Tema 3 - Clasificación

Aprendizaje Estadístico

Lorena Romero Mateo - 77857300T

Email: lorena.romerom@um.es

Índice

1. Clasificación

- Las variables cualitativas toman valores en un conjunto desordenado C, como por ejemplo:
  - color de ojos: {marrón, azul, verde}
  - correo electrónico: {spam, ham}
- Dado un vector de características X y una respuesta cualitativa Y que toma valores en el conjunto C, la tarea de clasificación es construir una función C(X) que tome como entrada el vector de características X y prediga su valor para Y; es decir, $C(X) \in C$.
- A menudo estamos más interesados en estimar las probabilidades de que X pertenezca a cada categoría en C.
- Por ejemplo, es más valioso tener una estimación de la probabilidad de que una reclamación de seguro sea fraudulenta, que una clasificación de fraudulenta o no.

1.1. Ejemplo - Credit Card Default

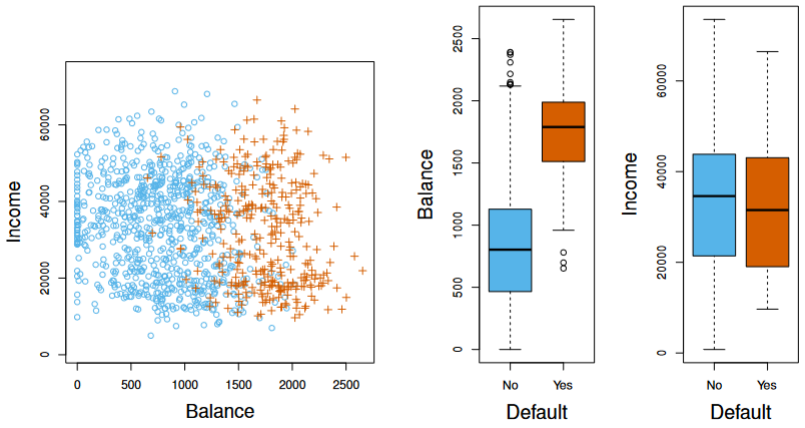


Figura 1: Credit Card Default

2. ¿Podemos usar regresión lineal?

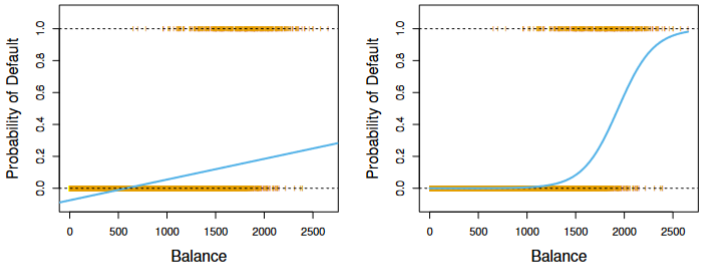
Supongamos para la tarea de clasificación de Default que codificamos

$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Sí} \end{cases}$$

¿Podemos simplemente realizar una regresión lineal de Y sobre X y clasificar como Sí si $\hat{Y} > 0{,}5$?

Recordamos que la recta de regresión tiende a la media, y va a tender que los valores bajen.

- En el caso de un resultado binario, la regresión lineal hace un buen trabajo como clasificador y es equivalente al **análisis discriminante lineal**, que discutiremos más adelante.
- Dado que en la población $E(Y|X = x) = \Pr(Y = 1|X = x)$, podríamos pensar que la regresión es perfecta para esta tarea.
- Sin embargo, la regresión lineal podría producir probabilidades menores que cero o mayores que uno. La regresión logística es más apropiada.



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Figura 2: Linear versus logistic regression

Ahora supongamos que tenemos una variable de respuesta con tres valores posibles. Un paciente se presenta en la sala de emergencias y debemos clasificarlo según sus síntomas.

$$Y = \begin{cases} 1 & \text{if es un derrame cerebral} \\ 2 & \text{if es una sobredosis de drogas} \\ 3 & \text{if es una convulsión epiléptica} \end{cases}$$

Esta codificación sugiere un orden y, de hecho, implica que la diferencia entre un derrame cerebral y una sobredosis de drogas es la misma que entre una sobredosis de drogas y una convulsión epiléptica.

