EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2023 — Clasificación.
Objetivo. Conceptos de clasificación. Regresión logística. Análisis discriminante. Clases desbalanceadas. Curva ROC.
Lectura recomendada: Capítulo 4 de [1]. Capítulo 11 de [2].

1. INTRODUCCIÓN

Nos interesa hacer predicciones sobre respuestas `qualitativas`. Donde suponemos que las respuestas corresponden a valores dentro de un conjunto $\mathcal{C} = \{1, \ldots, K\}$.

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 $C : \mathcal{X} \to \mathcal{C}$.

Usualmente resolvemos esto buscando poder predecir una medida de que tan seguros estamos que X pertenezca a la categoría $\mathcal{C}$. Esto lo denotamos por $p(X)$.

Los modelos que estudiaremos en esta sección no son tan computacionalmente intensivos como otras alternativas.

1.1. Ejercicio

¿Puedes pensar en situaciones que podrían interesarles como un problema de clasificación?

1.2. Créditos a estudiantes

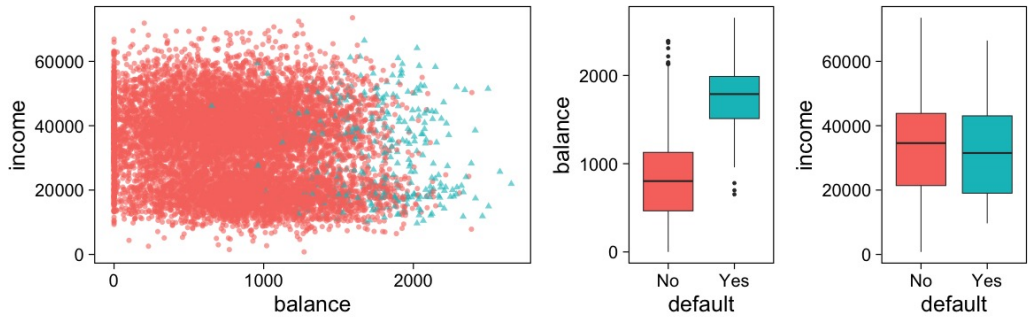


FIGURA 1. *Datos de tarjetas de crédito para estudiantes.*

1.3. ¿Podemos utilizar un modelo de regresión lineal?

Podemos `codificar` la respuesta como

$$Y = \begin{cases} 0, & \text{si } \texttt{Si paga a tiempo} \\ 1, & \text{si } \texttt{No paga a tiempo}\,, \end{cases} \tag{1}$$

por lo que podríamos definir una regla como

Clasificar que **no pagará a tiempo** si la predicción $\hat{Y} > 0.5$.

Podríamos ajustar un modelo lineal y calificar que `no pagará a tiempo` 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

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

1.4. El problema

La respuesta del modelo lineal de regresión **no** la podemos interpretar como una medida de qué tan seguros estamos o, idealmente, como una probabilidad. Además implícitamente estaríamos dispuestos a otorgar cierto orden a las categorías.

Por ejemplo, consideremos un problema de clasificación con mas categorías:

$$Y = \begin{cases} 1 & \text{llueve} \\ 2 & \text{está nublado} \\ 3 & \text{tiembla}\,. \end{cases} \tag{3}$$

Asumimos un orden. Suponemos que la condición de que el día esté nublado está entre lluvia y temblor. Además, reflejamos que la diferencia entre lluvia y nubes, y nubes y temblores es la misma.

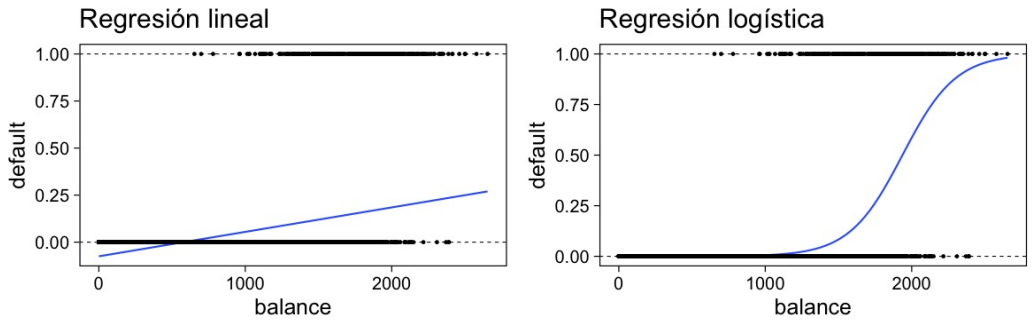


FIGURA 2. *Comparación de ajuste entre un modelo de regresión y un logístico.*

1.5. Otros ejemplos: Problemas multiclase.

Por supuesto, para estas situaciones con $K > 2$, un modelo de regresión o 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

