla.