Modelos Lineales de Clasificación

Regresión Logística y Discriminante Lineal

Gabriel Martos Venturini
gmartos@utdt.edu

Universidad Torcuato Di Tella

UNIVERSIDAD TORCUATO DI TELLA

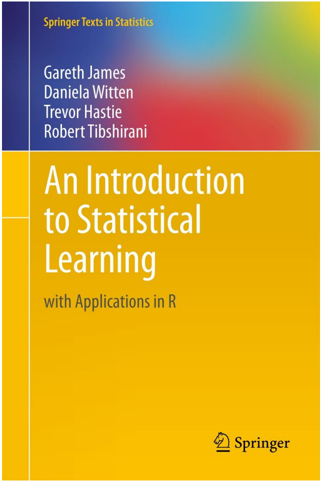
Agenda

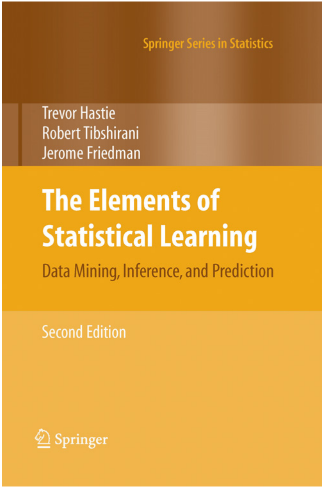
Regresión logística aditiva (GAM)

Discriminante lineal

Discriminante lineal de Fisher

Bibliografía recomendada





ISL: 4.
ESL: 4.1–4.4.

Agenda

Regresión logística aditiva (GAM)

Discriminante lineal

Discriminante lineal de Fisher

- Modelo logístico: $Y \in \{0,1\}$ e $Y|X \sim \text{Ber}(p(X))$.
  - $p(X) = P\{Y=1|X\} = E(Y|X) = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$.
- De manera general, el modelo de regresión logística propone:

$$p(X_1,\ldots,X_p) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

- Modelo no lineal en los parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.
- Fitting: Dada $S_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ resolvemos[1]:

$$\max_{b_0,\mathbf{b}} \underbrace{\sum_{i:y_i=1} \log\left(\frac{e^{b_0+\mathbf{x}_i^T\mathbf{b}}}{1+e^{b_0+\mathbf{x}_i^T\mathbf{b}}}\right) + \sum_{i:y_i=0} \log\left(\frac{1}{1+e^{b_0+\mathbf{x}_i^T\mathbf{b}}}\right)}_{\ell(b_0,\mathbf{b}|S_n)}.$$

- Computamos $\widehat{\boldsymbol{\beta}}$ con métodos numéricos escalables.

[1] $\ell(b_0,\mathbf{b}|S_n)$ es la log–verosimilitud del modelo logístico: *Probabilidad* de observar una muestra como S_n para cada valor que demos a b_0 y $\mathbf{b}$.

(BackUp slide)

- Para simplificar asumamos que hay una sola "X":

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{1}$$

- Dada $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \overset{iid}{\sim} \text{Bern}(p(x))$:

$$L(b_0, b_1 | S_n) = \Pi_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}.$$

- Tomamos logaritmos (transformación monótona):

$$\ell(b_0, b_1 | S_n) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)).$$

- Sustituir $p(x_i)$ por su expresión en 1 y recordá que $y_i \in \{0, 1\}$.
- Maximización de $\ell(b_0, b_1 | S_n)$ vía métodos numéricos estándar.

Predicciones

- Para una observación $\mathbf{x}_{\text{new}} = (X_1 = x_1, \ldots, X_p = x_p)$:

$$\underbrace{\widehat{P}(Y = 1 | X_1 = x_1, \ldots, X_p = x_p)}_{\text{Estimación de } E(Y|\mathbf{x}_{\text{new}})} = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p}}.$$

  - Comando `predict` de `R`.

- Cada predicción probabilística puede transformarse en una predicción $\{0, 1\}$ utilizando un umbral (threshold) $\nu \in [0, 1]$:

$$\widehat{Y} = \begin{cases} 1 \text{ cuando } \widehat{P}(Y = 1 | X_1 = x_1, \ldots, X_p = x_p) \geq \nu, \\ 0 \text{ en otro caso.} \end{cases}$$

- Rule of thumb: $\nu = \pi_1$. Es recomendable elegir ν por VC y atendiendo a las consecuencias de falsos negativos y positivos.
  - Ver caso de estudio con datos de default.

Selección de modelos logísticos

- Flavor estadístico: **pseudo-R^2, AIC, BIC, Deviance**:

$$\text{pseudo-}R^2\text{:}\ 1 - \frac{\ell(\widehat{\beta}|S_n)}{\ell(\widehat{\beta}_0|S_n)}, \quad \text{Deviance:} \ -2\ell(\widehat{\beta}|S_n) + C.$$

- Flavor ML: Tasa de acierto/error, **AUC**, F1–score, etc.

Pred / Realidad	$Y=1$	$Y=0$	Total
$\widehat{Y}_\nu = 1$	TP	FP	TP + FP
$\widehat{Y}_\nu = 0$	FN	TN	FN + TN
Total	P	N	n

Table: Matriz de confusión de un problema clasificación binario.

$$\text{Acierto} = \frac{TP + TN}{n}, \qquad \text{Error} = \frac{FN + FP}{n}.$$

- Utilizar métrica que NO dependen de ν (en negritas).

Otras métricas de evaluación

- $\mathrm{ET}_I = FP$ y $\widehat{P(\mathrm{ET}_I)} = \frac{FP}{n}$ (n tamaño del conjunto).
- $\mathrm{ET}_{II} = FN$ y $\widehat{P(\mathrm{ET}_{II})} = \frac{FN}{n}$.
- $\text{Precisión} = \frac{TP}{TP+FP}$.
- $\text{Recall} = \frac{TP}{TP+FN}$.
- $F_1 = 2\frac{\text{Precisión}\times\text{Recall}}{\text{Precisión}+\text{Recall}}$ (media armónica de Prec. y Rec.)

$$F_\beta = (1+\beta^2)\frac{\mathrm{Prec}\times\mathrm{Rec}}{(\beta^2\mathrm{Prec})+\mathrm{Rec}}, \text{ con } \beta \geq 0.$$

- Cuando se trata de seleccionar modelos, el inconveniente fundamental de estas métricas, es que dependen del umbral ν.

ROC (receiver operating characteristic curve)

$$\text{Sensibilidad (TPR)} = \frac{TP}{P} \ \ vs. \ \ \text{Especificidad (TNR)} = \frac{TN}{N}.$$

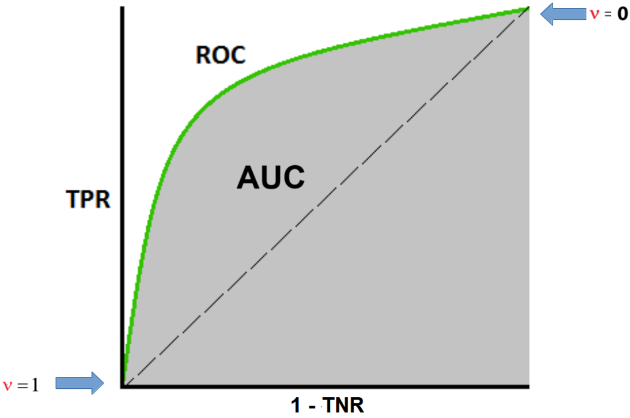
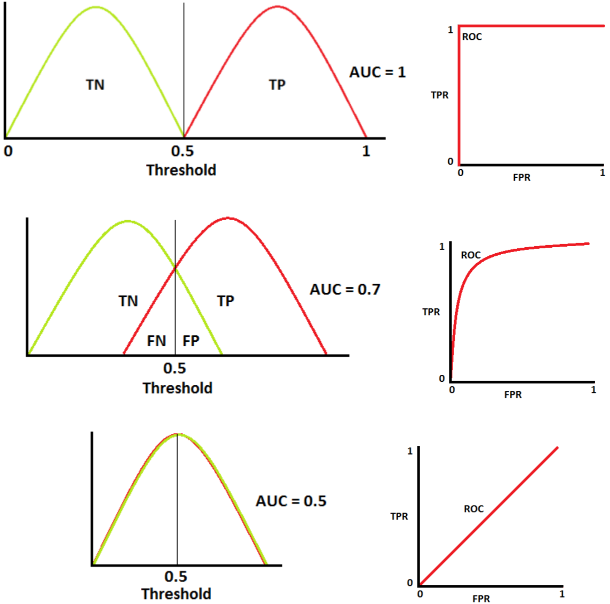
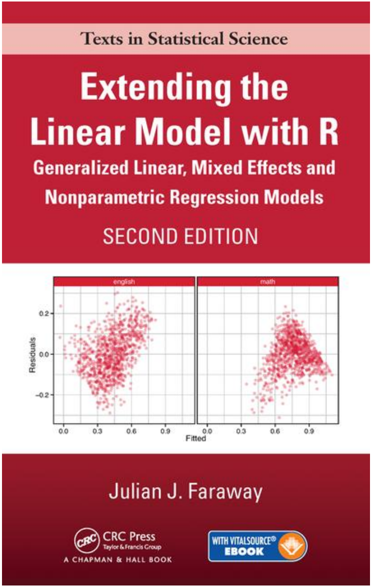


