Discriminación II

Graciela Boente

Estadístico U de Rao

Sean $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\Sigma} > 0$ y $\mathbf{W} \sim \mathcal{W}(\boldsymbol{\Sigma}, p, m)$ independientes entre sí, $m \geq p$.

Particionemos a $\mathbf{y}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ y $\mathbf{W}$ de la siguiente forma

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \qquad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix}$$

con $\mathbf{y}^{(i)}, \boldsymbol{\mu}^{(i)} \in \mathbb{R}^{p_i}$, $\boldsymbol{\Sigma}_{ii} \in \mathbb{R}^{p_i \times p_i}$, $\mathbf{W}_{ii} \in \mathbb{R}^{p_i \times p_i}$, $p_1 + p_2 = p$.

Sean

$$T^2_{p,m} = m\, \mathbf{y}^{\mathrm{T}} \mathbf{W}^{-1} \mathbf{y} \qquad T^2_{p_1,m} = m\, \mathbf{y}^{(1)^{\mathrm{T}}} \mathbf{W}_{11}^{-1} \mathbf{y}^{(1)}$$

Definamos

$$\lambda^2_p = \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \qquad \lambda^2_{p_1} = \boldsymbol{\mu}^{(1)^{\mathrm{T}}} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}^{(1)}\,.$$

Recordemos que

$$\begin{aligned} \frac{m-p+1}{p} \frac{T^2_{p,m}}{m} &\sim \mathcal{F}_{p,m-p+1}(\lambda^2_p) \\ \frac{m-p_1+1}{p_1} \frac{T^2_{p_1,m}}{m} &\sim \mathcal{F}_{p_1,m-p_1+1}(\lambda^2_{p_1})\,. \end{aligned}$$

Estadístico U de Rao

Usando que

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\Sigma}_{22.1}^{-1} \boldsymbol{\beta} & -\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\Sigma}_{22.1}^{-1} \\ -\boldsymbol{\Sigma}_{22.1}^{-1} \boldsymbol{\beta} & \boldsymbol{\Sigma}_{22.1}^{-1} \end{pmatrix}$$

con

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \in \mathbb{R}^{p_2 \times p_1}$$

es fácil ver que

$$\lambda_p^2 - \lambda_{p_1}^2 = \boldsymbol{\mu}_{2.1}^{\mathrm{T}} \, \boldsymbol{\Sigma}_{22.1}^{-1} \, \boldsymbol{\mu}_{2.1}$$

con

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}^{(2)} - \boldsymbol{\beta} \boldsymbol{\mu}^{(1)}$$

$$\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$$

Nos interesará testear $H_0 : \lambda_p^2 = \lambda_{p_1}^2$ que es equivalente a $\mu_{2.1} = 0$.

Para testear H_0, nos basaremos en

$$T_{p,m}^2 - T_{p_1,m}^2 = m\, \mathbf{y}_{2.1}^{\mathrm{T}} \mathbf{W}_{22.1}^{-1} \mathbf{y}_{2.1}$$

donde

$$\begin{aligned} \mathbf{y}_{2.1} &= \mathbf{y}^{(2)} - \mathbf{B}\mathbf{y}^{(1)} \qquad & \mathbf{B} = \mathbf{W}_{21}\mathbf{W}_{11}^{-1} \in \mathbb{R}^{p_2 \times p_1} \\ \mathbf{W}_{22.1} &= \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12} \end{aligned}$$

El estadístico $U-$de Rao se define por

$$U = \left\{1 + \frac{T_{p_1,m}^2}{m}\right\} \left\{1 + \frac{T_{p,m}^2}{m}\right\}^{-1}$$

Como

$$\frac{T_{p,m}^2 - T_{p_1,m}^2}{m + T_{p_1,m}^2} = \frac{1}{U} - 1$$

Rechazaremos H_0 si $T_{p,m}^2 - T_{p_1,m}^2$ es grande o sea si U es chico.

Teorema 1

Sean $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\Sigma} > 0$ y $\mathbf{W} \sim \mathcal{W}(\boldsymbol{\Sigma}, p, m)$ independientes entre sí, $m \geq p$.

$$T^2_{2.1} = (m - p_1)\, \frac{T^2_{p,m} - T^2_{p_1,m}}{m + T^2_{p_1,m}}$$

a) Bajo $H_0 : \lambda^2_p = \lambda^2_{p_1}$ que es equivalente a $H_0 : \boldsymbol{\mu}_{2.1} = 0$ se tiene

$$\frac{m-p+1}{p-p_1}\,\frac{T^2_{2.1}}{m-p_1} = \frac{m-p+1}{p-p_1}\,\frac{T^2_{p,m} - T^2_{p_1,m}}{m + T^2_{p_1,m}} \sim \mathcal{F}_{p-p_1,m-p+1}$$

y $T^2_{2.1}$ es independiente de $T^2_{p_1}$

El factor $m + T^2_{p_1,m}$ aparece pues la

$$\mathrm{VAR}(\mathbf{y}_{2.1}\Big|_{\mathbf{y}^{(1)},\mathbf{W}_{11}}) = \boldsymbol{\Sigma}_{22.1}\,(1 + T^2_{p_1,m}/m)$$

Teorema 1

b) Si $\lambda_p^2 \neq \lambda_{p_1}^2$, la distribución de $T_{2.1}^2$ condicional a $T_{p_1}^2$ es un Hotelling no central, o sea,

$$\frac{m-p+1}{p-p_1}\frac{T_{2.1}^2}{m-p_1}\Big|_{T_{p_1}^2} \sim \mathcal{F}_{p-p_1,m-p+1}(\nu)$$

con

$$\nu = \frac{\lambda_p^2 - \lambda_{p_1}^2}{1+\frac{T_{p_1}^2}{m}}$$

Contribución de componentes a la discriminación

Sea $\mathbf{x} \in \mathbb{R}^p$ tal que

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \text{ en } \mathcal{P}_1 \qquad \mathbf{x} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \text{ en } \mathcal{P}_2$$

Queremos determinar la contribución de las últimas componentes de $\mathbf{x}$ para discriminar entre $\mathcal{P}_1$ y $\mathcal{P}_2$.

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix} \qquad \mathbf{x}^{(1)} = (x_1, \ldots, x_d)^{\mathrm{T}} \qquad \mathbf{x}^{(2)} = (x_{d+1}, \ldots, x_p)^{\mathrm{T}}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} \boldsymbol{\mu}_1^{(1)} \\ \boldsymbol{\mu}_1^{(2)} \end{pmatrix} \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} \boldsymbol{\mu}_2^{(1)} \\ \boldsymbol{\mu}_2^{(2)} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

$$\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \begin{pmatrix} \boldsymbol{\delta}_1^{(1)} \\ \boldsymbol{\delta}_1^{(2)} \end{pmatrix} \qquad \boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{pmatrix} \boldsymbol{\alpha}_1^{(1)} \\ \boldsymbol{\alpha}_1^{(2)} \end{pmatrix}$$

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$$

Contribución de componentes a la discriminación

- $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}$ independientes $\mathbf{x}_{1i} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$

- $\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}$ independientes $\mathbf{x}_{2i} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$

Las dos muestras independientes entre sí. Sean

- $\mathbf{d} = \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 \sim N(\boldsymbol{\delta}, \frac{n_1 + n_2}{n_1\, n_2}\, \boldsymbol{\Sigma})$

- $\mathbf{y} = \sqrt{\frac{n_1\, n_2}{n_1+n_2}}\mathbf{d} \sim N(\sqrt{\frac{n_1\, n_2}{n_1 + n_2}}\, \boldsymbol{\delta}, \boldsymbol{\Sigma})$, $\mathbf{y}^{(2)} = \sqrt{\frac{n_1\, n_2}{n_1+n_2}}\mathbf{d}^{(2)}$,

- $\mathbf{y}^{(1)} = \sqrt{\frac{n_1\, n_2}{n_1+n_2}}\, \mathbf{d}^{(1)} = \sqrt{\frac{n_1\, n_2}{n_1+n_2}}\, (\overline{x}_{1,1} - \overline{x}_{2,1}, \ldots, \overline{x}_{1,d} - \overline{x}_{2,d})^{\mathrm{T}}$.

- $\widehat{\boldsymbol{\alpha}} = \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_1^{(1)} \\ \widehat{\boldsymbol{\alpha}}_1^{(2)} \end{pmatrix}$

Contribución de coordenadas a la discriminación

Definamos

$$\mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}} \qquad i = 1, 2$$

y sea

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2} = \frac{\mathbf{Q}_1 + \mathbf{Q}_2}{n_1 + n_2 - 2} = \frac{\mathbf{U}}{n_1 + n_2 - 2}$$

Luego, $\mathbf{U} \sim \mathcal{W}(\mathbf{\Sigma}, p, n - 2)$ donde $n = n_1 + n_2$.

$$\begin{aligned} \mathbf{W} &= \mathbf{U} = \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{21} \\ \mathbf{U}_{12}^{\mathrm{T}} & \mathbf{U}_{22} \end{pmatrix} \text{ con } \mathbf{U}_{11} = \mathbf{U}_{(d)} \in \mathbb{R}^{d \times d} \\ \mathbf{B} &= \mathbf{U}_{21}\mathbf{U}_{11}^{-1} = \mathbf{S}_{21}\mathbf{S}_{11}^{-1} \qquad \mathbf{B}^{\mathrm{T}} \in \mathbb{R}^{(p-d) \times d} \end{aligned}$$

Una medida de alejamiento entre $\mathcal{P}_1$ y $\mathcal{P}_2$ es

$$\Delta_p^2 = \boldsymbol{\alpha}^{\mathrm{T}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}$$

Tomando sólo las primeras d coordenadas obtenemos

- $\Delta_d^2 = \boldsymbol{\alpha}^{(1)\,\mathrm{T}} \left(\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)} \right) = \boldsymbol{\delta}^{(1)\,\mathrm{T}} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\delta}^{(1)}$
- $\Delta_p^2 = \Delta_d^2 + \left(\boldsymbol{\delta}^{(2)} - \boldsymbol{\beta}\, \boldsymbol{\delta}^{(1)} \right)^{\mathrm{T}} \boldsymbol{\Sigma}_{22.1}^{-1} \left(\boldsymbol{\delta}^{(2)} - \boldsymbol{\beta}\, \boldsymbol{\delta}^{(1)} \right)$