Consideremos el problema binario. Es decir, $Y \in \{0, 1\}$. Podríamos utilizar un predictor de $p(X)$ por medio de un modelo lineal

$$p_{\mathsf{L}}(X) = X^\top \beta\,. \tag{4}$$

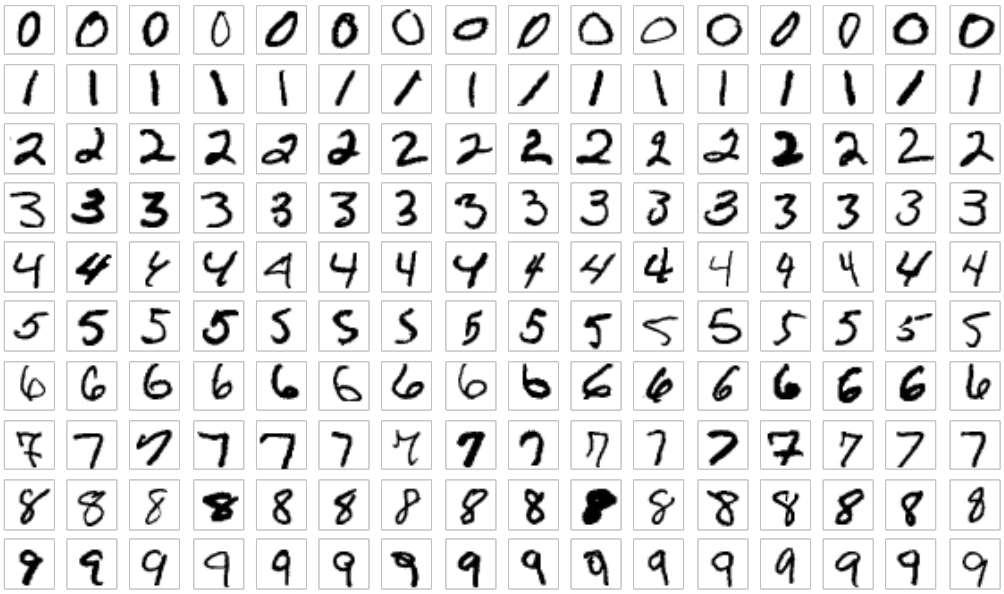


FIGURA 3. *Ejemplo multiclase típico (MNIST).*

El problema es que $p_{\mathsf{L}} : \mathbb{R} \to \mathbb{R}$, cosa que no podemos interpretar como un *score* para la clase Y. Por lo tanto usaremos la transformación

$$p_{\mathsf{L}}(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\,. \tag{5}$$

La función $p(X)$ es la podemos interpretar como un *score*. Pues, en el contexto de nuestro ejemplo, un banco estaría dispuesto a ser conservador para clasificar a un cliente como un cliente que no pagará si $p(X) > 0.15$. 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

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X\,. \tag{6}$$

A esta transformación se le llama `logit` o `log-momio`.

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

2.1. Estimando los parámetros

Una vez definido nuestro predictor para Y en función de características X. Necesitamos definir el objetivo a minimizar para encontrar la mejor configuración dada la muestra que tenemos para ajustar dicho modelo.

Podríamos seguir el camino recorrido en nuestros modelos de regresión y buscar minimizar

$$L(\mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n}(y_i - p(x_i))^2\,. \tag{7}$$