Figure: Si $\downarrow \nu \Rightarrow \uparrow$ TPR & $\downarrow$ TNR (y viceversa).

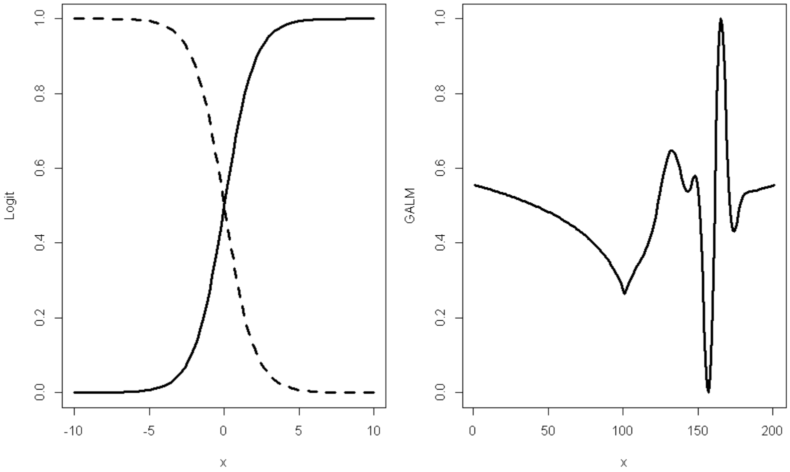
Comparando modelos





- Capítulo 2 (tiene código y casos de estudio en R).

Limitaciones del modelo logístico



- No capturamos efectos no lineales entre x y $P(Y = 1|X = x)$.

Modelos logísticos aditivos

- Regresión logística aditiva:

$$p(X_1, \ldots, X_p) = \frac{e^{\beta_0 + g_1(X_1) + \cdots + g_p(X_p)}}{1 + e^{\beta_0 + g_1(X_1) + \cdots + g_p(X_p)}}.$$

- Cada función g_j se puede *parametrizar* como:

$$g_j(x) = \sum_{b=1}^{B} \beta_{jb}\phi_b(x), \quad j = 1, \ldots, p$$

  - Modelo Aditivo Generalizado (GAM).

- Polinomios o Splines para modelizar cada una de las g's.
  - Función `poly` ó `bs` de la librería `splines`.
- Regularización para evitar el overfitting.
  - Paquete `glmnet`.

Regularización (`glmnet`)

- Cuando $p \gg 0$ el modelo tiene mucha varianza.
- Dada $S_n : \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ aprendemos los parámetros del modelo resolviendo el siguiente problema de optimización

$$\max_{b_0, \mathbf{b}} \left\{ \ell(b_0, \mathbf{b} | S_n) - \lambda \Big((1-\alpha) ||\mathbf{b}||_2^2 + \alpha ||\mathbf{b}||_1 \Big) \right\}.$$

  Penalizas la log–verosimilitud que intentas maximizar con un término que cuantifica la complejidad del modelo logístico.
- $\lambda = 0 \Rightarrow$ Regresión logística sin regularización.
- Ridge: $\alpha = 0$, Lasso: $\alpha = 1$ y ENets: $\alpha \in (0, 1)$.
  - Típicamente elegimos (λ, α) por VC.
- Alternativamente puedes utilizar métodos automáticos de selección de modelos. En particular cuando necesitas interpretar los parámetros del modelo con fines descriptivos.