El aumento en la distancia por la contribución de $(x_{d+1}, \ldots, x_p)$ está dado

$$\Delta_p^2 - \Delta_d^2 = \left(\boldsymbol{\delta}^{(2)} - \boldsymbol{\beta}\, \boldsymbol{\delta}^{(1)} \right)^{\mathrm{T}} \boldsymbol{\Sigma}_{22.1}^{-1} \left(\boldsymbol{\delta}^{(2)} - \boldsymbol{\beta}\, \boldsymbol{\delta}^{(1)} \right) = \boldsymbol{\delta}_{2.1}^{\mathrm{T}} \boldsymbol{\Sigma}_{22.1}^{-1} \boldsymbol{\delta}_{2.1}$$

Como se estiman

$$D_p^2 = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right) = \mathbf{d}^{\mathrm{T}} \mathbf{S}^{-1} \mathbf{d}$$

Tomando sólo las primeras d coordenadas obtenemos

- $D_d^2 = \widehat{\boldsymbol{\alpha}}^{(1)\,\mathrm{T}} \left(\overline{\mathbf{x}}_1^{(1)} - \overline{\mathbf{x}}_2^{(1)}\right) = \mathbf{d}^{(1)\,\mathrm{T}} \mathbf{S}_{11}^{-1} \mathbf{d}^{(1)}$
- $D_p^2 - D_d^2 = \left(\mathbf{d}^{(2)} - \mathbf{B}\,\mathbf{d}^{(1)}\right)^{\mathrm{T}} \mathbf{S}_{22.1}^{-1} \left(\mathbf{d}^{(2)} - \mathbf{B}\,\mathbf{d}^{(1)}\right)$

Recordemos que si $n = n_1 + n_2$

$$\mathbb{E} D_p^2 = \frac{n-2}{n-p-3} \left(\Delta_p^2 + \frac{pn}{n_1 n_2}\right)$$

Luego, un estimador insesgado de $\Delta_p^2 - \Delta_d^2$ es

$$\frac{1}{n-2} \left((n-p-3) D_p^2 - (n-d-3) D_d^2\right) - \frac{(p-d)n}{n_1 n_2}$$

Contribución de coordenadas a la discriminación

Queremos estudiar la hipótesis

$$H_0 : \boldsymbol{\alpha}^{(2)} = \mathbf{0}$$

Como

$$\boldsymbol{\alpha}^{(2)} = \boldsymbol{\Sigma}_{22.1}^{-1} \left(\boldsymbol{\delta}^{(2)} - \boldsymbol{\beta}\, \boldsymbol{\delta}^{(1)} \right) = \boldsymbol{\Sigma}_{22.1}^{-1} \boldsymbol{\delta}_{2.1}$$

H_0 es equivalente a

$$H_0^\star : \boldsymbol{\delta}^{(2)} = \boldsymbol{\beta}\, \boldsymbol{\delta}^{(1)}$$

es decir $\mathbb{E}\left(\mathbf{x}^{(2)} | (\mathbf{x}^{(1)}\right)$ es la misma en $\mathcal{P}_1$ y en $\mathcal{P}_2$.

Lo que significa que $(x_{d+1}, \ldots, x_p)$ no da ningún valor discriminante adicional una vez que ya consideramos $(x_1, \ldots, x_d)$.

Contribución de coordenadas a la discriminación

Un test para H_0 se basa en

$$U = \frac{1 + \frac{c^2\, D_d^2}{m}}{1 + \frac{c^2\, D_p^2}{m}}$$

con $m = n - 2$ y

$$c^2 = \frac{n}{n_1\, n_2}$$

Bajo H_0

$$\frac{n-p-1}{p-d}\left(\frac{D_p^2 - D_d^2}{m\, c^2 + D_d^2}\right) = \frac{m-p+1}{p-d}\left(\frac{1}{U} - 1\right) \sim \mathcal{F}_{p-d,n-p-1}$$

Selección de variables

- $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{1,n_i}$ independientes $\mathbf{x}_{i,j} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$

Las muestras independientes entre sí.

$$\mathbf{Q}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}} \qquad \overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{i,j}$$

La suma de cuadrados dentro de grupos es

$$\mathbf{U}_p = \mathbf{Q}_1 + \cdots + \mathbf{Q}_k = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}}$$

La suma de cuadrados entre poblaciones es

$$\mathbf{H}_p = \sum_{i=1}^{k} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^{\mathrm{T}} \qquad \overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{\mathbf{x}}_i$$

Selección de variables

Tenemos que

- $\mathbf{U}_p \sim \mathcal{W}(\mathbf{\Sigma}, p, n-k)$ con $n = \sum_{i=1}^{k} n_i$
- Bajo $H_0 : \boldsymbol{\mu}_1 = \ldots, \boldsymbol{\mu}_k$
  - $\mathbf{H}_p \sim \mathcal{W}(\mathbf{\Sigma}, p, k-1)$
  - $\mathbf{H}_p$ es independiente de $\mathbf{U}$.

Sean $\mathbf{U}_d$ y $\mathbf{H}_d$ los estadísticos basados en las primeras d componentes.

