Análisis Discriminante

José A. Perusquía Cortés

Análisis Multivariado, Semestre 2025 - II

Motivación

- Sabiendo que un objeto viene de uno de k grupos distintos se busca:
  - Asignar el objeto utilizando p características
  - La regla de asignación sea óptima en algún sentido

- Cuatro casos a considerar:
  - La distribución es conocida (imposible en la realidad)
  - La distribución es conocida salvo algunos parámetros
  - La distribución es parcialmente conocida
  - La distribución es desconocida

Distribución conocida (2 grupos)

Formulación

- Suponemos:
  - 2 grupos con proporciones π_1 y $\pi_2 = 1 - \pi_1$ y densidades f_1 y f_2
  - Asignamos al grupo G_i si $\mathbf{x} \in R_i$ con $R_1 \cup R_2 = R$
- Se puede cometer el error de:
  - Asignar $\mathbf{x}$ a G_2 cuando $\mathbf{x} \in G_1$ (o viceversa)
  - Las probabilidades de error se definen como

$$P(2 \mid 1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x} \qquad P(1 \mid 2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$

  - La probabilidad de mis-clasificación es:

$$p = P(1 \mid 2)\pi_2 + P(2 \mid 1)\pi_1$$

Formulación

Lema

La integral $\int_{R_1} g(\mathbf{x})d\mathbf{x}$ se minimiza con respecto a R_1 cuando $R_1 = R_{01} = \{\mathbf{x} : g(\mathbf{x}) < 0\}$

Observaciones

- La región R_{01} no es única
- Los puntos frontera $B = \{\mathbf{x} : g(\mathbf{x}) = 0\}$ se pueden asignar arbitrariamente a R_{01} o a R_{02}

Criterios de asignación

a. Minimizar la probabilidad total de mis-clasificación

- La probabilidad total de error es

$$p = P(1 \mid 2)\pi_2 + p(2 \mid 1)\pi_1 = \pi_1 + \int_{R_1} \left[\pi_2 f_2(\mathbf{x}) - \pi_1 f_1(\mathbf{x})\right] d(\mathbf{x})$$

- Se minimiza en $R_{01} = \{\mathbf{x} : \pi_2 f_2(\mathbf{x}) - \pi_1 f_1(\mathbf{x}) < 0\}$

- Asignar a G_1 si

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

Criterios de asignación

b. Maximizar la función de verosimilitud

- Si π_1 es desconocida

- Asignar a G_1 si

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1$$

- Caso particular de **a.** con

$$\pi_1 = \pi_2 = \frac{1}{2}$$

Criterios de asignación

c. Minimizar el costo de mis-clasificación

- Sean $C(1 \mid 2), C(2 \mid 1)$ los costos de clasificar mal a los miembros de G_1, G_2

- El costo total esperado es:

$$C_T = C(2 \mid 1)P(2 \mid 1)\pi_1 + C(1 \mid 2)P(1 \mid 2)\pi_2$$

- C_T se minimiza cuando $C(1 \mid 2)\pi_2 f_2(\mathbf{x}) < C(2 \mid 1)\pi_1 f_1(\mathbf{x})$

- Asignamos a G_1 si

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2 C(1 \mid 2)}{\pi_1 C(2 \mid 1)}$$

Criterios de asignación

d. Maximizar la probabilidad posterior

- La probabilidad posterior de G_i dado $\mathbf{x} = \mathbf{x}_0$

$$q_i(\mathbf{x}_0) = \frac{f_i(\mathbf{x}_0)\pi_i}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}$$

- Asignamos a G_1 si

$$q_1(\mathbf{x}) > q_2(\mathbf{x})$$

Criterios de asignación

e. Minimax

- Si $\pi_1 << \pi_2$ asignar un objeto para minimizar la máxima probabilidad individual de mis-clasificación

- Para $\alpha \in [0,1]$ se tiene que $\max\{P(1 \mid 2), P(2 \mid 1)\} \geq (1-\alpha)P(2 \mid 1) + \alpha P(1 \mid 2)$

- Se minimiza cuando

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\alpha}{1-\alpha} = c$$

- A c se puede elegir de tal forma que en R_{01} se cumpla $P_0(1 \mid 2) = P_0(2 \mid 1)$

Ejemplo 1: Discriminante lineal (lda)

Sea $f_i = N_p(\mu_i, \Sigma)$

- Asignamos a G_1 si

$$D(\mathbf{x}) = (\mu_1 - \mu_2)^T \Sigma^{-1} \left[\mathbf{x} - \frac{1}{2}(\mu_1 + \mu_2) \right] > \log \left(\frac{\pi_2}{\pi_1} \right)$$

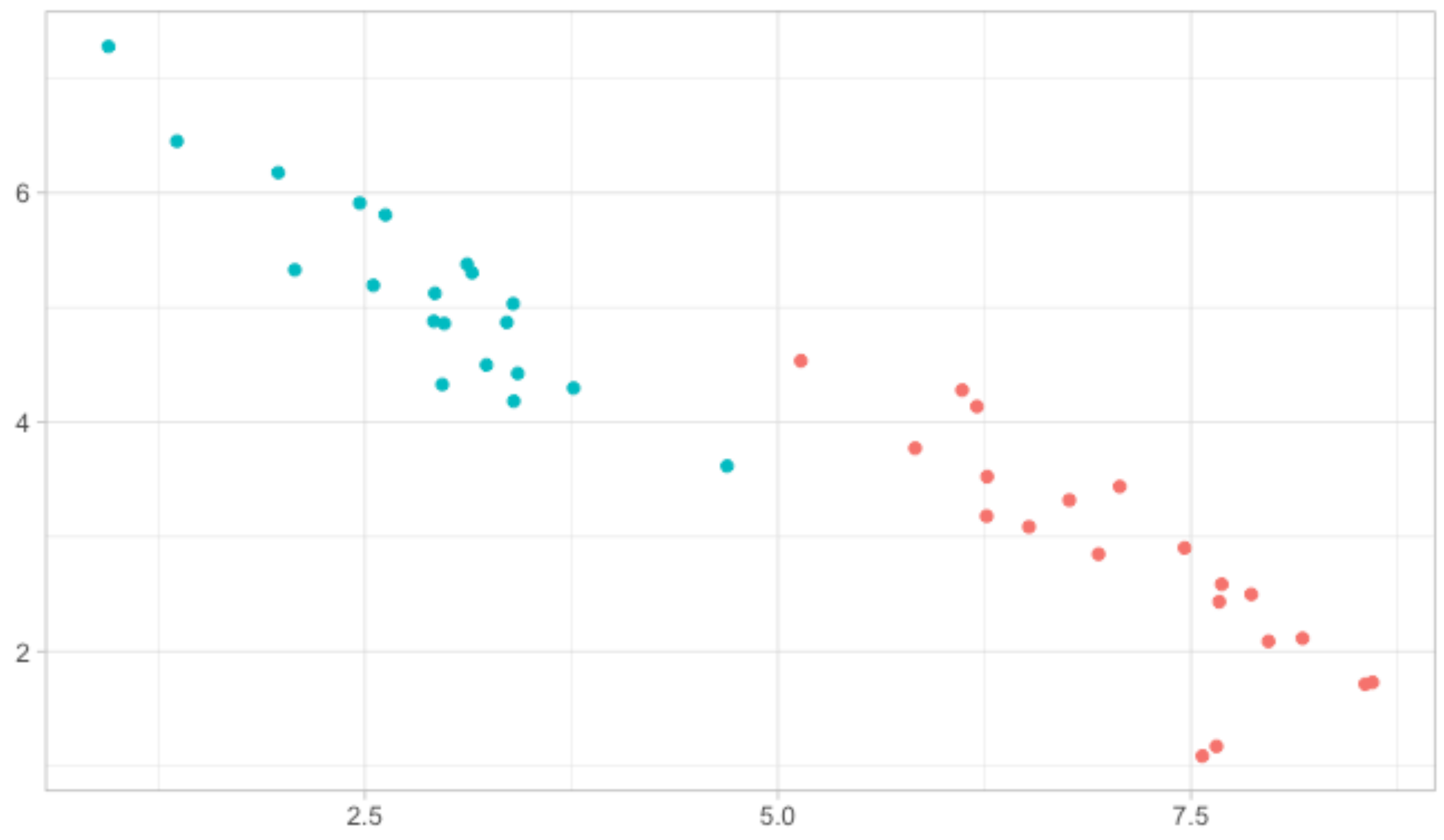
- Las probabilidades de mis-clasificación son:

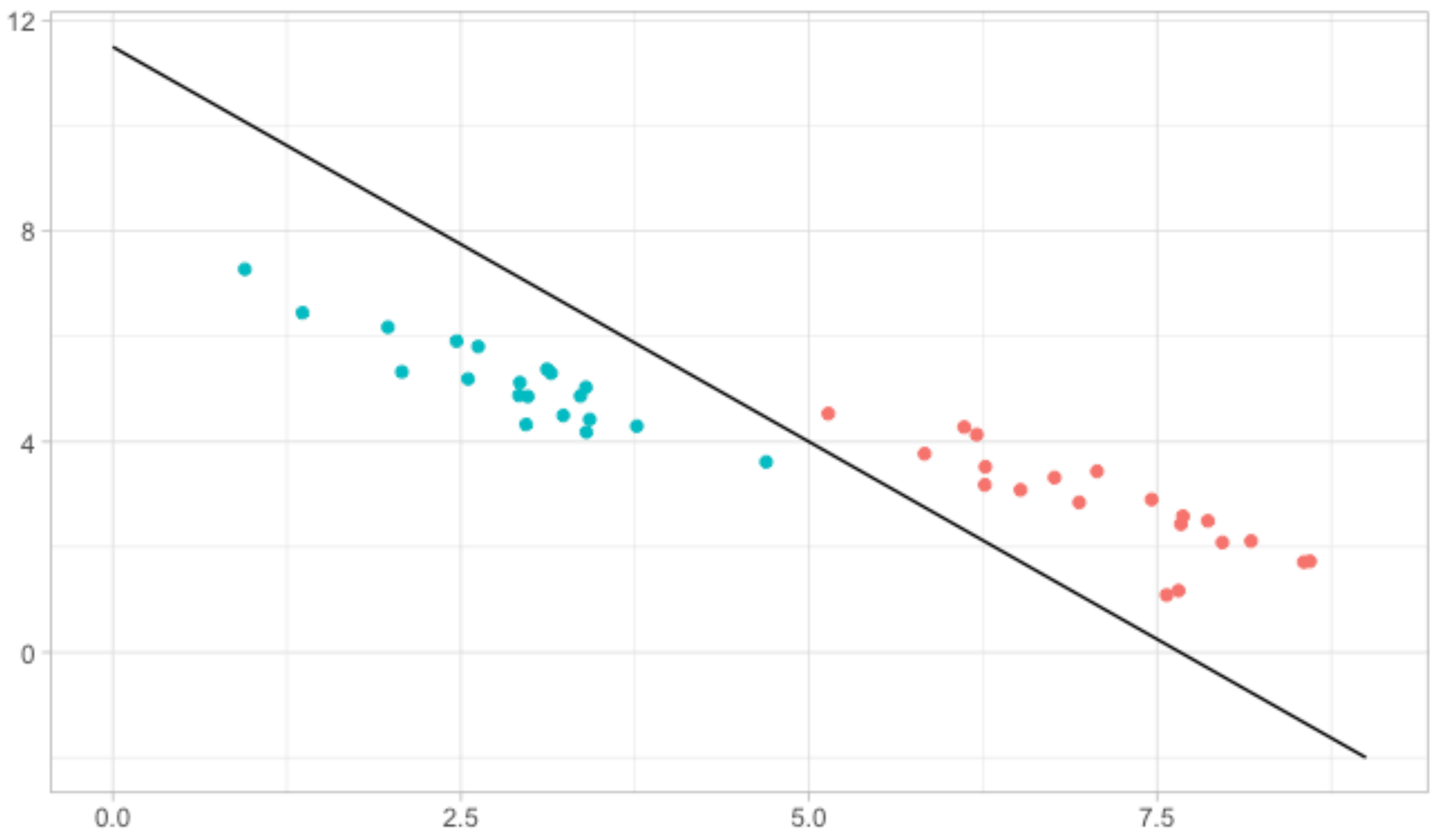
$$P(2 \mid 1) = \Phi \left(\frac{\log \left[\frac{\pi_2}{\pi_1} \right] - \frac{1}{2}\Delta^2}{\Delta} \right) \qquad P(1 \mid 2) = \Phi \left(\frac{\log \left[\frac{\pi_1}{\pi_2} \right] - \frac{1}{2}\Delta^2}{\Delta} \right)$$

- El caso $\pi_1 = \pi_2$ fue estudiado por Fisher (1936)

- En R en la librería MASS existe función lda()

Ejemplo 1: Discriminante lineal (lda)





Ejemplo 2: Discriminante cuadrático (qda)