Regresión logística en R

```
### Librería glm:

glm(formula, family , data, weights, subset, na.action,
    method = binomial(link = "logit"))

# Parámetros sensibles:
formula, data: especificación y datos.

weights, subset, na.action.

family =
  - Modelo logístico: binomial(link = "logit")
  - Regresión lineal: gaussian(link = "identity")
  - Modelo de conteo: poisson(link = "log")
```

- Regresión logística + regularización: `glmnet`.

Defaults TAIWAN Bank (2005)

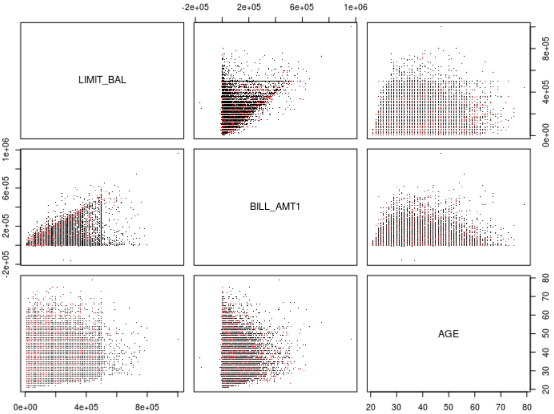


Figure: Instancias en rojo = Clientes en default.

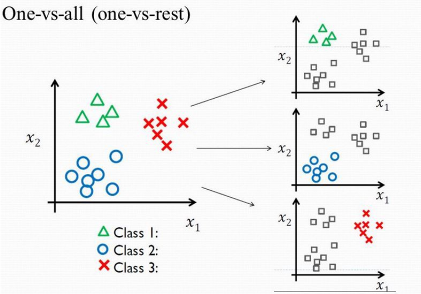
Source: UCI ML repository.

- *default next month*: response variable (Yes = 1, No = 0).
- *LIMIT BAL*: Amount of the given credit: Includes individual consumer credit and his/her family credit.
- SEX: 1 = male; 2 = female.
- EDUCATION: 1 = graduate school; 2 = university; 3 = high school; 4 = others.
- MARRIAGE: 1 = married; 2 = single; 3 = others.
- AGE: Age in years.
- PAY_j: Tracked past monthly repayment status: -1 = pay duly; i = payment delay for i month (i from 1 to 9). (Incoherencias)
- $BILLAMT_j$: Amount of bill statement.
- $PAYAMT_j$: Payment amount.
- j: Time index ranging from April (j=6) to September (j=1).

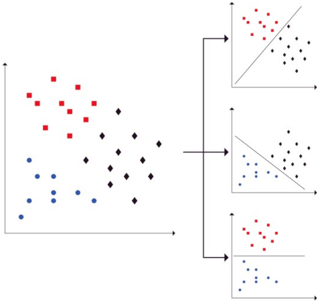
Extensiones (I)

Cuando $Y \in \{1, 2, \ldots, C\}$:

- **One–vs–all**: C modelos de una clase contra las restantes. Predecimos la observación nueva en la clase con mayor probabilidad estimada.



- **One–vs–one**: Entrenamos $C(C-1)/2$ modelos de una clase contra otra. Predicciones por voto y mayoría (puedes utilizar los tresholds *a–priori*).



- Logit multinomial.

Extensiones (II)

Estrategias para problemas de clasificación desbalanceados:

- **Thresholding**: Modificar el umbral de corte (prob apriori).
- **Undersampling**: Eliminar casos en clase mayoritaria.
- **Oversampling**: Remuestreo sobre la clase minoritaria.
- **Reweigthing**: Más peso observaciones en clase minoritaria

$$\ell_{\mathbf{w}}(b_0, \mathbf{b}|S_n) \equiv \sum_{i=1}^{n} w_i \big(y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))\big).$$

- Ninguna estrategia es siempre mejor que las otras. Explorar cada una de estas alternativas (o hacer un mix de todas ellas).