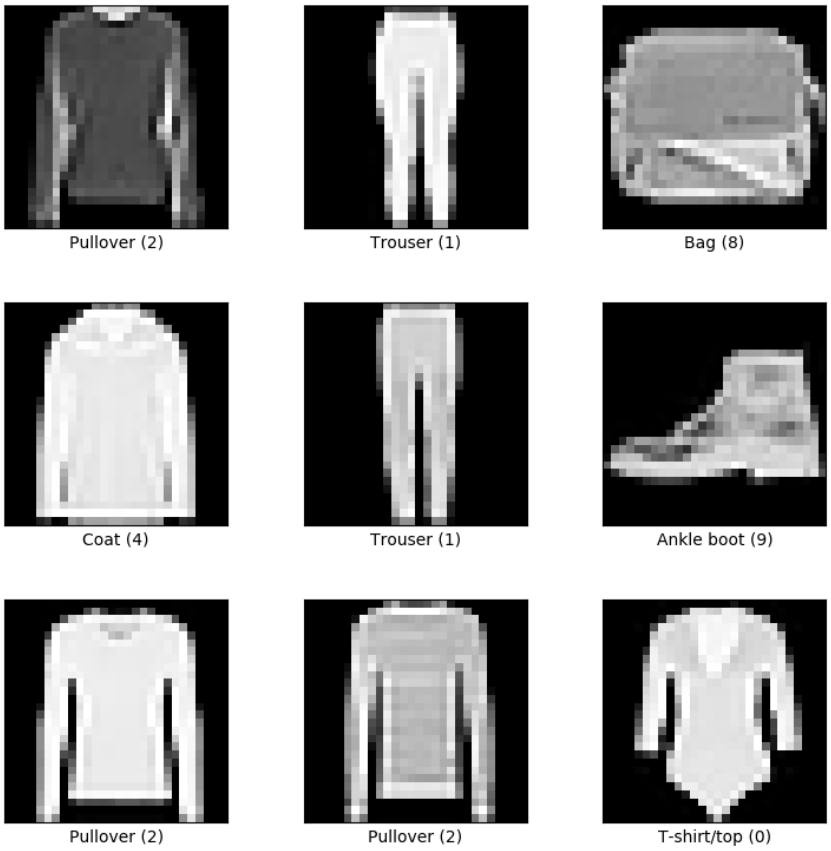


FIGURA 4. *Ejemplo multiclase (Fashion MNIST).*

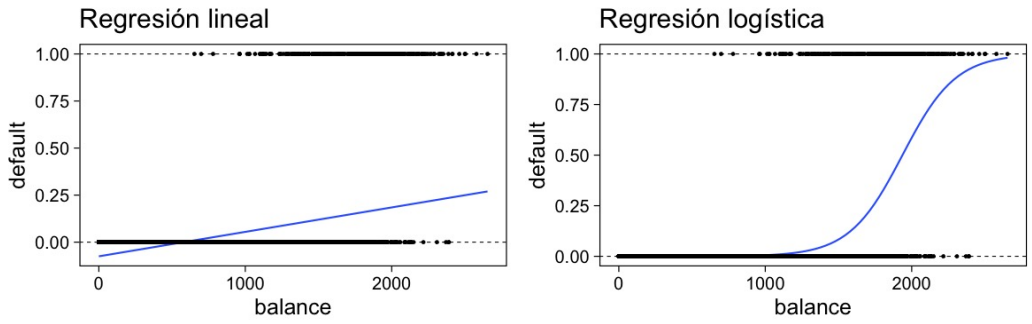


FIGURA 5. *La salida del modelo logistico está restringido gracias a la* `transformación no lineal`.

A esta pérdida se le conoce como la `función de pérdida Brier`. La cual tiene propiedades teóricas deseables. Por ejemplo, ¿qué función minimiza la pérdida de Brier?

Sin embargo, la pérdida de Brier penaliza de manera muy laxa aquellas situaciones donde nos equivocamos al predecir casos positivos, $Y = 1$, por medio de un *score* muy bajo $\hat{p}(X) = \epsilon$.

```
error_1    error_2 diferencia
0.98010    0.99998    0.01988
```

Una métrica adecuada, podríamos argumentar, sería aquella que:

1. Sea una función continua y decreciente en el dominio $[0, 1]$.
2. Si no nos equivocamos, entonces la pérdida es 0.
3. Si nos equivocamos con una $p(X)$ muy pequeña entonces la pérdida es muy grande.

La opción analítica que satisface estos puntos es la pérdida logarítmica:

$$L(y, p(x)) = -\log(p(x))\,, \tag{8}$$

que se muestra en Fig. 6.

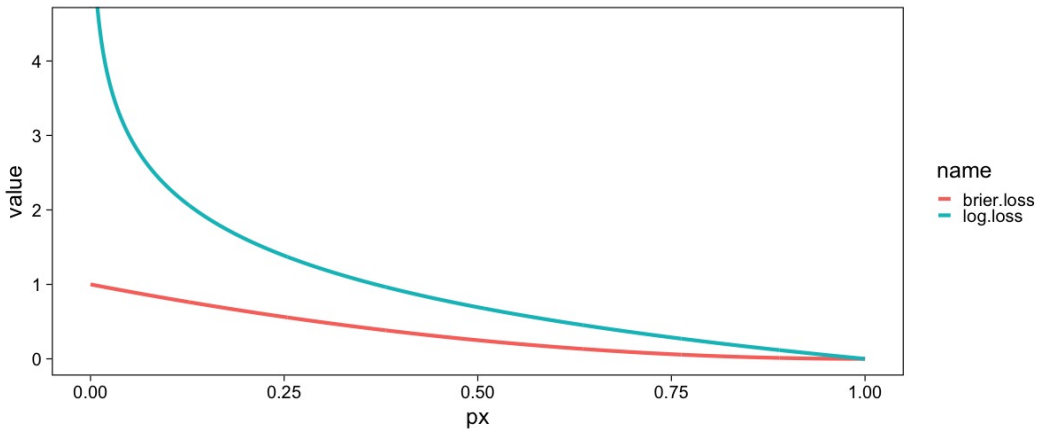


FIGURA 6. *Comparación entre pérdida de Brier y pérdida logarítmica.*

Por supuesto, sólo estamos mostrando los errores cuando $Y = 1$ pero utilizando un argumento análogo podemos definir la pérdida por medio de

$$L(y, p(x)) = -y \log(p(x)) - (1 - y) \log(1 - p(x))\,. \tag{9}$$

La cual se conoce como `pérdida entrópica` o `devianza binomial`.

