

EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Clasificación.

Objetivo. Conceptos de clasificación. Regresión logística. Análisis discriminante. Clases desbalanceadas. Curva ROC.

Lectura recomendada: Capítulo 4 de [1].

1. INTRODUCCIÓN

Nos interesa hacer predicciones sobre respuestas **qualitativas**. Donde suponemos que las respuestas corresponden a valores dentro de un conjunto \mathcal{C} .

La tarea de predicción es: considerando que tenemos un vector de **características** X queremos predecir la **respuesta** $Y \in \mathcal{C}$ por medio de una función $C: \mathcal{X} \rightarrow \mathcal{C}$.

Usualmente resolvemos buscando poder predecir la **probabilidad** de que X pertenezca a la categoría \mathcal{C} .

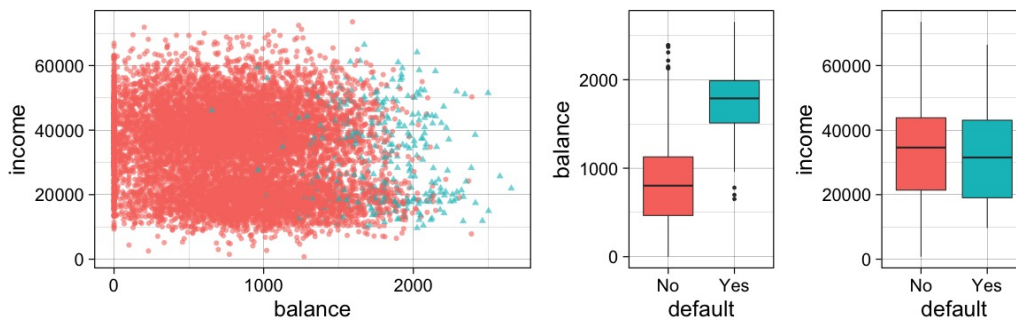


FIGURA 1. Datos de tarjetas de crédito para estudiantes.

1.1. ¿Podemos utilizar un modelo de regresión lineal?

Podemos codificar la respuesta como

$$Y = \begin{cases} 0, & \text{si No} \\ 1, & \text{si Si} \end{cases} \quad (1)$$

por lo que podríamos definir una regla como

Clasificar Si si la predicción $\hat{Y} > 0.5$.

Podríamos ajustar un modelo lineal y calificar como Si si tenemos $\hat{Y} > 0.5$. Este modelo es equivalente a un clasificador por medio de análisis discriminante lineal (LDA). La *garantía* que tenemos es que

$$\mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x). \quad (2)$$

1.2. El problema

La respuesta del modelo lineal no la podemos interpretar como una probabilidad.

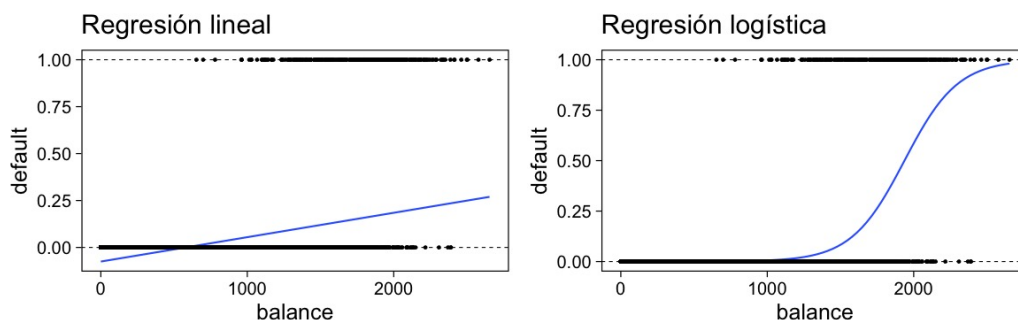


FIGURA 2. Comparación de ajuste entre un modelo de regresión y un logístico.

1.3. Extensiones

Antes habíamos considerado predictores multiclase. Lo mismo podríamos pensar para una predicción en múltiples clases.

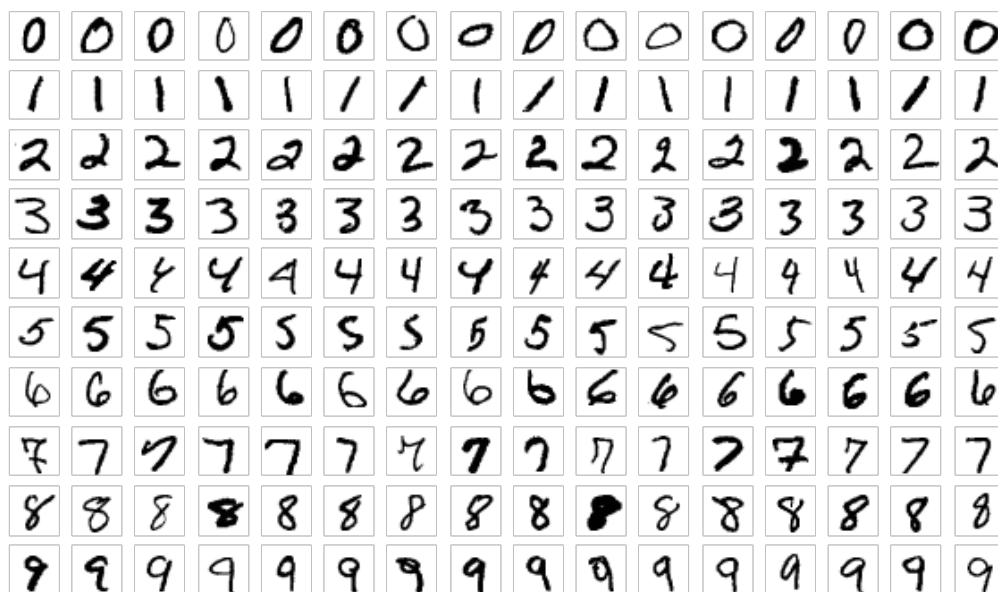


FIGURA 3. Ejemplo multiclase típico (MNIST).

Por supuesto un predictor binario no es apropiado y necesitamos considerar otros modelos (que veremos mas adelante) como Regresión logística multiclase o Análisis Discriminante.

2. REGRESIÓN LOGÍSTICA

Escribiremos $\mathbb{P}(Y = 1|X) = p(X)$ y consideraremos el escenario *simple*. La regresión logística utiliza la transformación

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}. \quad (3)$$

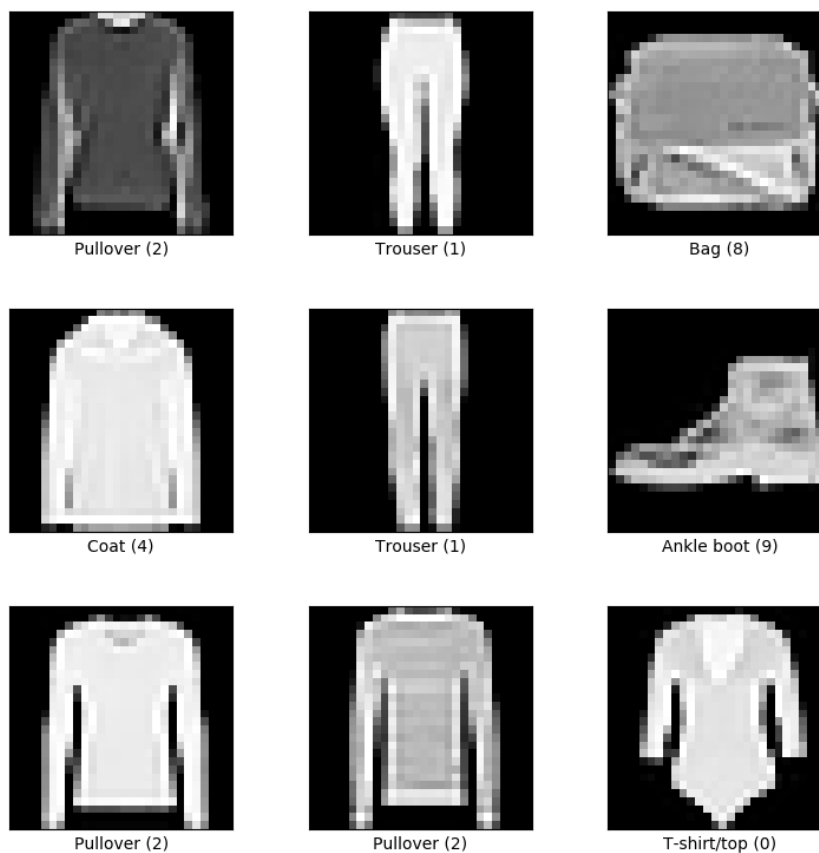


FIGURA 4. Ejemplo multiclase (*Fashion MNIST*).

La función $\sigma(x) = e^x / (1 + e^x)$ se conoce como **función logística**. Análogamente, la función $\sigma(x) = 1 / (1 + e^{-x})$ se conoce como la **función sigmoide**. Son... lo mismo.

Con un poco de álgebra podemos escribir

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X. \quad (4)$$

A esta transformación se le llama **logit** o **log-monio**.

Regresión logística nos ayuda a restringir la salida de nuestro modelo predictivo.

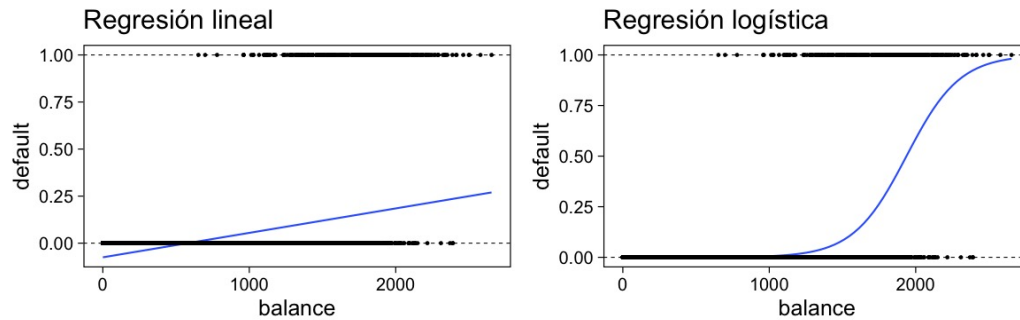


FIGURA 5. La salida del modelo logístico está restringido gracias a la transformación no lineal.

2.1. Estimando los parámetros

Utilizamos el principio de máxima verosimilitud para expresar nuestra función objetivo como

$$\mathcal{L}_n(\beta_0, \beta_1) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}. \quad (5)$$

La verosimilitud es la función de densidad (masa de probabilidad) conjunta de una muestra de n observaciones. Representa el **proceso generador de datos** y la consideramos una función de los parámetros de interés. Con este enfoque, se convierte en la función que dadas las observaciones explica el *origen* de los datos bajo el modelo supuesto.

El objetivo es encontrar

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \max_{\beta_0, \beta_1} \mathcal{L}_n(\beta_0, \beta_1). \quad (6)$$

```
1 modelo <- glm(default ~ balance,
2               family = "binomial",
3               data = data)
```

LISTING 1. Ajuste de modelo logístico.

```
1 modelo >
2   summary()
```

```
1
2 Call:
3 glm(formula = default ~ balance, family = "binomial", data = data)
4
5 Deviance Residuals:
6     Min       1Q   Median       3Q      Max
7  -2.270  -0.146  -0.059  -0.022   3.759
8
9 Coefficients:
10      Estimate Std. Error z value Pr(>|z|)
```

```

11 (Intercept) -10.65133    0.36116   -29.5    <2e-16 ***
12 balance      0.00550    0.00022    24.9    <2e-16 ***
13 ---
14 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
15
16 (Dispersion parameter for binomial family taken to be 1)
17
18 Null deviance: 2920.6  on 9999  degrees of freedom
19 Residual deviance: 1596.5  on 9998  degrees of freedom
20 AIC: 1600
21
22 Number of Fisher Scoring iterations: 8

```

```

1 modelo >
2   broom::tidy() >
3   as.data.frame()

```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-10.6513	0.36116	-29	3.6e-191
2	balance	0.0055	0.00022	25	2.0e-137

LISTING 2. Resumen de modelo logístico (*tidy*).

	balance	response	link	sigma(link)
1	1000	0.0058	-5.15	0.0058
2	2000	0.5858	0.35	0.59

LISTING 3. Tipos de respuesta de un modelo logístico con *glm*.

```

1 modelo <- glm(default ~ balance + income + student,
2               data = data,
3               family = "binomial")

```

LISTING 4. Ajuste de modelo logístico.