Sea $f_i = N_p(\mu_i, \Sigma_i)$

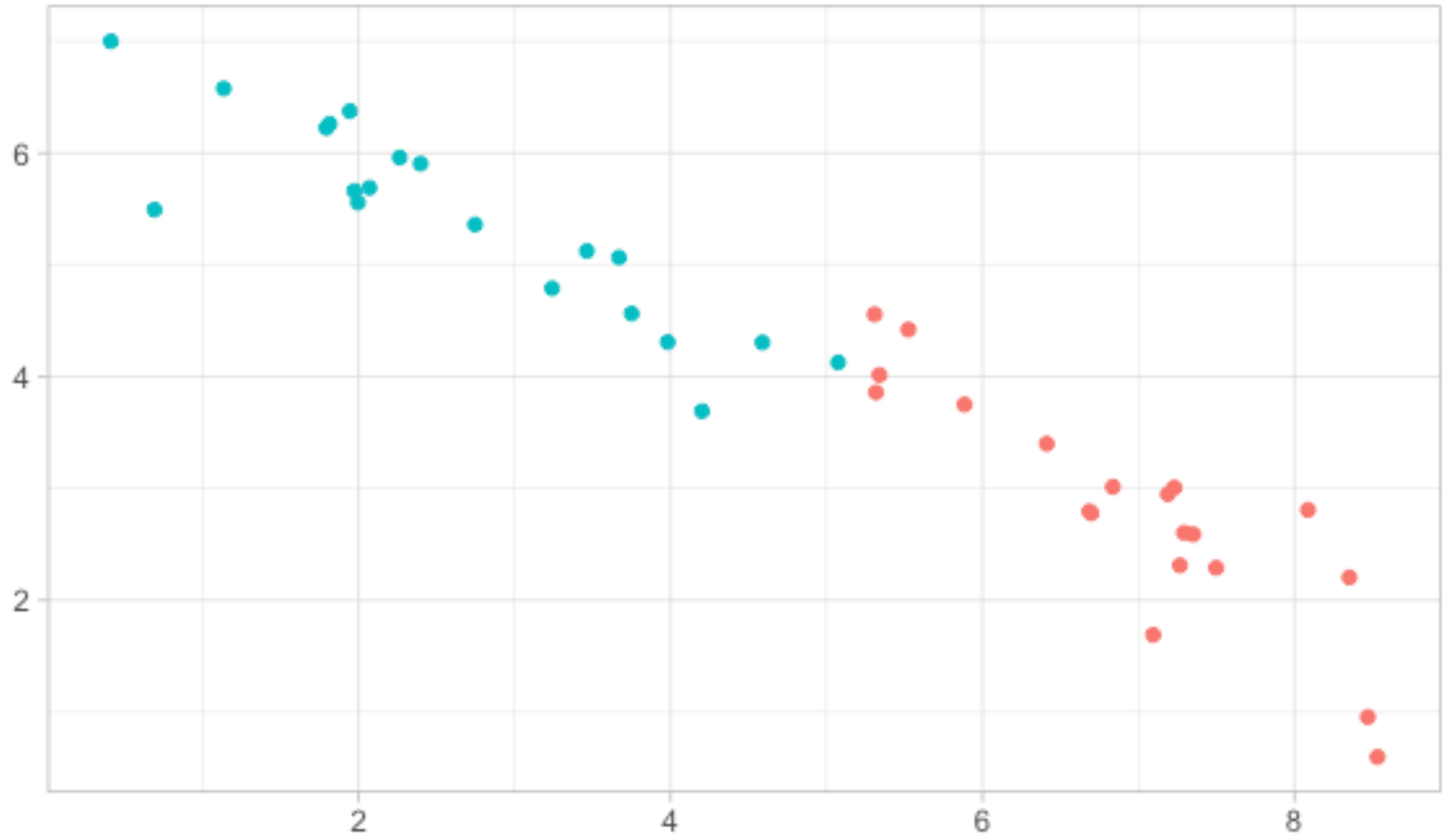
- Asignamos a G_1 si

$$Q(\mathbf{x}) = c_0 - \frac{1}{2}\left[\mathbf{x}^T(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x} - 2\mathbf{x}^T(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)\right] > \log\left(\frac{\pi_2}{\pi_1}\right)$$

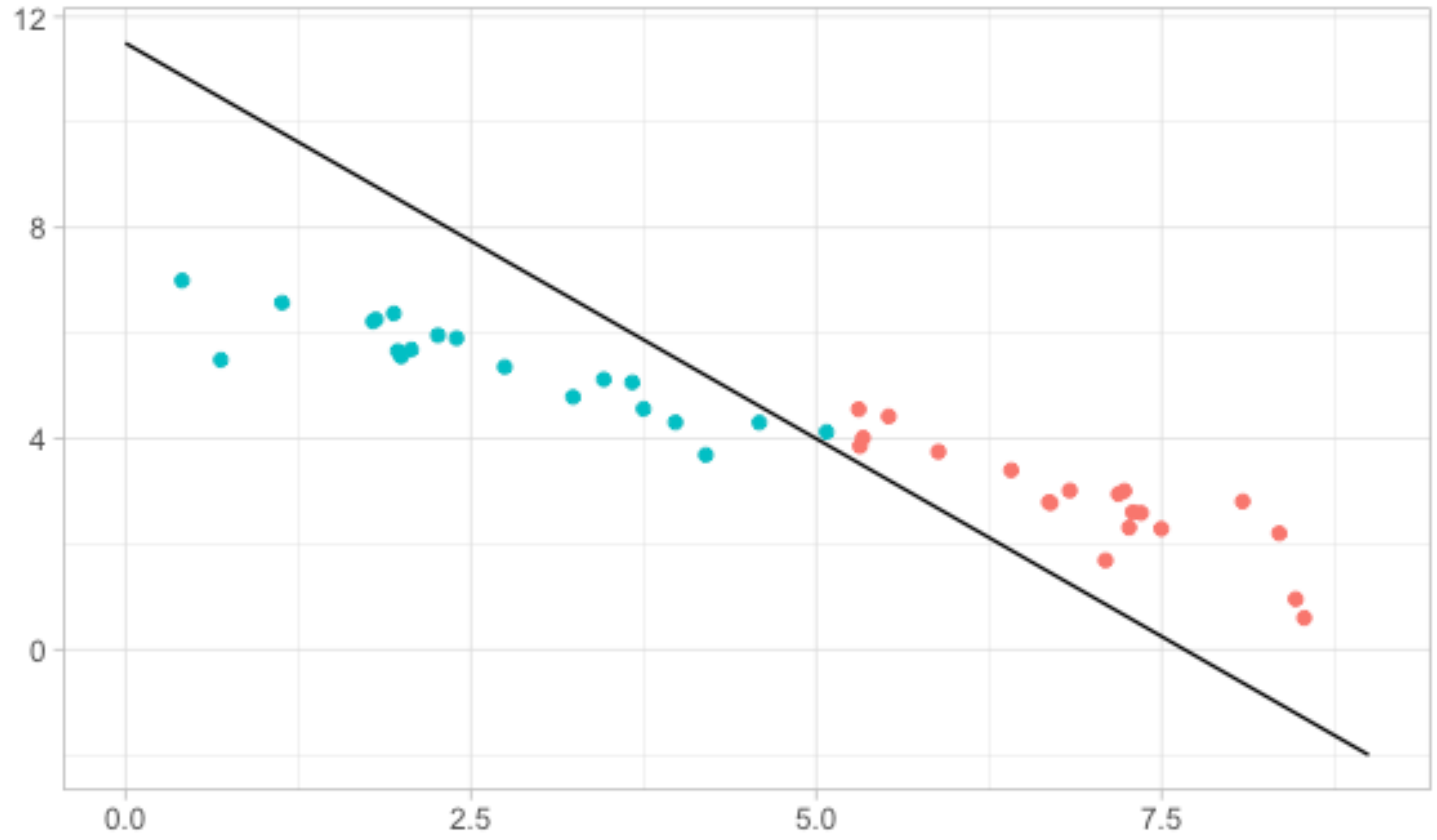
- La función ahora es cuadrática en lugar de lineal como $D(\mathbf{x})$