También podemos ligar dicha pérdida con el principio de máxima verosimilitud para expresar nuestra función objetivo como

$$\mathcal{L}_n(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}\,. \tag{10}$$

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

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0,\beta_1} \mathcal{L}_n(\beta_0, \beta_1)\,. \tag{11}$$

```
modelo ← glm(default ~ balance, family = "binomial", data = data)
```

Listing 1. *Ajuste de modelo logistico.*

```
modelo ▷
  summary()
```

```

Call:
glm(formula = default ~ balance, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.270   -0.146   -0.059   -0.022    3.759

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.65133    0.36116   -29.5   <2e-16 ***
balance       0.00550    0.00022    24.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600

Number of Fisher Scoring iterations: 8
```

```
modelo ▷
  broom::tidy()
```

```
# A tibble: 2 × 5
  term          estimate std.error statistic   p.value
  <chr>            <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept) -10.7       0.361         -29.5 3.62e-191
2 balance       0.00550   0.000220       25.0 1.98e-137
```

Listing 2. *Resumen de modelo logistico (`tidy`).*

```
# A tibble: 2 × 4
  balance response   link ‘sigma(link)‘
    <dbl>    <dbl>  <dbl> <list>
1    1000  0.00575 -5.15  <dbl [1]>
2    2000  0.586    0.347 <dbl [1]>
```

Listing 3. *Tipos de respuesta de un modelo logistico con `glm`.*

```
modelo ← glm(default ~ balance + income + student,
             data = data,
             family = "binomial")
```

Listing 4. *Ajuste de modelo logistico.*

```
# A tibble: 4 × 5
  term             estimate  std.error statistic   p.value
  <chr>               <dbl>      <dbl>     <dbl>     <dbl>
1 (Intercept) -10.9         0.492         -22.1   4.91e-108
2 balance       0.00574     0.000232       24.7   4.22e-135
3 income        0.00000303  0.00000820      0.370 7.12e-  1
4 studentYes   -0.647       0.236          -2.74  6.19e-  3
```

Listing 5. *Resumen del modelo logistico multivariado.*

2.2. Una situación interesante

```
# A tibble: 2 × 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   -3.50     0.0707    -49.6   0
2 studentYes     0.405    0.115       3.52  0.000431
```

```
# A tibble: 4 × 5
  term             estimate  std.error statistic   p.value
  <chr>               <dbl>      <dbl>     <dbl>     <dbl>
1 (Intercept) -10.9         0.492         -22.1   4.91e-108
2 balance       0.00574     0.000232       24.7   4.22e-135
3 income        0.00000303  0.00000820      0.370 7.12e-  1
4 studentYes   -0.647       0.236          -2.74  6.19e-  3
```

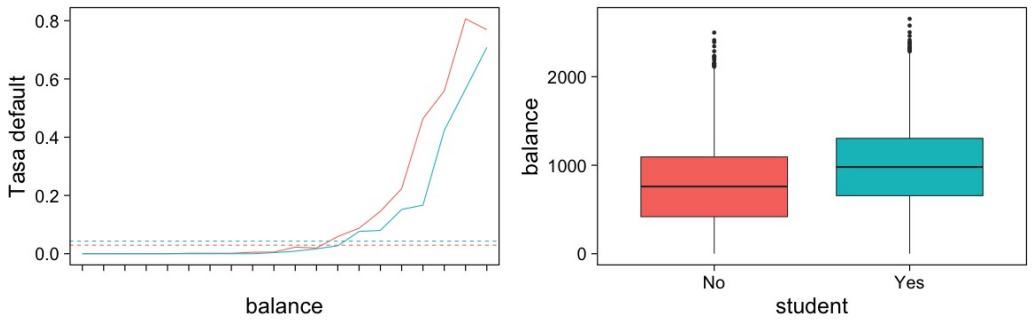


Figura 7. *Aparente paradoja para la tasa de* Default.

3. CLASIFICACIÓN PARA MAS DE DOS CLASES

Podemos extender a un problema `multi-clase`

$$\mathbb{P}(Y = k|X) = \frac{e^{\beta_{0,k}+\beta_{1,k}X_1+\cdots+\beta_{p,k}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0,\ell}+\beta_{1,\ell}X_1+\cdots+\beta_{p,\ell}X_p}} \tag{12}$$

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

$$\mathbb{P}(Y = k|X = x) = \frac{\mathbb{P}(X = x|Y = k) \cdot \mathbb{P}(Y = k)}{\mathbb{P}(X = x)}. \tag{13}$$

En el contexto de análisis discriminante utilizamos

$$\mathbb{P}(Y = k|X = x) = \frac{\pi_k \, f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell \, f_\ell(x)}, \tag{14}$$

donde

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

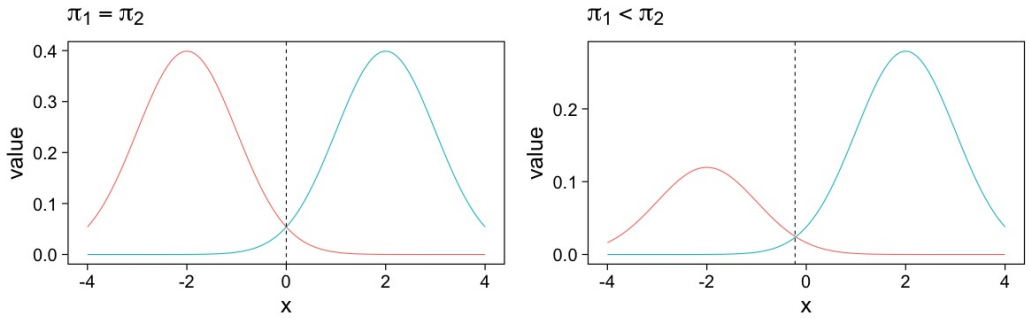


FIGURA 8. *Analisis discriminante con densidades Gaussianas.*

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

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma_2} + \log(\pi_k)\,. \tag{15}$$

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

$$x = \frac{\mu_1 + \mu_2}{2}\,. \tag{16}$$

4.4. ¿Y en la vida real?