La regresión lineal no es apropiada aquí.

La Regresión Logística Multiclase o el Análisis Discriminante son más apropiados.

3. Regresión Logística

Escribamos $p(X) = \Pr(Y = 1|X)$ para abreviar y consideremos usar el balance para predecir el `default`. La regresión logística utiliza la forma

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

Con Y = 0 ó 1 y un predictor. $p(X) = \Pr(Y = 1|X)$
La probabilidad estará entre 0 y 1, nunca llegando a ser igual.
la probabilidad de que Y sea 0 entonces es , $P(Y = 0|X = X)$

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

El número e es

$$e = \lim_{x\to\infty}(1 + \frac{1}{x})^x$$

(donde $e \approx 2{,}71828$ es una constante matemática conocida como el número de Euler.)
Es fácil ver que, sin importar los valores que tomen β_0, β_1 o X, $p(X)$ tendrá valores entre 0 y 1.
¿Diferencia entre logistic regression and linear regression?
En la figura **??** la regresión logística muestra que nuestra estimación para $p(X)$ es siempre entre 0 y 1.

3.1. Maximum Likelihood

Usamos la máxima verosimilitud para estimar los parámetros.

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

Esta verosimilitud da la probabilidad de los ceros y unos observados en los datos. Elegimos β_0 y β_1 para maximizar la verosimilitud de los datos observados.
La mayoría de los paquetes estadísticos pueden ajustar modelos de regresión logística lineal mediante máxima verosimilitud. En R usamos la función `glm`.

	Coeficiente	Error Estándar	Estadístico Z	Valor P
Intercepto	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Cuadro 1: Resultados del modelo de regresión logística

3.2. Estimación de Probabilidades

¿Cuál es nuestra probabilidad estimada de incumplimiento para alguien con un balance de $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10,6513+0,0055\times 1000}}{1+e^{-10,6513+0,0055\times 1000}} = 0,006$$

Con un balance de $2000:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10,6513+0,0055\times 2000}}{1+e^{-10,6513+0,0055\times 2000}} = 0,586$$

3.2.1. Usando el predictor `student`

Ahora consideremos usar el predictor `student` para predecir el `default`. Los resultados del modelo de regresión logística son los siguientes:

	Coeficiente	Error Estándar	Estadístico Z	Valor P
Intercepto	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Cuadro 2: Resultados del modelo de regresión logística usando `student` como predictor

La probabilidad estimada de incumplimiento para alguien que es estudiante (`student=Yes`) es:

$$\hat{Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3,5041+0,4049\times 1}}{1+e^{-3,5041+0,4049\times 1}} = 0,0431$$

Para alguien que no es estudiante (`student=No`):

$$\hat{Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3,5041+0,4049\times 0}}{1+e^{-3,5041+0,4049\times 0}} = 0,0292$$

3.3. Regresión logística con varias variables

Consideremos ahora un modelo de regresión logística con múltiples variables predictoras. La forma general del modelo es:

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

Los resultados del modelo de regresión logística con las variables `balance`, `income` y `student` son los siguientes:

	Coeficiente	Error Estándar	Estadístico Z	Valor P
Intercepto	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

¿Por qué el coeficiente para `student` es negativo, mientras que antes era positivo?

Podemos ver que entonces existen correlaciones entre la variable, podemos entender por el coeficiente negativo que si eres estudiante y teniendo en cuenta balance e income es menos probable que incumpla que si no lo es.