- En R en la librería MASS existe función qda()

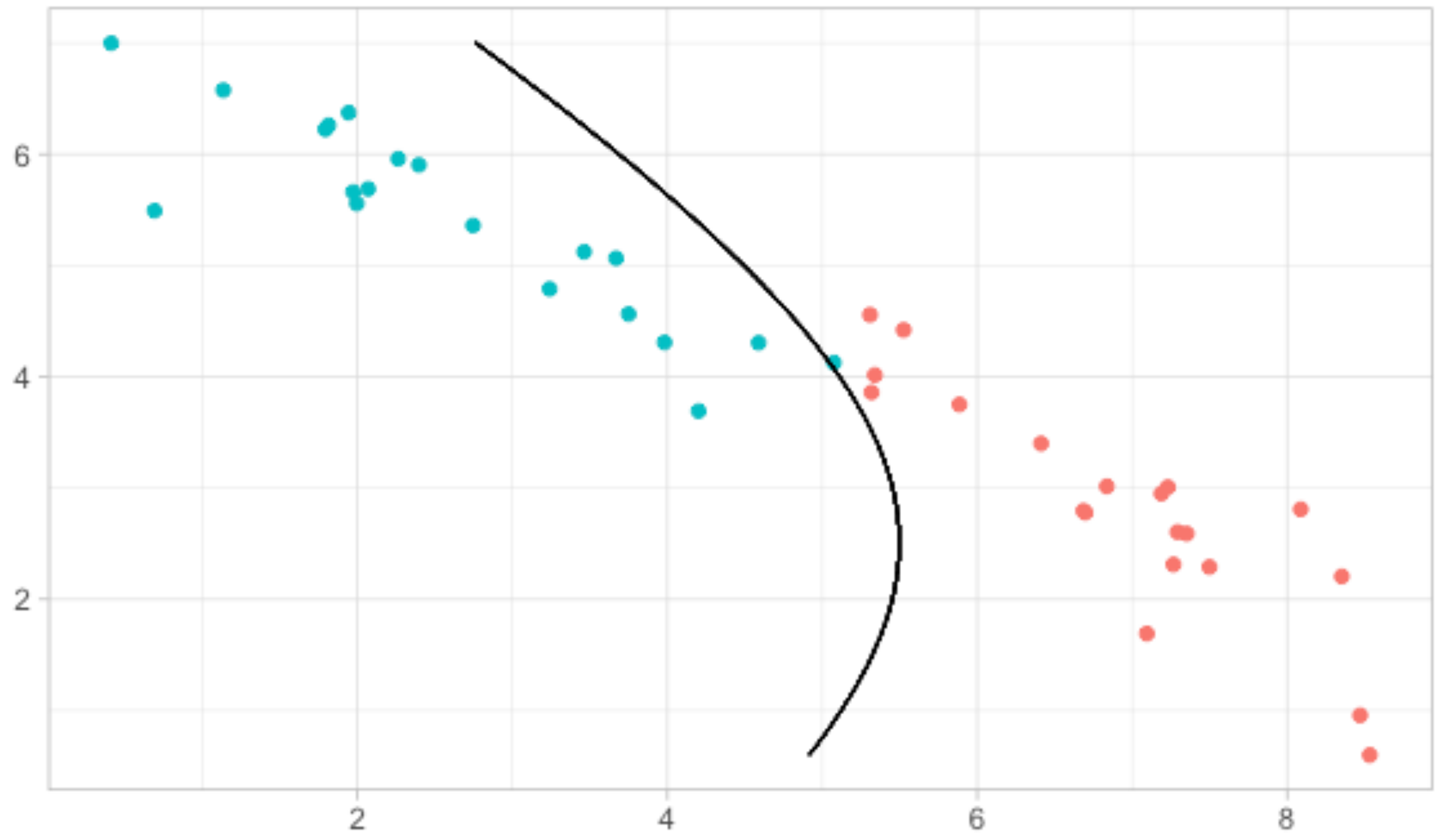
Ejemplo 2: Discriminante cuadrático (qda)



Ejemplo 2: Discriminante cuadrático (qda)



lda



qda

Parámetros Desconocidos

Metodología

- Sea $f_i(\mathbf{x} \mid \theta_i)$ la densidad del grupo G_i con parámetros desconocidos θ_i y una muestra de cada grupo

- Obtener $\hat{\theta}_i$ (e.g. máximos verosímiles)

- La región óptima del grupo G_1 es

$$\hat{R}_{01} = \left\{ \mathbf{x} : \frac{f_1\left(\mathbf{x} \mid \hat{\theta}_1\right)}{f_2\left(\mathbf{x} \mid \hat{\theta}_2\right)} > c \right\}$$

- Para $n >> 1$ se tiene que $\hat{R}_{01} \approx R_{01}$

Errores

Como en el caso de θ_i conocidos se deben considerar los siguientes errores de clasificación

- Error óptimo

$$e_{opt} = \pi_1 e_{1,opt} + \pi_2 e_{2,opt}, \qquad e_{i,opt} = \int_{R_{0j}} f_i(\mathbf{x} \mid \theta_i) d\mathbf{x}$$