Estimamos los parámetros con los criterios usuales.

Los parámetros que se ajustarán serán: $\pi_k, \mu_k, \sigma_k, \sigma$.

4.5. LDA con $p > 1$.

La función discriminante es

$$\delta_k(x) = x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log(\pi_k)\,. \tag{17}$$

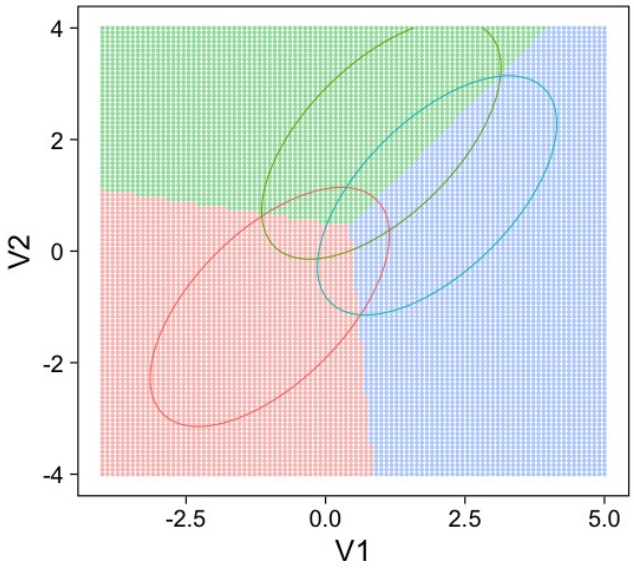


FIGURA 9. *LDA en dos dimensiones.*

4.6. Predicciones

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

$$\hat{\mathbb{P}}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{\ell=1}^{K} e^{\hat{\delta}_\ell(x)}}\,. \tag{18}$$

5. LDA EN DATOS

```
data ← Default
data ← data ▷ as_tibble() ▷
  mutate(default = ifelse(default == "Yes", "positive", "negative"))
data ▷ head()
```

```
# A tibble: 6 × 4
  default  student balance income
  <chr>    <fct>     <dbl>  <dbl>
1 negative No         730. 44362.
2 negative Yes        817. 12106.
3 negative No        1074. 31767.
4 negative No         529. 35704.
5 negative No         786. 38463.
6 negative Yes        920.  7492.
```

```
lda.model ← MASS::lda(default ~ balance, data)
```

Listing 6. *Modelo ajustado para los datos de crédito de estudiantes.*

6. EVALUACIÓN DE MODELOS

Con nuestro modelo entrenado podemos comparar lo que predecimos contra lo que realmente sucede.

```
library(yardstick)
data ← data ▷
  as_tibble() ▷
  mutate(predicted = predict(lda.model)$class,
         probability = predict(lda.model)$posterior[,1])
data ▷
  conf_mat(truth = default, estimate = predicted)
```

```
          Truth
Prediction positive negative
  positive       76       24
  negative      257     9643
```

Listing 7. *Comparación predicciones contra etiquetas verdaderas.*

Lo cual es una realización de un concepto que podemos establecer por medio de Fig. 10.

La capacidad predictiva usualmente está medida en términos de la exactitud (*acurracy*). Esto es la tasa de aciertos.

```
data ▷
  accuracy(truth = default, estimate = predicted)
```

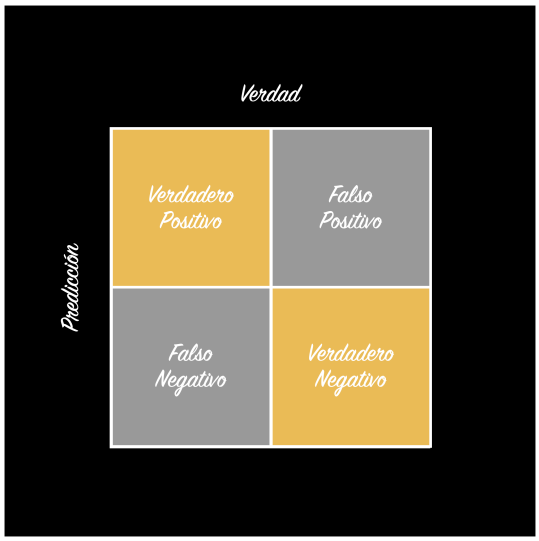


FIGURA 10. *Matriz de confusión.*

```
# A tibble: 1 × 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.972
```

LISTING 8. *Exactitud del modelo.*

Con esta exactitud, la `tasa de errores de clasificación` es: $(24+257)/10,000 \approx 0.028$ (complemento).

La exactitud la podemos ilustrar en el contexto de nuestros datos como se muestra en Fig. 11.

FIGURA 11. *Matriz de confusión con predicciones de* LDA.

Esto nos muestra una moraleja: En problemas donde tenemos la misma proporción para cada clase (`datos balanceados`) es una buena métrica de ajuste. Pero en caso donde no (`datos desbalanceados`) entonces puede ser una métrica engañosa.