Consideraciones finales sobre modelos logísticos aditivos

- Aspectos positivos:
  - Interpretabilidad de parámetros.
  - Outputs probabilísticos.
  - Es factible modelar relaciones no lineales (versión aditiva).
  - Estimación veloz de parámetros (fácil de escalar en n y p).
- Aspectos negativos:
  - La estimación de los parámetros del modelo resulta sensiblemente afectada cuando hay datos atípicos en la muestra, la muestra presenta fuertes desbalanceos de clases y/o hay datos perdidos (`R` omite las filas con datos faltantes).

Preguntas y aplicaciones

- Reflexionar sobre ventajas y desventajas (computacionales, en términos de la utilidad de los outpus del modelo, etc) tiene resolver un problema de clasificación con una regresión logística en comparación con el modelo de k–vecinos.

- Los parámetros estimados de tu modelo logístico son: $\widehat{\beta}_0 = -6$, $\widehat{\beta}_1 = 0.05$ y $\widehat{\beta}_2 = 1$.

  - Indica la predicción del modelo cuando $X_1 = 40$ y $X_2 = 0.5$.

  - Considerando que en la población $P(Y = 1) = 0.5$, ¿cómo clasificas a la instancia del punto anterior?

  - Manteniendo fijo $X_1 = 40$ ¿cuánto tiene que valer X_2 para que la probabilidad estimada por el modelo de que $Y = 1$ sea $1/2$?

- Resuelve el ejercicio aplicado 10 en § 4.7 (pp 171 ISLR) en `R`.

Agenda

Regresión logística aditiva (GAM)

Discriminante lineal

Discriminante lineal de Fisher

Conceptos generales

- $Y \in \{1, \dots, K\}$ y $\underbrace{\pi_k = p(Y = k)}_{\text{Probabilidad Apriori}}$, para $k = 1, \dots, K$.
- El modelo discriminante asume (para $k = 1, \dots, K$):

$$(X_1, \dots, X_p)|Y = k \sim N_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- De la regla de Bayes:

$$\underbrace{p(Y = k|X = \mathbf{x})}_{\text{Probabilidad Aposteriori}} \propto \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right] \pi_k$$

- Clasificamos una instancia nueva $\mathbf{x}_{\text{new}}$ analizando las magnitudes de cada **función discriminante**:

$$\delta_k(\mathbf{x}_{\text{new}}) = \ln(p(Y = k|X = \mathbf{x}_{\text{new}})).$$

- Bayes: Si $\delta_i(\mathbf{x}_{\text{new}}) \geq \delta_j(\mathbf{x}_{\text{new}})$ para $j \neq i$, entonces $\widehat{y}_{\text{new}} = i$.

Caso univariante ($p = 1$) y binario ($K = 2$)

$$\delta_2(x) - \delta_1(x) = C(\sigma_1, \sigma_2, \pi_1, \pi_2) - \frac{(x_{\text{new}} - \mu_2)^2}{2\sigma_2^2} + \frac{(x_{\text{new}} - \mu_1)^2}{2\sigma_1^2}$$

- Parámetros del modelo: $(\pi_1, \mu_1, \sigma_1^2)$ y $(\pi_2, \mu_2, \sigma_2^2)$.

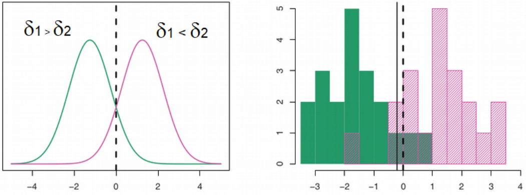


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

- Cuando "$\sigma_1 = \sigma_2$" la regla de clasificación es una función lineal de x_{new}: $\delta_2(x_{\text{new}}) - \delta_1(x_{\text{new}}) = w^T x_{\text{new}} + c$.