- Error actual

$$e_{act} = \pi_1 e_{1,act} + \pi_2 e_{2,act}, \qquad e_{i,act} = \int_{\hat{R}_{0j}} f_i(\mathbf{x} \mid \theta_i) d\mathbf{x}$$

Estimación de errores

- Usando $\hat{\theta}_i$

$$\hat{e}_{i,act} = \int_{\hat{R}_{0j}} f_i(\mathbf{x} \mid \hat{\theta}_i) d\mathbf{x}$$

- Errores aparentes usando los elementos mal clasificados m_i

$$e_{i,app} = \frac{m_i}{n_i}$$

- Validación cruzada

$$e_{i,val} = \frac{a_i}{n_i}$$

- Métodos bootstrap

Discriminación Logística

Metodología

- El modelo logístico asume que

$$\log\left[\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right] = \alpha + \beta^T\mathbf{x}$$

- Asignamos a G_1 si $\alpha + \beta^T\mathbf{x} > \log(\pi_2/\pi_1)$

- Las probabilidades posteriores son

$$q_1(\mathbf{x}) = \frac{\exp[\alpha + \log(\pi_1/\pi_2) + \beta^T\mathbf{x}]}{\exp[\alpha + \log(\pi_1/\pi_2) + \beta^T\mathbf{x}] + 1} \qquad q_2 = 1 - q_1$$

Metodología

- Estimar menos parámetros

- No necesitamos especificar las densidades de cada grupo

- Muchas familias satisfacen la relación lineal

- Particularmente útil para diagnósticos

Distribuciones Desconocidas

Método del kernel

- Estimar $f(\mathbf{x})$ a partir de los datos como

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} K(\mathbf{x} \mid \mathbf{x}_j, \lambda)$$

- Donde $K(\mathbf{y} \mid \mathbf{z}, \lambda)$ es un kernel o una densidad con moda $\mathbf{z}$ y parámetro de suavidad λ

- Asignar a G_1 si

$$\frac{\hat{f}_1(\mathbf{x})}{\hat{f}_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

Tipo de datos

- Para datos continuos

$$K_1(\mathbf{y} \mid \mathbf{z}, \lambda) = (2\pi\lambda^2)^{-\frac{p}{2}} \mid \mathbf{S} \mid^{-\frac{1}{2}} \exp\left[\frac{1}{2\lambda^2}(\mathbf{y}-\mathbf{z})^T\mathbf{S}^{-1}(\mathbf{y}-\mathbf{z})\right]$$

- Para datos binarios se sugiere (Aitchison y Aitken, 1976)

$$K_2(\mathbf{y} \mid \mathbf{z}, \lambda) = \lambda^{p-D(\mathbf{y},\mathbf{z})}(1-\lambda)^{D(\mathbf{y},\mathbf{z})} \qquad \frac{1}{2} \leq \lambda \leq 1 \qquad D(\mathbf{y}, \mathbf{z}) = \mid\mid \mathbf{y} - \mathbf{z} \mid\mid^2$$

- Para mezclas de continuos y discretos

$$K_3(\mathbf{y} \mid \mathbf{z}, \lambda) = K_1(\mathbf{y} \mid \mathbf{z}, \lambda)K_2(\mathbf{y} \mid \mathbf{z}, \lambda)$$

Otros métodos

- Vecino más cercano

- Particiones

- Distancias

- Rangos

Más de 2 grupos

Formulación

- Suponemos: k grupos con proporciones π_i con densidades f_i

- Queremos encontrar partición $\mathscr{R} = \{R_1, R_2, \ldots, R_k\}$ y asignar a G_i si $\mathbf{x} \in R_i$

- La probabilidad de asignar a G_j cuando viene de G_i

$$P(j \mid i) = \int_{R_j} f_i(\mathbf{x}) d\mathbf{x}$$

- La probabilidad de clasificar mal a un elemento de G_i

$$P(i) = \sum_{j \neq i}^{k} P(j \mid i) = 1 - P(i \mid i)$$

Formulación

- La probabilidad total de mis-clasificación

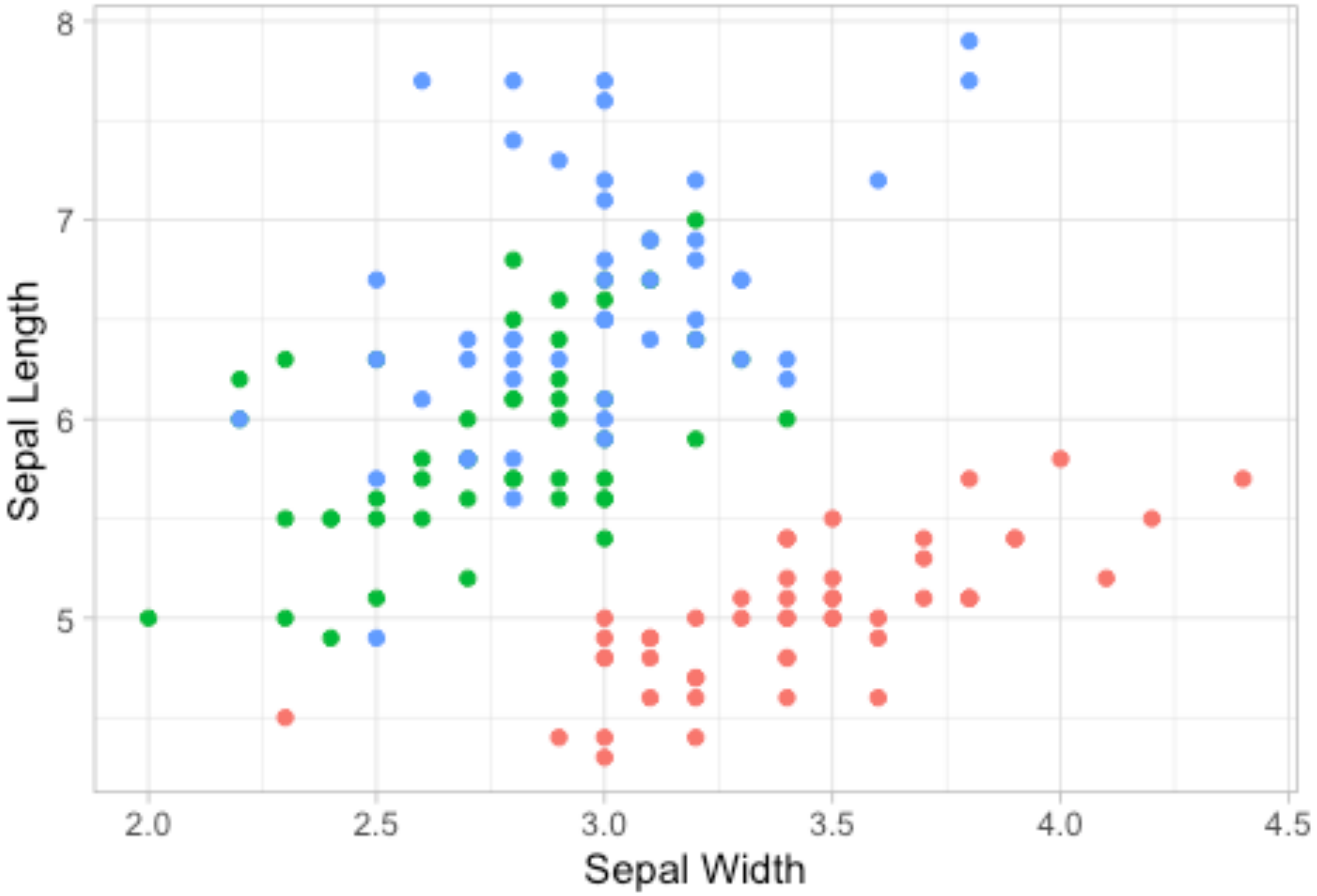
$$P(\mathcal{R}, \mathbf{f}) = \sum_{i=1}^{k} \pi P(i) = 1 - \sum_{i=1}^{k} \pi_i P(i \mid i)$$