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-1.1e+01	4.9e-01	-22.08	4.9e-108
2	balance	5.7e-03	2.3e-04	24.74	4.2e-135
3	income	3.0e-06	8.2e-06	0.37	7.1e-01
4	studentYes	-6.5e-01	2.4e-01	-2.74	6.2e-03

LISTING 5. Resumen del modelo logístico multivariado.

3. CLASIFICACIÓN PARA MAS DE DOS CLASES

Podemos extender a un problema multi-clase

$$\mathbb{P}(Y = k|X) = \frac{e^{\beta_{0,k} + \beta_{1,k}X_1 + \dots + \beta_{p,k}X_p}}{\sum_{\ell=1}^K e^{\beta_{0,\ell} + \beta_{1,\ell}X_1 + \dots + \beta_{p,\ell}X_p}} \quad (7)$$

El modelo de arriba se puede reducir para tener $K - 1$ ecuaciones.

4. ANÁLISIS DISCRIMINANTE

Modelamos la distribución de las características en cada una de las clases de manera separada. Luego, utilizamos el **teorema de Bayes** para obtener la probabilidad $\mathbb{P}(Y|X)$.

Se puede utilizar cualquier distribución, pero nos quedaremos en el caso Gaussiano.

4.1. La regla de Bayes

La regla de Bayes (o teorema de Bayes) lo expresamos en términos de probabilidades condicionales

$$\mathbb{P}(Y = k|X = x) = \frac{\mathbb{P}(X = x|Y = k) \cdot \mathbb{P}(Y = k)}{\mathbb{P}(X = x)}. \quad (8)$$

En el contexto de análisis discriminante utilizamos

$$\mathbb{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^K \pi_\ell f_\ell(x)}, \quad (9)$$

donde

- f_k es la densidad de X para la clase k ,
- π_k es la proporción de datos en la clase k .

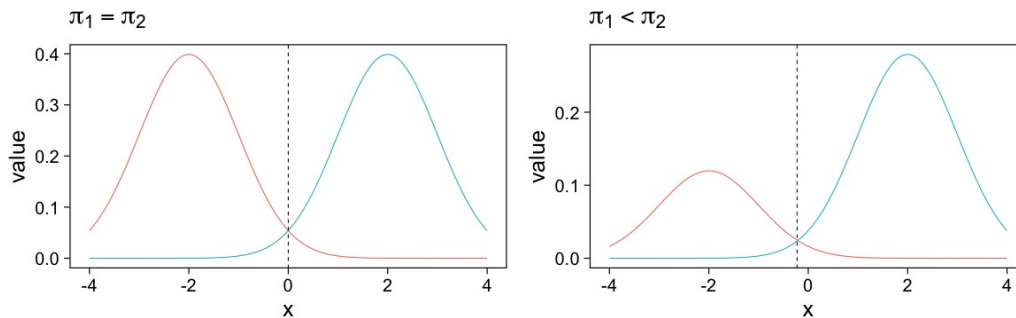


FIGURA 6. Analisis discriminante con densidades Gaussianas.

4.2. ¿Por qué utilizar un LDA?

- En casos con clases **separables**, los estimadores de regresión logística son inestables.
- Si n es pequeña y las densidades son aproximadamente normales en cada una de las clases entonces LDA es mas estable.
- LDA nos permite visualizaciones de dimensiones bajas.

4.3. LDA con $p = 1$.

Asumimos $\sigma_k = \sigma$ para toda k , para poder escribir nuestra $p_k(x)$.

Los términos constantes se eliminan.

Como dijimos antes, clasificamos de acuerdo a cual p_k es la mas grande para x . Lo que nos lleva a buscar el *score* discriminante mas grande

$$\delta_k(x) = x \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k). \quad (10)$$

Tomamos logaritmos y eliminamos los términos que no dependen de k . Notemos que $\delta_k(\cdot)$ es una función *lineal* para x .