Selección de variables

Para testear $H_{0,d} : \boldsymbol{\mu}_1^{(1)} = \cdots = \boldsymbol{\mu}_k^{(1)}$ usábamos el estadístico de Wilks

$$\Lambda_d = \frac{|\mathbf{U}_d|}{|\mathbf{U}_d + \mathbf{H}_d|} = \Lambda(n-1, d, k-1)$$

Un test para la información adicional que brinda la $(d+1)-$ésima componente x_{d+1} está dado por

$$F_d = \frac{n-k-d}{k-1}\left(\frac{\Lambda_d}{\Lambda_{d+1}} - 1\right)$$

Luego, podemos elegir la variable que maximiza F_d o equivalentemente, que minimiza Λ_{d+1}.

Después de elegida esta variable, el conjunto se reexamina y se elimina la variable con menor valor de F

Selección de variables

Cuando $k = 2$, a mayores valores de F corresponden menores valores de e_{act}, para $k > 2$ esto no tiene porque ocurrir.

Este criterio, tiende a separar grupos que ya están bien separados más que a separar aquellos que no están bien separados. Por eso, no se recomienda.

Selección de variables

Otra forma consiste en

- Fijemos primero $d \leq p$
  - Se eligen d variables $x_{s_1}, \ldots, x_{s_d}$. Sea $\mathcal{S}_d = \{s_1, \ldots, s_d\}$
  - Construir con $x_{s_1}, \ldots, x_{s_d}$ la regla de clasificación.
  - Calcular el error de convalidación cruzada $e_{cv,\mathcal{S}_d}$.
  - Elegir el subconjunto $\mathcal{S}_d$ óptimo en el sentido que minimiza $e_{cv,\mathcal{S}_d}$.
  - Sea $e_{cv,d}$ el valor óptimo.
- Comparar para distintos valores de d el valor $e_{cv,d}$ de modo a elegir el d óptimo.

Problema: Alto costo computacional.

Habbema y Hermans (1977) mostraron que este método combinado con clasificación noparamétrica basada en núcleos es mejor entre varios elegidos por los autores.

Recordemos que la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\mathcal{G}_{i,0} \quad = \quad \{\mathbf{x} \in \mathbb{R}^p : \quad \pi_\ell \, f_\ell(\mathbf{x}) < \pi_i \, f_i(\mathbf{x}) \quad \forall \ell \neq i\}$$

En la mayoría de los casos f_i es desconocida y por lo tanto, en base a una muestra

- $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}$ independientes $\mathbf{x}_{i,j} \sim f_i$

debería estimarla en el punto $\mathbf{x}_0$ a clasificar.

Métodos basados en núcleos

Sean

- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}_i \sim f$, $\mathbf{x}_i \in \mathbb{R}^p$
- $h = h_n > 0$, $h_n \to 0$, $nh_n \to \infty$
- $\mathcal{V}_h(\mathbf{x}) = \{\mathbf{u} \in \mathbb{R}^p : d(\mathbf{x}, \mathbf{u}) \leq h\}$ y $\mathcal{V}_1 = \mathcal{V}_1(0)$
- Un estimador ingenuo de $f(\mathbf{x})$ es

$$\widehat{f}_n(\mathbf{x}) = \frac{\#\{\mathbf{x}_j : \mathbf{x}_j \in \mathcal{V}_h(\mathbf{x})\}}{nh^p \lambda(\mathcal{V}_1)} = \frac{1}{nh^p} \sum_{j=1}^{n} \frac{1}{\lambda(\mathcal{V}_1)} \mathbb{I}_{\mathcal{V}_h(\mathbf{x})}(\mathbf{x}_i)$$

Métodos basados en núcleos

El estimador de Nadaraya–Watson se define como

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{nh^p} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$$

con $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ tal que $\int \mathcal{K}(\mathbf{u}) d\mathbf{u} = 1$.

Un caso particular, corresponde a tomar

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{nh^p} \sum_{j=1}^{n} K\left(\frac{\|\mathbf{x}_i - \mathbf{x}\|}{h}\right)$$

con $K : \mathbb{R} \to \mathbb{R}$.

Métodos basados en núcleos

Podríamos tomar como $\|\mathbf{x}-\mathbf{u}\|$ (o como $d(\mathbf{u},\mathbf{x})$) la norma (o distancia) de Mahalanobis, o sea,

$$d(\mathbf{x},\mathbf{u}) = \|\mathbf{x}-\mathbf{u}\|^2 = (\mathbf{x}-\mathbf{u})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{u})$$

Eligiendo

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$$

resulta

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{nh^p|\mathbf{S}|^{\frac{1}{2}}}\sum_{j=1}^{n}\exp\left(-\frac{1}{2h^2}(\mathbf{x}_i-\mathbf{x}_0)^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_i-\mathbf{x}_0)\right)$$

Métodos basados en núcleos

Basados en $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}$ independientes $\mathbf{x}_{i,j} \sim f_i$, consideremos

$$\widehat{f}_{i,n_i}(\mathbf{x}) = \frac{1}{n_i h^p} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}_{i,j} - \mathbf{x}}{h}\right)$$

Clasifico $\mathbf{x}_0 \in \mathcal{P}_i$ si

$$\pi_i \, \widehat{f}_{i,n_i}(\mathbf{x}_0) = \max_{1 \leq \ell \leq k} \pi_\ell \, \widehat{f}_{\ell,n_\ell}(\mathbf{x}_0)$$

Si π_i son desconocidos tomamos $\widehat{\pi}_i = n_i/n$ con $n = \sum_{i=1}^{k} n_i$, resultando en:

Clasifico $\mathbf{x}_0 \in \mathcal{P}_i$ si

$$\frac{1}{nh^p} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}_{i,j} - \mathbf{x}_0}{h}\right) = \max_{1 \leq \ell \leq k} \frac{1}{nh^p} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}_{\ell,j} - \mathbf{x}_0}{h}\right)$$

Métodos basados en $m-$vecinos más cercanos

- Definamos una distancia entre puntos $d(\mathbf{x}, \mathbf{x}_{i,j})$, por ejemplo la distancia de Mahalanobis $d(\mathbf{x}, \mathbf{x}_{i,j}) = (\mathbf{x} - \mathbf{x}_{i,j})^{\mathrm{T}} \mathbf{S}_i^{-1} (\mathbf{x} - \mathbf{x}_{i,j})$

- Calculemos la distancia del punto $\mathbf{x}_0$ a clasificar a todos los puntos $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}$, $1 \leq i \leq k$.

$$d_{ij} = d(\mathbf{x}_0, \mathbf{x}_{i,j})$$

- Ordenemos las distancias de menor a mayor, $d^{(1)}, \ldots, d^{(n)}$ y sea $D = d^{(m)}$ la distancia de $\mathbf{x}_0$ a su $m-$ésimo vecino más cercano.

- Sea $M_i = \#\{\mathbf{x}_{i,j} : d(\mathbf{x}_0, \mathbf{x}_{i,j}) \leq D\}$, o sea la cantidad de puntos de la muestra i que está entre los m más próximos.

$$\widehat{f}_{i,n_i}(\mathbf{x}) = \frac{M_i}{n_i D^p \lambda(\mathcal{V}_1)}$$

Métodos basados en $m-$vecinos más cercanos

Clasifico $\mathbf{x}_0 \in \mathcal{P}_i$ si

$$\pi_i \, \widehat{f}_{i,n_i}(\mathbf{x}_0) \quad = \quad \max_{1 \leq \ell \leq k} \pi_\ell \, \widehat{f}_{\ell,n_\ell}(\mathbf{x}_0)$$

$$\Updownarrow$$

Clasifico $\mathbf{x}_0 \in \mathcal{P}_i$ si

$$\pi_i \, \frac{M_i}{n_i} = \max_{1 \leq \ell \leq k} \pi_\ell \, \frac{M_\ell}{n_\ell}$$

Si π_i son desconocidos tomamos $\widehat{\pi}_i = n_i/n$ con $n = \sum_{i=1}^{k} n_i$, resultando en:

Clasifico $\mathbf{x}_0 \in \mathcal{P}_i$ si $\quad \dfrac{M_i}{m} = \max\limits_{1 \leq \ell \leq k} \dfrac{M_\ell}{m}$

O sea, clasifico el punto $\mathbf{x}_0$ en la población con mayor frecuencia de puntos entre los m más cercanos.

Métodos basados en $m-$vecinos más cercanos

En particular, si $m = 1$, corresponde a asigna $\mathbf{x}_0$ al grupo al que pertenece la observación más cercana.

Cómo elegir m?

- $m = \sqrt{n_j}$ con n_j un tamaño del grupo promedio
- Probar con distintos m, obtener $e_{cv}(m)$ y elegir el valor de m que minimiza $e_{cv}(m)$.

Discriminación logística

$$\log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$$

Supongamos que $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_1$ si $\mathbf{x} \in \mathcal{G}_{1,0}$ donde

$$\begin{aligned}
\mathcal{G}_{1,0} &= \{\mathbf{x} \in \mathbb{R}^p : \quad \log\left(\frac{f_1(\mathbf{x})}{f_1(\mathbf{x})}\right) > \log\left(\frac{\pi_2}{\pi_1}\right)\} \\
&= \{\mathbf{x} \in \mathbb{R}^p : \quad \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} > \log\left(\frac{\pi_2}{\pi_1}\right)\} \\
&= \quad \{\mathbf{x} \in \mathbb{R}^p : \quad \beta_{0,0} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} > 0\}
\end{aligned}$$

donde

$$\beta_{0,0} = \beta_0 + \log\left(\frac{\pi_1}{\pi_2}\right)$$

Discriminación logística

La probabilidad a posteriori es

$$q_1(\mathbf{x}) = \mathbb{P}(G = 1|\mathbf{x} = \mathbf{x}_0) = \frac{\pi_1 f_1(\mathbf{x}_0)}{\pi_1 f_1(\mathbf{x}_0) + \pi_2 f_2(\mathbf{x}_0)} = \frac{\exp(\alpha_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)}{\exp(\alpha_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0) + 1}$$

Como $q_2(\mathbf{x}) = 1 - q_1(\mathbf{x})$ se obtiene

$$\log \frac{q_1(\mathbf{x})}{q_2(\mathbf{x})} = \beta_{0,0} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$$

que se llama el modelo de *posterior odds* (Cox, 1966)

Discriminación logística

Ventaja de discriminación logística

- Sólo necesitamos estimar $d + 1$ parámetros.
- No requiere la especificación de $x|G = j \sim f_j$
- La familia de modelos que incluye es amplia:
  - Normal multivariada con $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$,

$$\beta_0 = -\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \qquad \boldsymbol{\beta} = \boldsymbol{\alpha} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

  - variables dicotómicas independientes,
  - distribuciones multivariadas discretas que siguen un modelo log–lineal, etc.