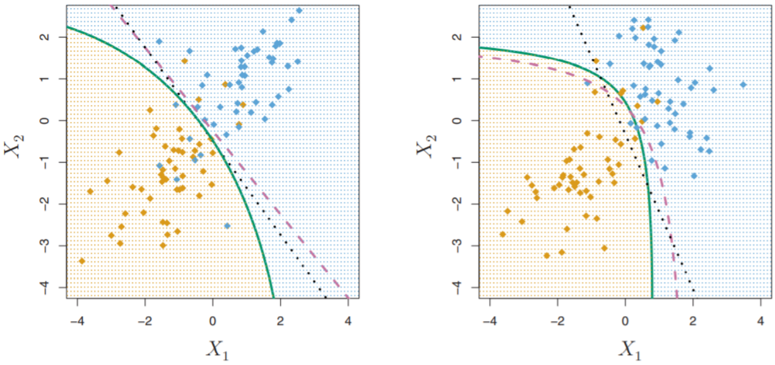
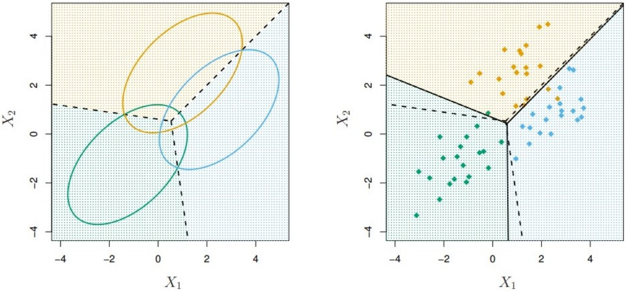


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Caso multivariante y multiclase (BackUp slide)

$$\delta_k(\mathbf{x}) \propto -\frac{1}{2}\left[\ln(|\mathbf{\Sigma}_k|)+(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right]+\ln(\pi_k), \text{ para } k=1,\dots,K.$$

- Reemplazamos $(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ por estimaciones (train).
- LDA: $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_K$ (regla lineal).



- QDA: $\mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j$ (regla cuadrática).

LDA (con $K = 2$) en detalles (BackUp slide)

- Con los datos de Train estimas $\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2$ y computamos:
  - $\widehat{\mathbf{w}} = (\widehat{\boldsymbol{\mu}_2} - \widehat{\boldsymbol{\mu}}_1)^T \widehat{\boldsymbol{\Sigma}}^{-1}$.
  - $\widehat{c} = \log(\widehat{\pi}_2/\widehat{\pi}_1) - 1/2(\widehat{\boldsymbol{\mu}}_2^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_1)$.
- Dada $\mathbf{x}$, la regla de clasificación requiere la evaluación de la función lineal $\widehat{\mathbf{w}}^T \mathbf{x} + \widehat{c}$; luego:

$$\widehat{y} = \begin{cases} 2 \text{ si } \widehat{\mathbf{w}}^T \mathbf{x} + \widehat{c} > 0, \\ 1 \text{ en otro caso.}. \end{cases}$$

- Cuando $p \gg 0$, se suele regularizar eligiendo una estructura parsimónica para $\boldsymbol{\Sigma}$ (por ejemplo asumiendo diagonalidad).
- En contextos de alta dimensión podemos sustituir los features originales por las primeras $r \ll p$ componentes principales.

Discriminante lineal y cuadrático en R

```
# Discriminante lineal
lda(formula, data, ..., subset, na.action)

# Discriminante cuadrático
qda(formula, data, ..., subset, na.action)

# Parámetros sensibles:
formula, data: especificación y datos.
weights, subset, na.action: opciones standard de mod lin.
control = glm.control(epsilon = 1e-8, maxit = 25)
method = método para estimar medias y covarianzas
  - 'mle': máxima verosimilitud.
  - 'moment': método de momentos.
  - 'mve': método robusto (aconsejable).
CV = si es igual a TRUE estima por LOOCV.
```

Caso de estudio

Comentarios finales

- Features deben ser continuos (y aproximadamente normales).
  - Bajo estas hipótesis con n pequeño este es el mejor modelo predictivo.
  - Features mixtos: Discretización y Naive–Bayes.
- Sensibilidad a datos atípicos.
- La *complejidad del modelo* crece a una tasa de p^2.
  - Cuando $p \gg 0$ hay que regularizar (o asumir estructura en $\boldsymbol{\Sigma}$).
    - Algunas estrategias en ESL § 4.3.1–4.3.3.
- Alternativamente hacer PCA (reducir p).
  - Discutimos esta técnica más adelante.
- Estimar el error del modelo utilizando técnicas de VC.