3.4. Confounding

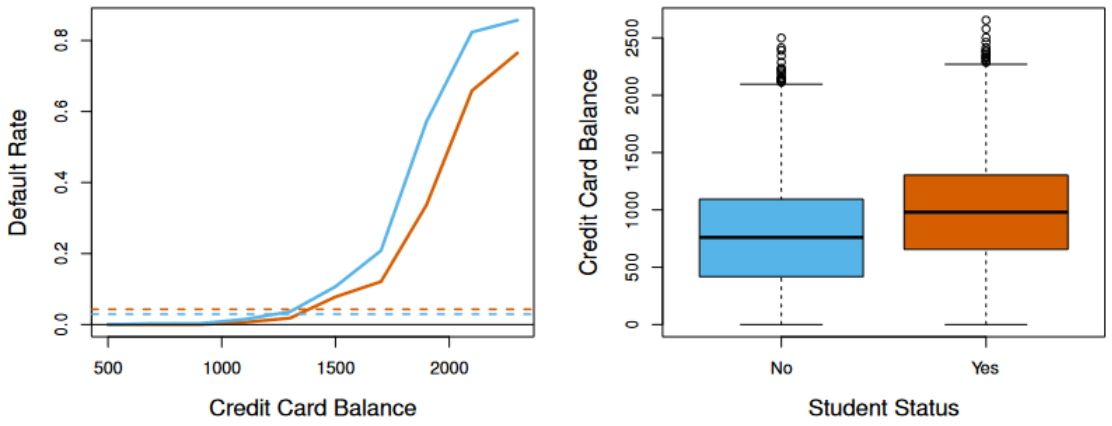


Figura 3: Confounding

Los estudiantes tienden a tener saldos más altos que los no estudiantes, por lo que su tasa de incumplimiento marginal es mayor que la de los no estudiantes.

- Pero para cada nivel de saldo, los estudiantes incumplen menos que los no estudiantes.
- La regresión logística múltiple puede desentrañar esto.

3.5. Regresión logística con más de dos clases

Hasta ahora hemos discutido la regresión logística con dos clases. Se generaliza fácilmente a más de dos clases. Una versión (utilizada en el paquete R glmnet) tiene la forma simétrica

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{pk}X_p}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_{1l}X_1+\cdots+\beta_{pl}X_p}}$$

Aquí hay una función lineal para cada clase. (Los estudiantes más matemáticos reconocerán que es posible cierta cancelación, y solo se necesitan $K-1$ funciones lineales como en la regresión logística de 2 clases.)

La regresión logística multiclase también se conoce como regresión **multinomial**.

4. Análisis Discriminante

Aquí el enfoque es modelar la distribución de X en cada una de las clases por separado, y luego usar el teorema de Bayes para invertir las cosas y obtener $\Pr(Y|X)$.

Cuando usamos distribuciones normales (Gaussianas) para cada clase, esto conduce al análisis discriminante lineal o cuadrático.

Sin embargo, este enfoque es bastante general, y también se pueden usar otras distribuciones. Nos centraremos en distribuciones normales.

4.1. Teorema de Bayes para clasificación

Thomas Bayes fue un matemático famoso cuyo nombre representa un gran subcampo del modelado estadístico y probabilístico. Aquí nos centramos en un resultado simple, conocido como el teorema de Bayes:

$$\Pr(Y=k|X=x) = \frac{\Pr(X=x|Y=k)\cdot\Pr(Y=k)}{\Pr(X=x)}$$

Este teorema nos permite invertir las probabilidades condicionales y es fundamental en muchos métodos de clasificación y estimación de probabilidades.

$$P(A\cap B) = P(B)P(A|B) = P(A)P(B|A)$$

Por lo tanto,

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

Escribimos esto de manera ligeramente diferente para el análisis discriminante:

$$\Pr(Y=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)},$$

donde

- $f_k(x) = \Pr(X=x|Y=k)$ es la densidad de X en la clase k. Aquí usaremos densidades normales para estas, por separado en cada clase. (dificil)
- $\pi_k = \Pr(Y=k)$ es la probabilidad marginal o a priori para la clase k. (fácil)

4.1.1. Condiciones LDA o Bayes

Si fijamos $Y = k$, la variable X tal que $Y = k$ se supone que sigue una distribución normal con media μ_k y varianza σ^2. Se supone que la varianza es la misma para todas las clases.