Discriminación logística

La estimación de $\boldsymbol{\beta} = \boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ mediante el modelo logístico no es eficiente en el caso normal ya que en lugar de estimar los $d(d+1)/2$ términos de $\boldsymbol{\Sigma}$ y los $2d$ elementos de $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$ con el modelo logístico estimamos solamente $d+1$ parámetros. Por esta razón en caso de normalidad se obtiene un mejor procedimiento con la regla de Fisher, lo que ocurre es que como

$$f_{\mathbf{x},G}(\mathbf{x}, g) = f(g|\mathbf{x}) f_{\mathbf{x}}(\mathbf{x})$$

perdemos información al considerar solamente la distribución condicional $f(g|\mathbf{x})$ que como veremos es lo que hace el modelo logístico en lugar de la conjunta que es lo que hacíamos en el caso que vimos antes para la distribución normal.

Discriminación logística

Podemos pensar en el siguiente modelo:

Sea $y = 1$ si el individuo pertenece al grupo 2 y $y = 0$ si pertenece al grupo 1.

Supongamos que $y_i|\mathbf{x}_i \sim Bi(1, p_i)$, $1 \leq i \leq n$ donde

$$p_i = \mathbb{P}(y_i = 1|\mathbf{x}_i)$$

El modelo logístico modela la probabilidad de éxito a través de la función de distribución logística como

$$p_i = p(\mathbf{x}_i) = \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\right)}$$

Discriminación logística

Como

$$1 - p_i = \frac{1}{1 + \exp\left(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right)}$$

resulta que

$$g_i = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i$$

La variable g_i representa en escala logarítmica la diferencia de las probabilidades de pertenecer a la población 1 y 2.

$$O_i = \frac{p_i}{1 - p_i} = \exp(\beta_0) \prod_{j=1}^{d} \exp(\beta_j)^{x_{i,j}}$$

Luego, $\exp(\beta_0)$ y $\exp(\beta_j)$ indican cuanto se modifican los probabilidades por unidad de cambio en las variables x_j.

Discriminación logística

Por otra parte, si $p_i = 1/2$, el valor

$$x_{i,1} = \frac{\beta_0}{\beta_1} - \sum_{j=2}^{d} \frac{\beta_j}{\beta_1} x_{i.j}$$

representa el valor de la variable x_1 que hace igualmente probable que la observación i cuyas restantes variables son $x_{i,2}, \ldots, x_{i,d}$ pertenezca a $\mathcal{P}_1$ o $\mathcal{P}_2$.

Estimación MV

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

Luego,

$$\begin{aligned} \log(L(\beta_0, \boldsymbol{\beta})) &= \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \log(1-p_i) \\ &= \sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i) \end{aligned}$$

Estimación MV

- $D(\beta_0, \boldsymbol{\beta}) = -2\log(L(\beta_0, \boldsymbol{\beta}))$ se llama la desviación del modelo y
- $d_i = -2\{y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\}$ se llama la desviación del dato i y mide el ajuste del modelo al dato $(y_i, \mathbf{x}_i)$
  - Si $y_i = 1$, la observación pertenece a $\mathcal{P}_2$, y $d_i = -2\log(p_i)$, por lo que la observación tendrá una *desviación o deviance* grande si la probabilidad de pertenecer a $\mathcal{P}_2$ es pequeña, o sea, cuando una observación está mal explicada por el modelo.
  - Si $y_i = 0$, la observación pertenece a $\mathcal{P}_1$, y $d_i = -2\log(1 - p_i)$, por lo que la observación tendrá una *desviación o deviance* grande si p_i es grande, o sea, el modelo ajusta mal ese dato.

Estimación MV

$$\ell(\beta_0, \boldsymbol{\beta}) = \log(L(\beta_0, \boldsymbol{\beta})) = \sum_{i=1}^{n} y_i \left[\beta_0 + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right] + \sum_{i=1}^{n} \log(1 - p_i)$$

Como

$$\frac{\partial p_i}{\partial \beta_0} = p_i(1 - p_i) \qquad \frac{\partial p_i}{\partial \boldsymbol{\beta}} = \mathbf{x}_i p_i(1 - p_i)$$

Obtenemos que

$$S_0(\beta_0, \boldsymbol{\beta}) = \frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right)}$$

$$\mathbf{S}(\beta_0, \boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right)} \right)$$

Estimación MV

Para encontrar el EMV necesitamos resolver $S_0(\beta_0, \boldsymbol{\beta}) = 0$ y $\mathbf{S}(\beta_0, \boldsymbol{\beta}) = 0$ y se puede utilizar Newton–Raphson.

$$\begin{aligned} 0 &= \sum_{i=1}^{n} y_i - \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right)} \\ 0 &= \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right)} \right) \end{aligned}$$

Luego,

$$\begin{aligned} \sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right)} = \sum_{i=1}^{n} \widehat{y}_i \\ \sum_{i=1}^{n} \mathbf{x}_i y_i &= \sum_{i=1}^{n} \mathbf{x}_i \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i\right)} = \sum_{i=1}^{n} \widehat{y}_i \mathbf{x}_i \end{aligned}$$

Ejemplo

Dos variedades de Mosquitos depredadores *Amerohelea Fasciata* (Af) y *A. pseudofasciata* (Apf). Como son parecidas se los intenta clasificar usando características de f"acil medición. Se usó el largo de la antena y el largo del ala en mm.