- Usamos el enfoque bayesiano, i.e., asignar al grupo con mayor probabilidad posterior

$$q_i(\mathbf{x}) = \frac{\pi_i f_i(\mathbf{x})}{\sum_{j=1}^{k} \pi_j f_j(\mathbf{x})}$$

- Los puntos frontera se asignan de forma arbitraria

Ejemplo iris



Ejemplo iris

- Los vectores de medias

$$\hat{\mu}_{set} = \begin{pmatrix} 3.428 & , & 5.006 \end{pmatrix} \qquad \hat{\mu}_{ver} = \begin{pmatrix} 2.770 & , & 5.936 \end{pmatrix} \qquad \hat{\mu}_{vir} = \begin{pmatrix} 2.974 & , & 6.588 \end{pmatrix}$$

- Las matrices de covarianzas

$$\hat{S}_{set} = \begin{pmatrix} 0.14360.0992 \\ 0.09920.1242 \end{pmatrix} \qquad \hat{S}_{ver} = \begin{pmatrix} 0.09840.0851 \\ 0.08510.2664 \end{pmatrix} \qquad \hat{S}_{vir} = \begin{pmatrix} 0.10400.0937 \\ 0.09370.4043 \end{pmatrix}$$

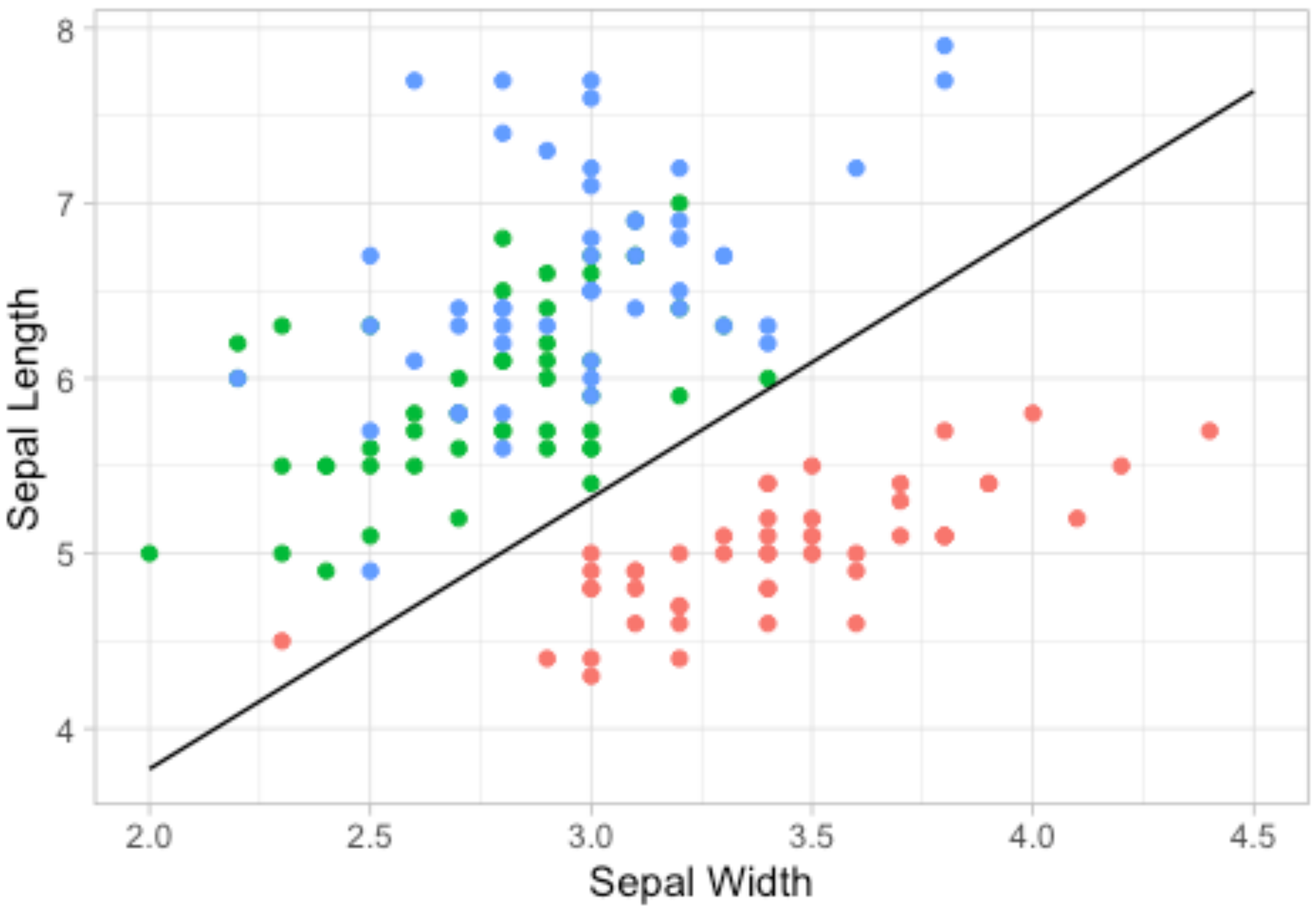
- Asumiendo que son la misma (se debe probar) tomamos la varianza compartida

$$\hat{S}_p = \frac{49}{147}\left(\hat{S}_{set} + \hat{S}_{ver} + \hat{S}_{vir}\right) = \begin{pmatrix} 0.1153 & 0.0927 \\ 0.0927 & 0.2650 \end{pmatrix}$$