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

Por ejemplo, la capacidad de acertar en la identificación para los casos `positivos` es:

```
data ▷
  recall(truth = default, estimate = predicted)
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 recall  binary         0.228
```

A esta métrica le llamamos **exhaustividad** (*recall*) y expresamos de manera esquemática en Fig. 12.

FIGURA 12. *Matriz de confusión con predicciones de* LDA.

> El término de exhaustividad se interpreta como qué tan hábil es una entidad en recuperar los elementos importantes o de interés dentro de una población.

La proporción de aciertos para la clase positiva es lo que denominamos **precisión** (*precision*) y para nuestro modelo tenemos

```
data ▷
  precision(truth = default, estimate = predicted)
```

```
# A tibble: 1 × 3
  .metric   .estimator .estimate
  <chr>     <chr>          <dbl>
1 precision binary          0.76
```

Cuya representación esquemática se muestra en Fig. 13.
Por otro lado, la capacidad de identificación de para la clase **negativa** es:

```
data ▷
  recall(truth = default, estimate = predicted, event_level = 'second')
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 recall  binary         0.998
```

Esta métrica también tiene un nombre particular y es el de **specificidad**.

FIGURA 13. *Matriz de confusión con predicciones de* LDA.

```
data ▷
  spec(truth = default, estimate = predicted)
```

También podemos fijarnos en la proporción de aciertos para la clase negativa es

```
data ▷
  precision(truth = default, estimate = predicted, event_level = 'second')
```

```
# A tibble: 1 × 3
  .metric   .estimator .estimate
  <chr>     <chr>          <dbl>
1 precision binary         0.974
```

La pregunta natural es: ¿en qué métrica nos fijamos? La respuesta es: tenemos que hacer un compromiso.

Nos puede interesar ambas, y podemos construir métrica que las combine:

```
data ▷
  f_meas(truth = default, estimate = predicted)
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 f_meas  binary         0.351
```

La métrica F es un compromiso entre las dos métricas que hemos visto anteriormente. Es decir, el *recall* (la tasa con la que podemos identificar los objetos que buscamos) y la **precisión** (la tasa con la que correctamente identificamos ambos casos). De tal manera, que la métrica F (el caso particular para F_1) es

$$F = 2 \cdot \frac{\text{precision} \quad \text{recall}}{\text{precision} + \text{recall}}. \tag{19}$$

6.1. Sensibilidad al punto de corte

Las métricas anteriores esconden una peculiaridad: se esconde detrás una selección para tomar una decisión.

```
data ← data ▷
  mutate(predicted.score = factor(
           ifelse(probability ≥ .2,"positive", "negative"),
           levels = c("positive", "negative")
         ))
```

```
# A tibble: 1 × 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.963
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 recall  binary         0.586
```

```
# A tibble: 1 × 3
  .metric   .estimator .estimate
  <chr>     <chr>          <dbl>
1 precision binary         0.452
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 f_meas  binary         0.510
```

Entonces, lo que necesitaríamos es tener una visión global para cada punto de corte que podamos elegir. Para esto podemos construir el espacio ROC (*receiver operating characteristic*) de clasificadores posibles con el modelo entrenado.