Especie	Af	Af	Af	Af	Af	Af	Af	Af	Af
Antena	1.38	1.40	1.24	1.36	1.38	1.48	1.54	1.38	1.56
Ala	1.64	1.70	1.72	1.74	1.82	1.82	1.82	1.90	2.08
Especie	Apf	Apf	Apf	Apf	Apf	Apf			
Antena	1.14	1.20	1.18	1.30	1.26	1.28			
Ala	1.78	1.86	1.96	1.96	2.00	2.00			

Ejemplo

Se consideró como variable

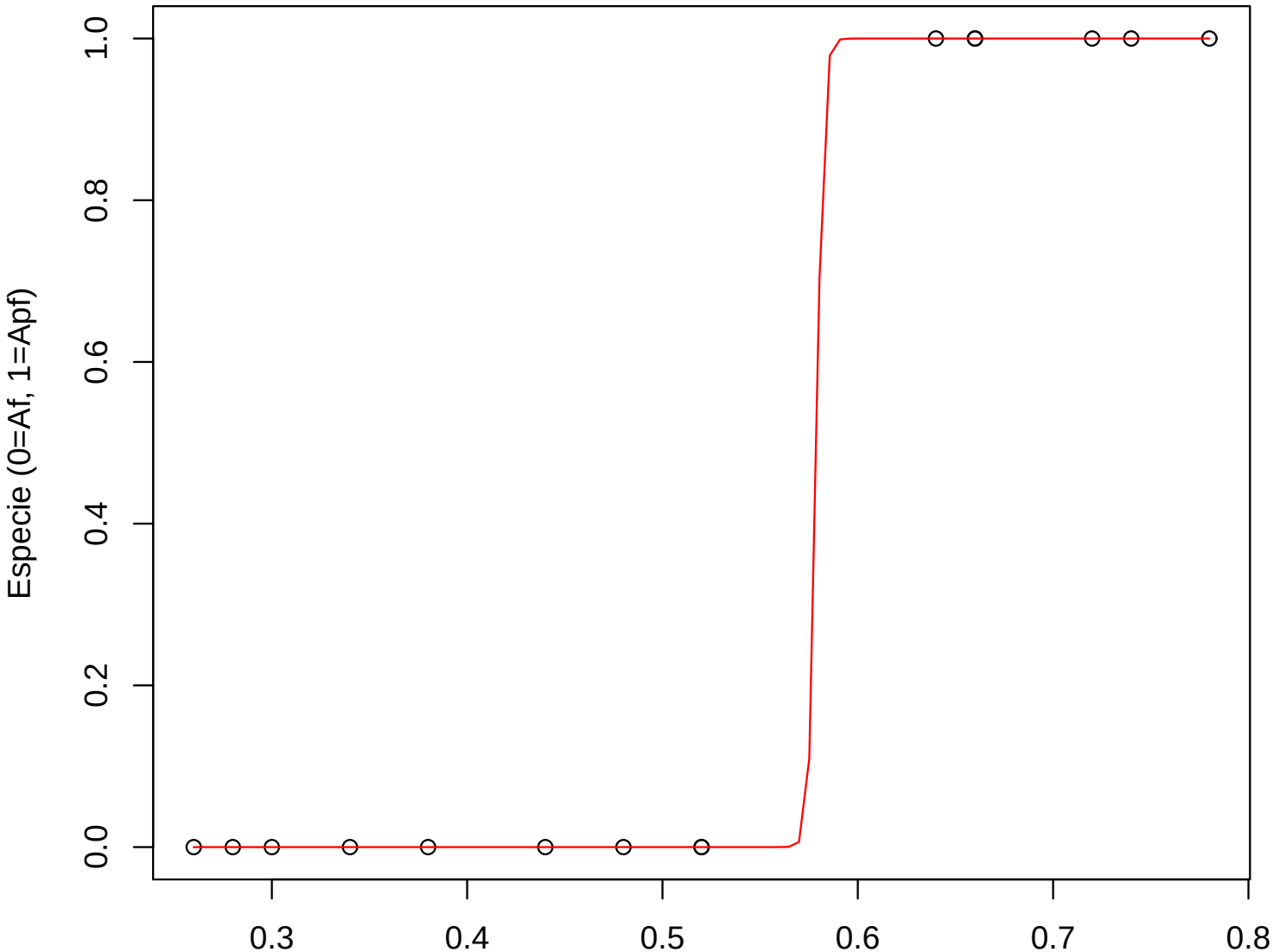
$$x = \text{Largo del Ala} - \text{Largo de la Antena}$$

Al hacer el ajuste el R devuelve

Mensajes de aviso perdidos
glm.fit: fitted probabilities numerically 0 or 1 occurred

Qué ocurre?

Ejemplo


Especie (0=Af, 1=Apf)
0.0
0.2
0.4
0.6
0.8
1.0
0.3
0.4
0.5
0.6
0.7
0.8
Largo Ala−Largo Antena

Ejemplo

Existe c tal que $y_i = 0$ para $x_i < c$ y $y_i = 1$ para $x_i > c$, en este caso

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i:x_i<c} (1-p_i) \prod_{i:x_i>c} p_i$$

Si no hubiera restricciones sobre p_i, el máximo se obtendría tomando

- $p_i = 0$ para $x_i < c$ y
- $p_i = 1$ para $x_i > c$

y valdría 1.