4.3.1. *Tarea:* Prueba que para el caso $K = 2$ y $\pi_1 = \pi_2 = .5$ la frontera de la decisión está en

$$x = \frac{\mu_1 + \mu_2}{2}. \quad (11)$$

4.4. ¿Y en la vida real?

Estimamos los parámetros con los criterios usuales.

4.5. LDA con $p > 1$.

La función discriminante es

$$\delta_k(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log(\pi_k). \quad (12)$$

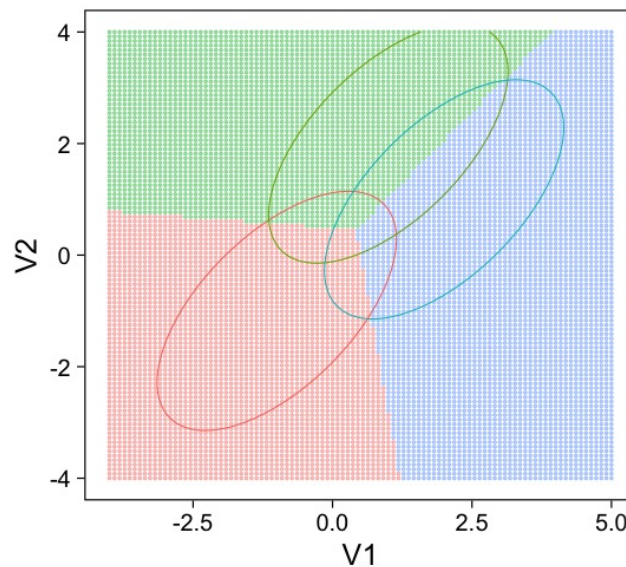


FIGURA 7. LDA en dos dimensiones.

4.6. Predicciones

Una vez que tenemos ajustadas nuestras $\hat{\delta}_k(x)$ podemos utilizarlas para asignar probabilidades de clase:

$$\hat{\mathbb{P}}(Y = k | X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{\ell=1}^K e^{\hat{\delta}_\ell(x)}}. \quad (13)$$

5. LDA EN DATOS

```
1
2 data ← Default
3 data ▷ head()
```

```

1 default student balance income
2 1      No      No      730  44362
3 2      No      Yes      817  12106
4 3      No      No     1074  31767
5 4      No      No      529  35704
6 5      No      No      786  38463
7 6      No      Yes      920   7492

```

```
1 lda.model <- MASS::lda(default ~ balance, data)
```

```

1 library(yardstick)
2
3 data <- data >
4   as_tibble() >
5   mutate(predicted = predict(lda.model)$class,
6           probability = predict(lda.model)$posterior[,1])
7
8 data >
9   conf_mat(truth = default, estimate = predicted)

```

```

1           Truth
2 Prediction No  Yes
3           No 9643 257
4           Yes  24  76

```

LISTING 6. Matriz de confusión.

```

1 data >
2   accuracy(truth = default, estimate = predicted) >
3   as.data.frame()

```

```

1 .metric .estimator .estimate
2 1 accuracy      binary      0.972

```

LISTING 7. Precisión del modelo.

¿Qué hubiera pasado si clasificamos a todos con la clase mayoritaria?

5.1. Evaluación de modelos

La proporción de aciertos para la clase Si es:

```

1 data >
2   recall(truth = default, estimate = predicted) >
3   as.data.frame()

```

```

1 .metric .estimator .estimate
2 1 recall      binary      0.998

```


La proporción de errores para la clase Si se le llama Tasa de Falsos Positivos (aprox. 0.02%).

La proporción de aciertos para la clase No es:

```
1 data >
2   recall(truth = default, estimate = predicted, event_level = 'second') >
3   as.data.frame()
```

```
1   .metric .estimator .estimate
2 1 recall      binary      0.228
```

La proporción de errores para la clase No se le llama Tasa de Falsos Negativos (aprox. 73%).