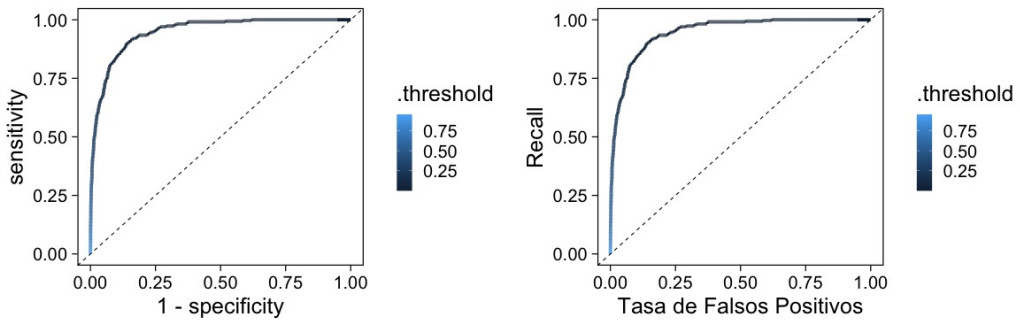


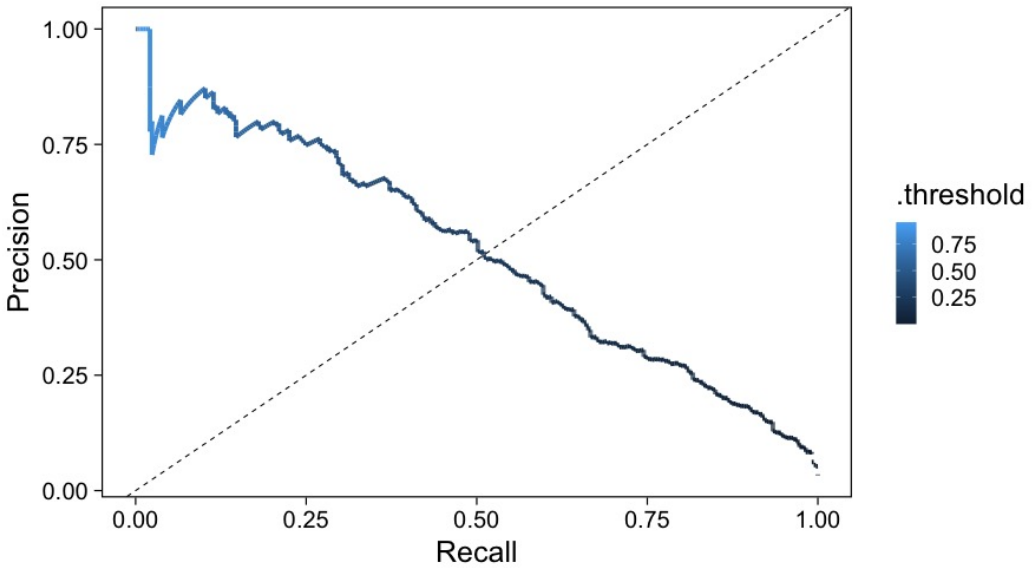
FIGURA 14. *Gráfico ROC (*Receiver Characteristic Curve*).*

Esta curva la podemos resumir por medio del área bajo la curva ROC

```
data ▷
  roc_auc(default, probability)
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.948
```

De la misma manera podemos mostrar también gráfico con todas las posibles combinaciones de `recall` y `precision` dados los diferentes puntos de corte que podemos realizar.



6.2. Post-procesando las probabilidades

La predicción de probabilidad de clase (por ejemplo, $\hat{p}_1(x)$) en general no puede entenderse como una probabilidad. Lo podemos interpretar cómo un *score* de pertenencia a la clase 1. Nos encantaría poder interpretar dicha predicción como una probabilidad. En el sentido frecuentista nos encantaría buscar que la frecuencia relativa de la categoría 1 dentro de los individuos con las características x sea cercana a $\hat{p}_1(x)$. Esto lo podemos estudiar a través de un `gráfico de calibración` de probabilidades, donde evaluamos la *cobertura* de dichos *scores* (ver mas en Capítulo 11 de [2]).

En la práctica es importante contextualizar los costos de una mala clasificación. Por ejemplo, el costo de no identificar a los clientes que te van a dejar de pagar un crédito, o los pacientes que no necesitan un tratamiento médico. La curva *lift* nos ayuda a contextualizar esto y en consecuencia buscar un punto de corte apropiado para el problema de predicción de clases. Si pensamos en que estudiaremos con mayor cuidado las predicciones mas seguras, querríamos que nuestro modelo sea capaz de *encontrar* a los individuos de interés con tan sólo ordenarlos por esas *probabilidades*. Puedes consultar mas de esto en el Capítulo 11 de [2].

