Modelos simples en Machine Learning

Fundamentos de regresión logísica

German Rosati
german.rosati@gmail.com

UNTREF - UNSAM - CONICET

5 de octubre de 2018

Repaso...

Problemas de Clasificación

- Y es una variable cualitativa.
- Toma valores sin orden en un conjunto C
- Dado un vector de features X y una variable target cualitativa Y, un problema de clasificación busca construir una función $f(X)$ que prediga los valores de Y, es decir, $f(X) \in C$
- Muchas veces estamos interesados en un predecir *probabilidades*

Regresión Logística

Fundamentos - Ejemplo

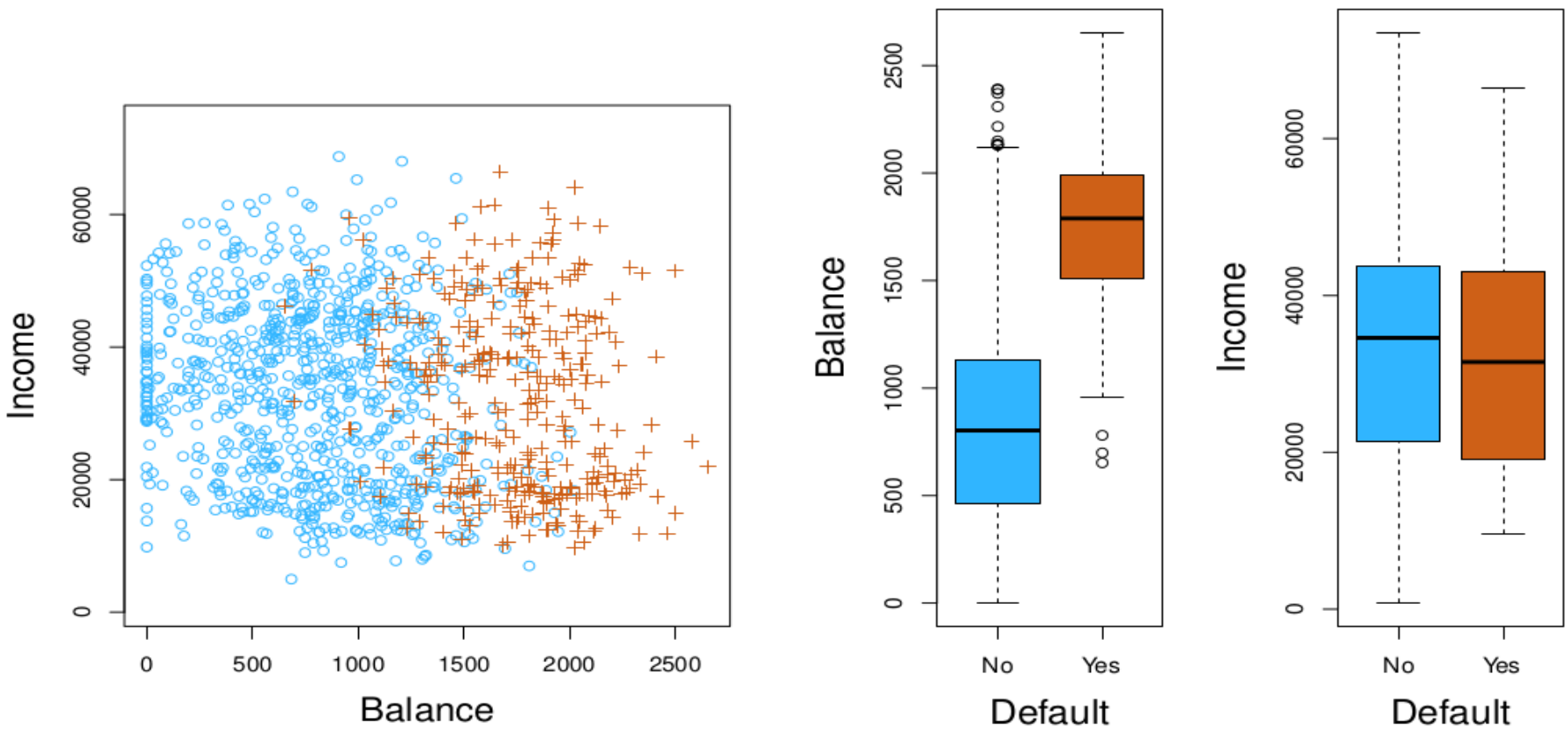

Figura: Scatterplot y BoxPlot Default dataset [1]

Regresión Logística

¿Por qué no usar una regresión lineal?

- Podríamos codificar Y de la siguiente forma

$$Y = \begin{cases} 1 & \text{si no paga} \\ 0 & \text{si paga} \end{cases}$$

- ¿Podríamos realizar una regresión de Y en X y predecir que "SI" si $\hat{Y} \geq 0{,}5$?
- La regresión lineal produce probabilidades mayores a 1 o menores a cero. En ese sentido, una regresión logística es más apropiada.