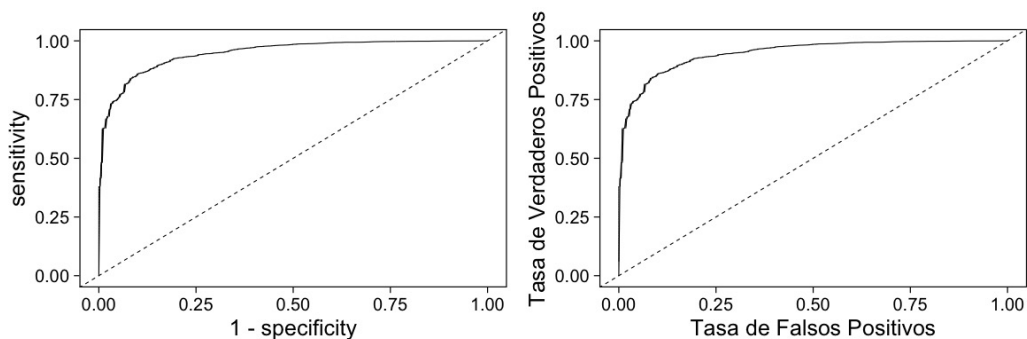
Una combinación de ambos

```
1 data >
2   f_meas(truth = default, estimate = predicted) >
3   as.data.frame()
```

```
1   .metric .estimator .estimate
2 1 f_meas      binary      0.986
```

5.2. El punto de corte

Para las métricas anteriores consideramos que si $\hat{p}(x) > .5$ entonces la predicción de clase será Si. Si cambiamos el punto de corte podemos modificar la tasa de error en ambas.



También podemos pedir un resumen de la gráfica por medio del área bajo la curva (más alto mejor).

```
1 data >
2   roc_auc(default, probability) >
3   as.data.frame()
```

```
1   .metric .estimator .estimate
2 1 roc_auc      binary      0.948
```

LISTING 8. Resumen curva ROC.

6. OTROS MODELO DISCRIMINANTES

Si asumimos diferentes formas para $f_k(x)$ podemos recuperar diferentes modelos discriminantes clásicos.

- Si consideramos un modelo Gaussiano con distintas Σ_k entonces tenemos un **modelo discriminante cuadrático**.
- Si consideramos un modelo Gaussiano con Σ_k una matriz diagonal tenemos el **clasificador Bayesiano ingenuo**.
- Hay muchos mas que se pueden explorar considerando estimadores no-paramétricos.

6.1. Análisis discriminante cuadrático

Si dejamos que el término de varianzas cambie con respecto a k entonces

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k) + \log \pi_k - \frac{1}{2} \log |\Sigma_k|. \quad (14)$$

6.2. Clasificador ingenuo Bayesiano

Cada atributo es independiente de los demás. Tiene muy buenas capacidades predictivas cuando p es grande.

$$\delta_k(x) \propto \log \left(\pi_k \prod_{j=1}^p f_{kj}(x_j) \right) = -\frac{1}{2} \sum_{j=1}^p \left(\frac{(x_j - \mu_{kj})^2}{\sigma_{kj}^2} + \log \sigma_{kj}^2 \right) + \log \pi_k. \quad (15)$$

Se puede utilizar con mezcla de atributos *mixtos*. Es decir, cuando tenemos atributos continuos y discretos.

7. RELACIÓN ENTRE CLASIFICADORES

En el caso binario se puede mostrar que LDA y la función *liga* de regresión logística tienen la misma forma. La diferencia es cómo se estiman los parámetros:

- Con regresión logística aprendemos $\mathbb{P}(Y|X)$ (que se conoce como **aprendizaje discriminante**).
- Con LDA aprendemos $\mathbb{P}(X, Y)$ (que se conoce como **aprendizaje generativo**).

En la práctica los resultados son muy similares.

8. RESUMEN

- Regresión logística es popular, especialmente en clasificación binaria.
- LDA es útil cuando n es pequeña o las clases son separables, y además los supuestos Gaussianos son razonables.
- El clasificador ingenuo Bayesiano es útil cuando tenemos muchas categorías.

9. TEMAS FUERA

- *Confounders*.
- Case-Control sampling.

REFERENCIAS

- [1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. . 1