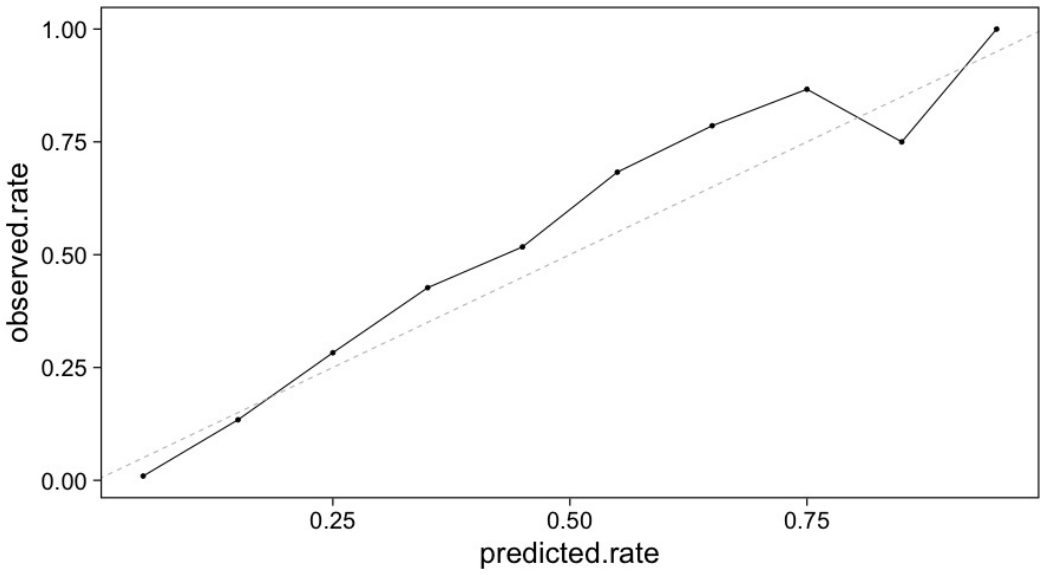
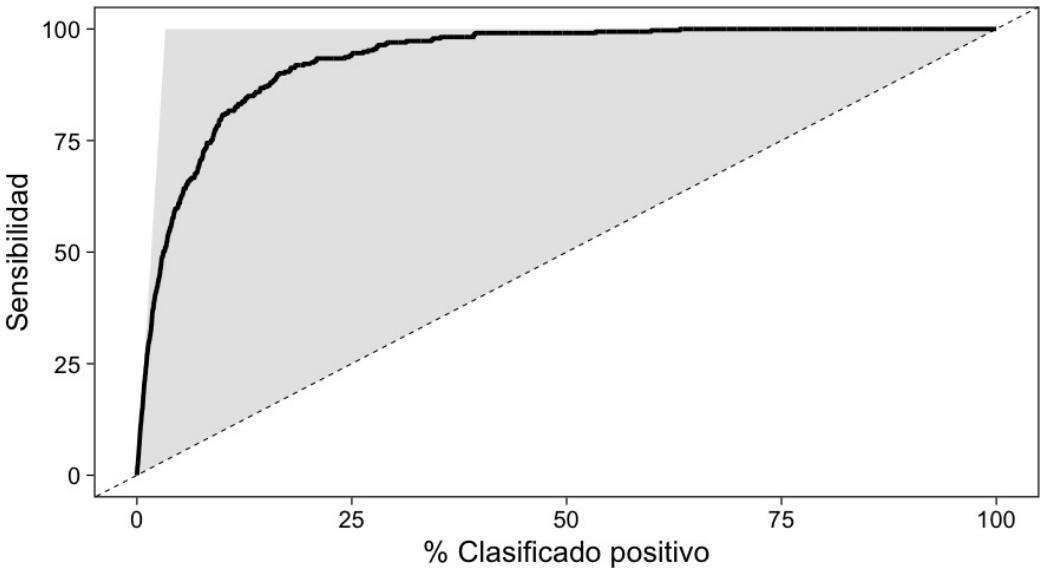
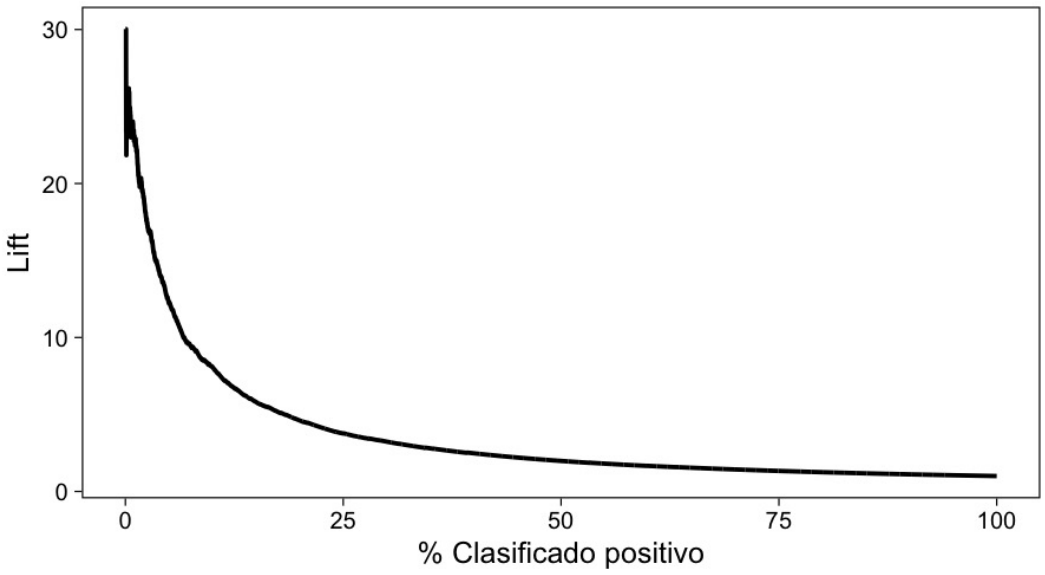


FIGURA 15. *Gráfico de calibración de probabilidades.*



Este gráfico nos ayuda a identificar con qué facilidad estamos encontrando los casos positivos si siguiéramos lo que el modelo nos indica de acuerdo a los *scores*. Esto puede determinar una estrategia de selección o intervención basada en un modelo predictivo.

Otra alternativa es graficar en el eje vertical la sensibilidad entre el % de clasificados como positivos.



En este último gráfico podemos identificar aquellos casos donde el *lift* es mayor a uno, cuando el modelo es superior a escoger los casos al azar.

7. OTROS MODELOS 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 que *dentro de cada clase* las *características son independientes* tenemos el `clasificador Bayesiano ingenuo`.
- Hay muchos mas que se pueden explorar considerando estimadores no-paramétricos.

7.1. Análisis discriminante cuadrático

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

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log \pi_k - \frac{1}{2}\log|\Sigma_k|\,. \tag{20}$$

7.2. Clasificador ingenuo Bayesiano

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

$$\delta_k(x) \propto \log\left(\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right) = -\frac{1}{2}\sum_{j=1}^{p}\left(\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log\sigma_{kj}^2\right) + \log\pi_k\,. \tag{21}$$

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

8. 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 entre un modelo logístico y un LDA son muy similares.

9. RESUMEN

- Regresión logistica 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.

10. OTROS MODELOS ÚTILES

- Modelos lineales generalizados.
- Vecinos más cercanos.

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. 1

[2] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-6848-6 978-1-4614-6849-3. . 1, 15