Regresión Logística

Fundamentos - Ejemplo

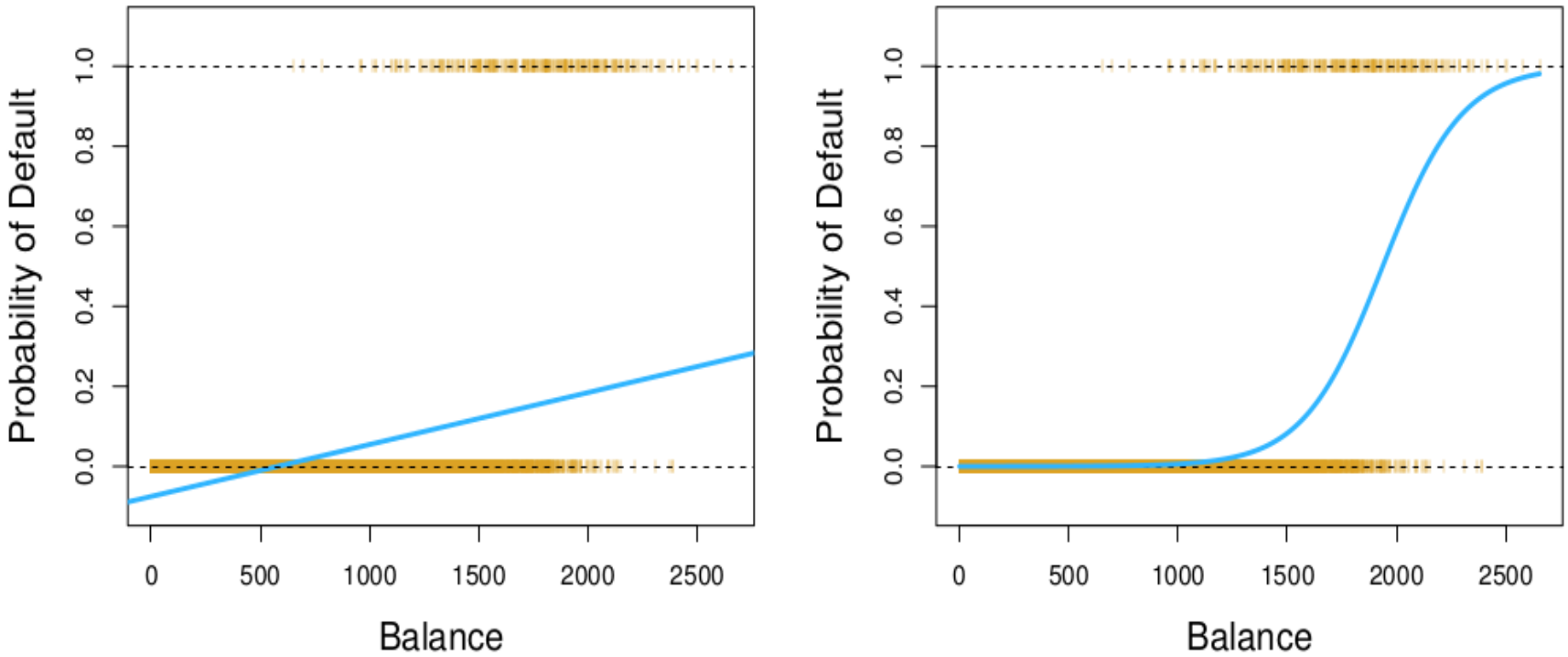

Figura: Regresión lineal y Logística en Default dataset [1]

Regresión Logística

Forma Funcional

- Escribamos la forma de una regresión logística.
- Queremos predecir la probabilidad de que una persona entre en default, condicionado a los valores de otras variables X, y hemos codificado este evento con 1 y 0 $P(Y = 1|X)$. Entonces,

$$P(Y = 1|X) = \frac{\exp^{\beta_0+\beta_1 X}}{1 + \exp^{\beta_0+\beta_1 X}} = \frac{1}{1 + \exp^{\beta_0+\beta_1 X}} \quad (1)$$

$$P(Y = 0|X) = 1 - P(Y = 1|X) \quad (2)$$

- Independientemente del valor que tomen β_0 y β_1, $P(Y = 1|X)$ nunca va a ser mayor a 1.
- Por convención y para simplificar usaremos $p(X) = P(Y = 1|X)$ y $1 - p(X) = P(Y = 0|X)$

- Tomando logaritmos y arreglando un poco la ecuación, tenemos que:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{3}$$

- Esta transformación se llama log odds o logit.

- La idea es entrenar un vector de parámetros (β_0, β_1) tal que el modelo estime $p(X)$ altas para valores positivos y bajas para valores negativos.
- Para un solo dato, la función de costo es

$$\ell(\beta_0, \beta_1) = \begin{cases} -\log(\hat{p}(x)) & \text{si } Y = 1 \\ \log(1 - \hat{p}(x)) & \text{si } Y = 0 \end{cases} \tag{4}$$

- Intuición: $-\log(t)$ crece mucho cuando t se acerca a 0.
- Costo grande si el modelo estima una probabilidad cercana a 0 para una instancia positiva, y también será muy grande si el modelo estima una probabilidad cercana a 1 para una instancia negativa