Ejemplo iris

- Asignamos a setosa en lugar de versicolor si

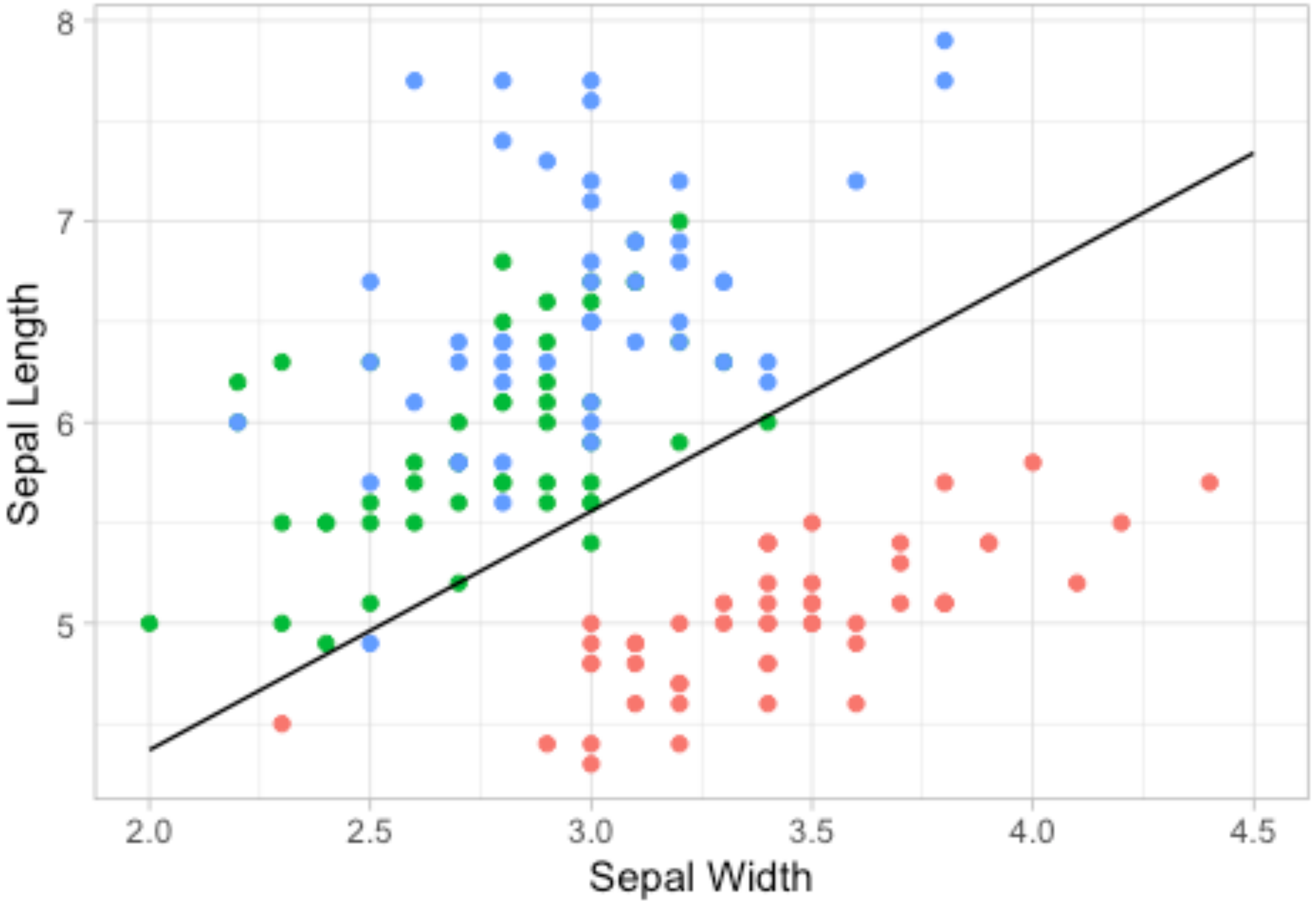
$$D(\mathbf{x}) = (\hat{\mu}_{set} - \hat{\mu}_{ver})^T \hat{S}_p^{-1} \left[\mathbf{x} - \frac{1}{2}(\hat{\mu}_{set} + \hat{\mu}_{ver}) \right] = 5.1528 - 7.6574 x_1 + 11.8557 x_2 > 0$$