Agenda

Regresión logística aditiva (GAM)

Discriminante lineal

Discriminante lineal de Fisher

- Existen K clases (o subpoblaciones) y en cada una de ellas

  $E(\mathbf{X}|Y=k)=\boldsymbol{\mu}_k \in \mathbb{R}^p \quad \mathrm{Var}(\mathbf{X}|Y=k)=\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p},\ k=1,\ldots,K.$

- Para un vector (de proyección) unitario $\mathbf{e}$ definimos:

$$b(\mathbf{e})=\sum_{i=1}^{K}|\mathbf{e}^T(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})|^2, \quad \text{y} \quad w(\mathbf{e})=\sum_{i=1}^{K}\mathrm{Var}(\mathbf{e}^T\mathbf{X}),$$

  donde $\overline{\boldsymbol{\mu}}=\sum_{i=1}^{K}\boldsymbol{\mu}_i/K$ (promedio de las medias).

- El score discriminante de Fisher $\mathfrak{d}$ se define como:

$$\mathfrak{d}=\max_{\{\mathbf{e}:\|\mathbf{e}\|=1\}}\frac{b(\mathbf{e})}{w(\mathbf{e})}.$$

- Se puede demostrar que $\mathfrak{d}$ es el autovalor más grande de $\mathbf{W}^{-1}\mathbf{B}$, con $\mathbf{W}=\sum_{i=1}^{K}\boldsymbol{\Sigma}_i$ y $\mathbf{B}=\sum_{i=1}^{K}(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})^T$.

- El vector propio $\boldsymbol{\eta}$ asociado a ∂ nos permite construir el score

$$Z = \boldsymbol{\eta}^T \mathbf{X}.$$

- Intuición: El DLF proyecta los datos en 1D de forma de maximizar la capacidad predictiva del modelo de clasificación.

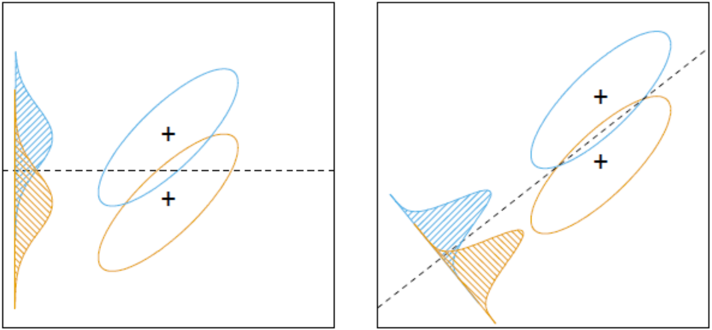


Figure: A la izquierda una proyección poco discriminativa, a la derecha la proyección que propone el Discriminante Lineal de Fisher.

- La instancia $\mathbf{X}$ pertenece a (es generada por) la clase ℓ si:

$$|\boldsymbol{\eta}^T\mathbf{X} - \boldsymbol{\eta}^T\boldsymbol{\mu}_\ell| < |\boldsymbol{\eta}^T\mathbf{X} - \boldsymbol{\eta}^T\boldsymbol{\mu}_v| \text{ para todo } v \neq \ell.$$

- Trabajando con la expresión anterior y asumiendo que $k = 2$, llegamos a la expresión de la función discriminante:

$$h(\mathbf{X}) = \left[\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]^T \boldsymbol{\eta}.$$

  - Luego $\widehat{Y} = 1$ si $h(\mathbf{X}) > 0$.

- En la práctica los parámetros de esta función discriminante son desconocidos. Los estimamos con la muestra de train para luego reemplazarlos en la función discriminante de Fisher.

- Detalles adicionales se pueden consultar por ejemplo en Koch, *Analysis of Multivariate and High–Dimensional Data* (2020).