- $-\log(t)$ está cerca de 0 cuando t está cerca de 1, entonces, el costo será cercano a 0 si la probabilidad estimada es cercana a 0 para un negativo instancia o cerca de 1 para una instancia positiva.
- La función de costo en todo el conjunto de entrenamiento es simplemente el costo promedio de todo el entrenamiento.

$$\ell(\beta_0, \beta 1) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{p}(x)) + (1 - y_i)\log(1 - \hat{p}(x))\right] \quad (5)$$

- Buscamos los (β_0, β_1), que minimicen esa función.

- Fiteemos una regresión lineal con la variable *balance* -cuantitativa-

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.6513	0.3612	-29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

Figura: Regresión Logística en Default dataset [1]

- ¿Qué se puede decir de estos coeficientes?

- ¿Cuál sería nuestra estimación de la probabilidad de *default* para un *balance* de $1000?

$$\hat{p}(x) = \frac{exp^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + exp^{\hat{\beta}_0+\hat{\beta}_1 \hat{X}}} = \frac{exp^{-10{,}6513+0{,}0055\times 1000}}{1 + exp^{-10{,}6513+0{,}0055\times 1000}} = 0{,}006 \quad (6)$$

- ¿Ypara un *balance* de $2000?

$$\hat{p}(x) = \frac{exp^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + exp^{\hat{\beta}_0+\hat{\beta}_1 \hat{X}}} = \frac{exp^{-10{,}6513+0{,}0055\times 2000}}{1 + exp^{-10{,}6513+0{,}0055\times 2000}} = 0{,}586 \quad (7)$$

- Fiteemos una regresión lineal con la variable *student* -cualitativa-

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-3.5041	0.0707	-49.55	< 0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Figura: Regresión Logística en Default dataset [1]

- ¿Qué se puede decir de estos coeficientes?