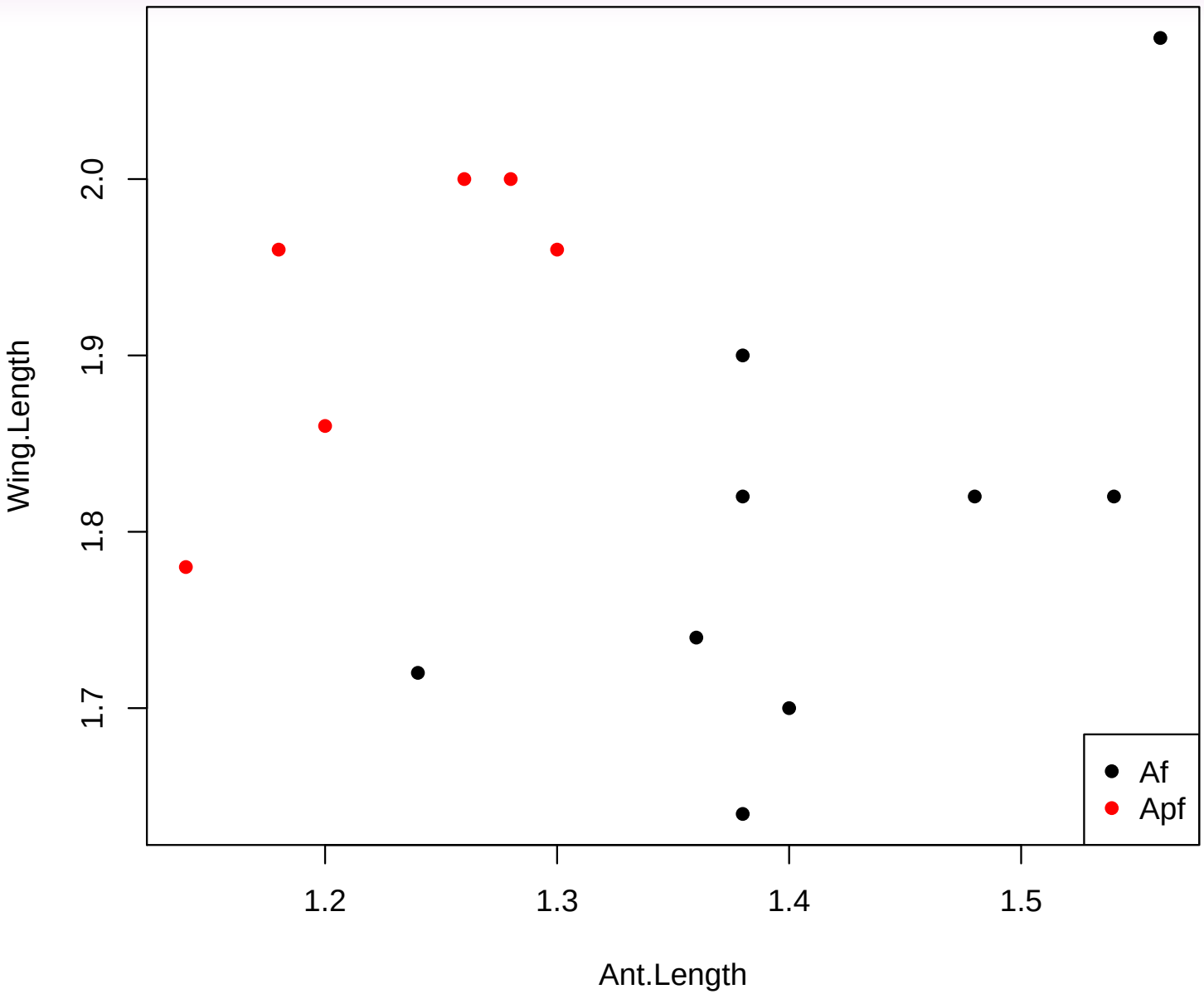
Ejemplo

Podemos lograracercarnos a ese valor haciendo $\beta \to \infty$ y tomando $\beta_0 = -c\beta$, ya que

$$\begin{aligned} \lim_{\alpha\to\infty,\beta_0=-c\beta} p_i(\beta_0,\beta) &= \lim_{\beta\to\infty} \frac{\exp\left[\beta(x_i-c)\right]}{1+\exp\left[\beta(x_i-c)\right]} \\ &= \begin{cases} 0 & \text{si } x_i < c \\ 1 & \text{si } x_i > c \end{cases} \end{aligned}$$

Este problema es llamado separación perfecta. Los datos de las poblaciones $\mathcal{P}_1$ y $\mathcal{P}_2$ yacen a un lado y al otro de un hiperplano.

Para que converja el EMV es necesario que haya superposición de puntos.


Scatterplot of midge data
Wing.Length
Ant.Length
1.7
1.8
1.9
2.0
1.2
1.3
1.4
1.5
Af
Apf

Otros Métodos

- Árboles de clasificación
- Redes Neuronales
- Máquinas del vector soporte