Ejemplo iris

- Asignamos a setosa en lugar de virginica si

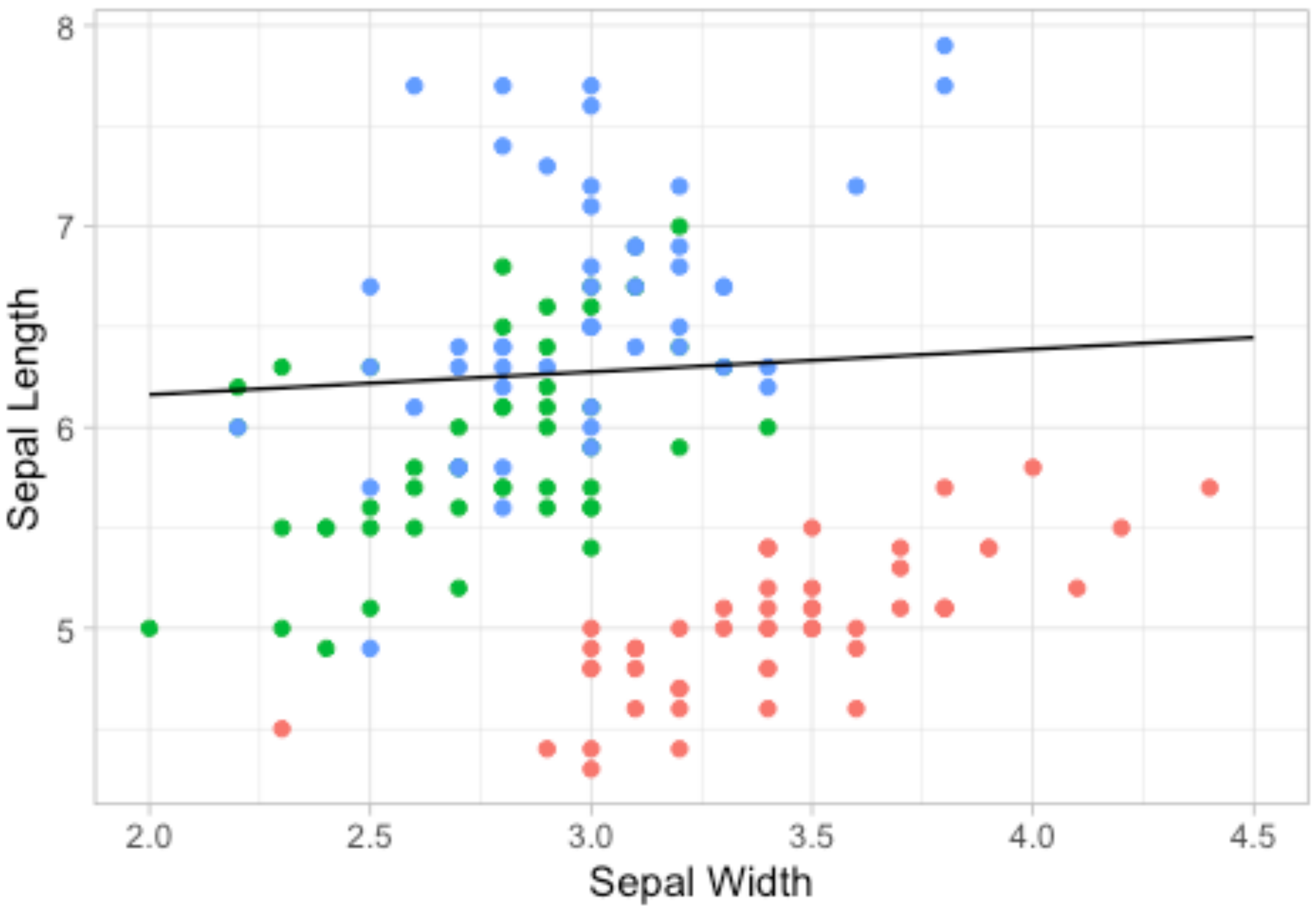
$$D(\mathbf{x}) = (\hat{\mu}_{set} - \hat{\mu}_{vir})^T \hat{S}_p^{-1} \left[\mathbf{x} - \frac{1}{2}(\hat{\mu}_{set} + \hat{\mu}_{vir})\right] = 20.3612 - 10.2914x_1 + 12.1465x_2 > 0$$



Ejemplo iris

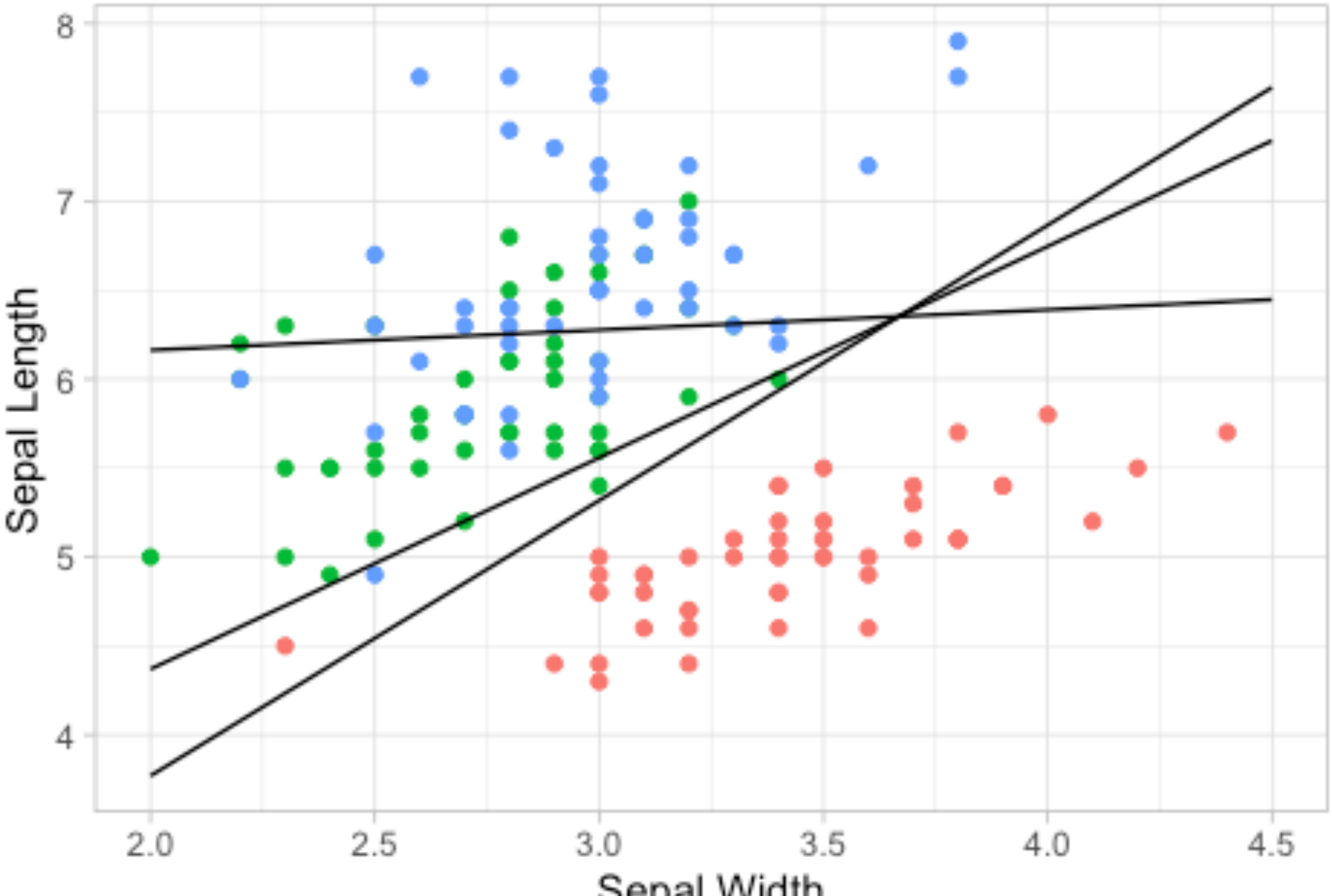
- Asignamos a versicolor en lugar de virginica si

$$D(\mathbf{x}) = (\hat{\mu}_{ver} - \hat{\mu}_{vir})^T \hat{S}_p^{-1} \left[\mathbf{x} - \frac{1}{2}(\hat{\mu}_{ver} + \hat{\mu}_{vir})\right] = 15.2084 - 2.5621x_1 + 0.2908x_2 > 0$$



Ejemplo iris

- Todas las regiones



Ejemplo iris

- Simplificando las regiones