- ¿Cuál sería nuestra estimación de la probabilidad de *default* para los estudiantes $student = 1$?

$$\hat{p}(default = 1|student = 1) = \frac{exp^{-3,5041+0,4049\times 1}}{1 + exp^{-3,5041+0,4049\times 1}} = 0,0431 \quad (8)$$

- ¿Y para los no estudiantes $student = 0$?

$$\hat{p}(default = 1|student = 1) = \frac{exp^{-3,5041+0,4049\times 0}}{1 + exp^{3,5041+0,4049\times 0}} = 0,0292 \quad (9)$$

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad (10)$$

$$P(Y=1|X) = \frac{exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}}{1+exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}} \quad (11)$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-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

Figura: Regresión Logística en Default dataset [1]

- Los estudiantes tienden a tener balances mayores que los no estudiantes. Por lo cual, las tasas de default çrudas"son mayores.
- Pero una vez que controlamos por el balance... entonces, los estudiantes defaultean menos.

Regresión Logística

Múltiples categorías en Y

- Cuando Y presenta más de dos valores se fitea una regresión para cada una de las k clases.

$$P(Y = k|X) = \frac{\exp^{\beta_{0k}+\beta_{1k}X_1+\beta_{2k}X_2+...+\beta_{pk}X_p}}{\sum_{k=1}^{K} \exp^{\beta_{0k}+\beta_{1k}X_1+\beta_{2k}X_2+...+\beta_{pk}X_p}} \quad (12)$$

- Es decir, se fitean $K-1$ regresiones logísticas binarias.

Regresión Logística

Ejercicio

Con el dataset data_filt.csv

1. Generar una variable Y que divida en
$$Y = \begin{cases} 1 & \text{si } p47t \geq 6990 \\ 0 & \text{si } p47t < 6990 \end{cases}$$
2. Generar features cualitativas de forma correcta
3. Entrenar una regresión logística para predecir
4. Evaluarla... (de forma correcta...)

- ¿Qué hacemos con el overfitting?
- Una forma es hacer *model selection*...
- Otra es utilizar técnicas de regularización.
- El objetivo es introducir una restricción en la función de costo (*RSS* - la función que se minimiza) y con eso forzar a los β a reducir su valor
- **Ridge:**

$$CF = \sum_{i=1}^{N}(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \beta_j^2 = RSS + \lambda \beta_j^2 \qquad (13)$$

- **LASSO:**

$$CF = \sum_{i=1}^{N}(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda |\beta_j| = RSS + \lambda |\beta_j| \qquad (14)$$

Modelo Lineal y Overfitting

Regularización

- Se parte de la minimización de *RSS* habitual + una restricción $\lambda|\sum_{j=1}^{p}\beta_j|$ que tiene el efecto de reducir los coeficientes β_j estimados
- λ es un hiperparámetro del modelo que controla el impacto de la penalización y se estima mediante *cross validation*
- Ridge se inventó originalmente para lidiar con el problema de la multicolinealidad. Sesga los coeficientes para reducir la varianza
- Ambos "encogen" los coeficientes hacia cero. LASSO, además, hace que algunos sean iguales a cero
- LASSO, entonces, realiza *model selection*
- Es una formalización de un proceso que muchas veces se hace artesanalmente

Referencias bibliográficas I

James, G., Witten, D., Hastie, T., and Tibshirani, R.
An Introduction to Statistical Learning – with Applications in R, vol. 103 of *Springer Texts in Statistics*.
Springer, New York, 2013.