- Cuando las clases están bien separadas, las estimaciones de los parámetros para el modelo de regresión logística son sorprendentemente inestables. El análisis discriminante lineal no sufre de este problema.
- Si n es pequeño y la distribución de los predictores X es aproximadamente normal en cada una de las clases, el modelo de análisis discriminante lineal es nuevamente más estable que el modelo de regresión logística.
- El análisis discriminante lineal es popular cuando tenemos más de dos clases de respuesta, porque también proporciona vistas de baja dimensión de los datos.

4.1.2. Análisis Discriminante Lineal cuando p = 1

La densidad Gaussiana tiene la forma

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Aquí μ_k es la media, y σ_k^2 la varianza (en la clase k). Asumiremos que todas las $\sigma_k = \sigma$ son iguales.

Al sustituir esto en la fórmula de Bayes, obtenemos una expresión bastante compleja para $p_k(x) = \Pr(Y = k|X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

Afortunadamente, hay simplificaciones y cancelaciones.

4.1.3. Funciones discriminantes

Para clasificar en el valor $X = x$, necesitamos ver cuál de los $p_k(x)$ es mayor. Tomando logaritmos y descartando términos que no dependen de k, vemos que esto es equivalente a asignar x a la clase con la mayor puntuación discriminante:

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

Note que $\delta_k(x)$ es una función lineal de x.

Si hay $K = 2$ clases y $\pi_1 = \pi_2 = 0{,}5$, entonces se puede ver que el límite de decisión está en

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

(Vea si puede demostrar esto)

4.1.4. Estimando los parámetros

Para estimar los parámetros en el análisis discriminante lineal, utilizamos las siguientes fórmulas:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma}_k^2$$

donde

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

es la fórmula usual para la varianza estimada en la clase k.

4.1.5. Análisis Discriminante Lineal cuando p > 1

La densidad Gaussiana multivariante tiene la forma

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

La función discriminante es:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

A pesar de su forma compleja,

$$\delta_k(x) = c_{k0} + c_{k1} x_1 + c_{k2} x_2 + \cdots + c_{kp} x_p$$

es una función lineal.

5. Tipos de error

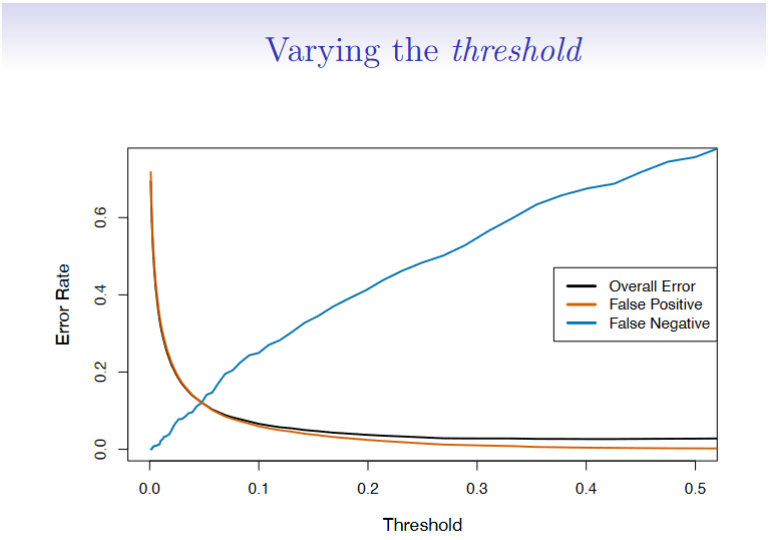


Figura 4: Tipos de error

Tasa de falsos positivos: La fracción de ejemplos negativos que son clasificados como positivos — 0.2 % en el ejemplo.

Tasa de falsos negativos: La fracción de ejemplos positivos que son clasificados como negativos — 75.7 % en el ejemplo.

Producimos esta tabla clasificando a la clase Sí si

$$\Pr(\text{Default} = \text{Sí}|\text{Balance}, \text{Student}) > 0{,}5$$

Podemos cambiar las dos tasas de error cambiando el umbral de 0.5 a algún otro valor en $[0, 1]$:

$$\Pr(\text{Default} = \text{Sí}|\text{Balance}, \text{Student}) > \text{umbral}$$

y variar el umbral.