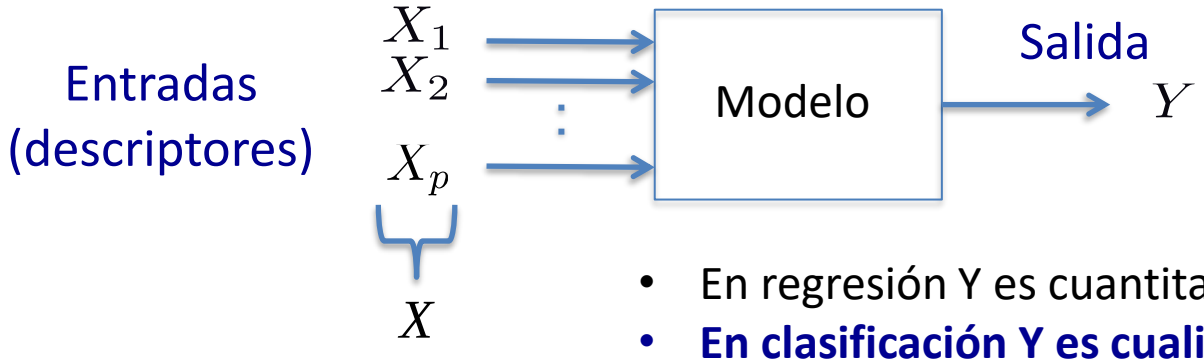
Clasificación y Regresión Logística

Santiago Alférez

edwin.alferez@urosario.edu.co

Diciembre 12 de 2020

Clasificación



- En regresión Y es cuantitativa
- **En clasificación Y es cualitativa (categórica)**

Muchas situaciones:

- Condición de paciente ∈ {sano, enfermo}
- Célula ∈ {*normal*, *drepanocito*, *esferocito*}
- ……

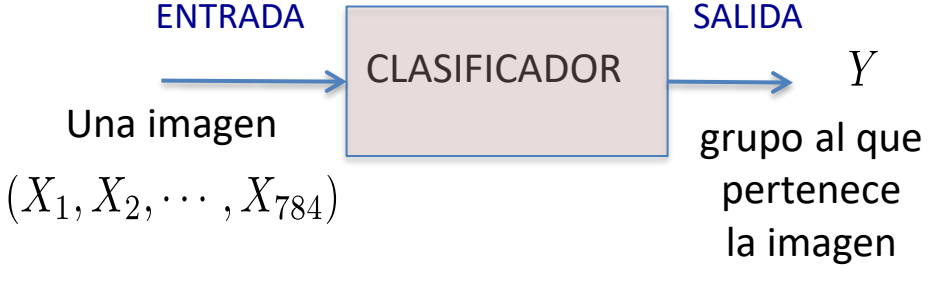
Dado un objeto caracterizado por un vector de descriptores, el **objetivo de un clasificador (modelo)** es predecir a qué clase pertenece el objeto dentro de un conjunto de clases predefinidas.

Clasificación - ejemplo

Base de datos MNIST: 70000 imágenes de los **dígitos (0,1,…., 9)** escritos a mano

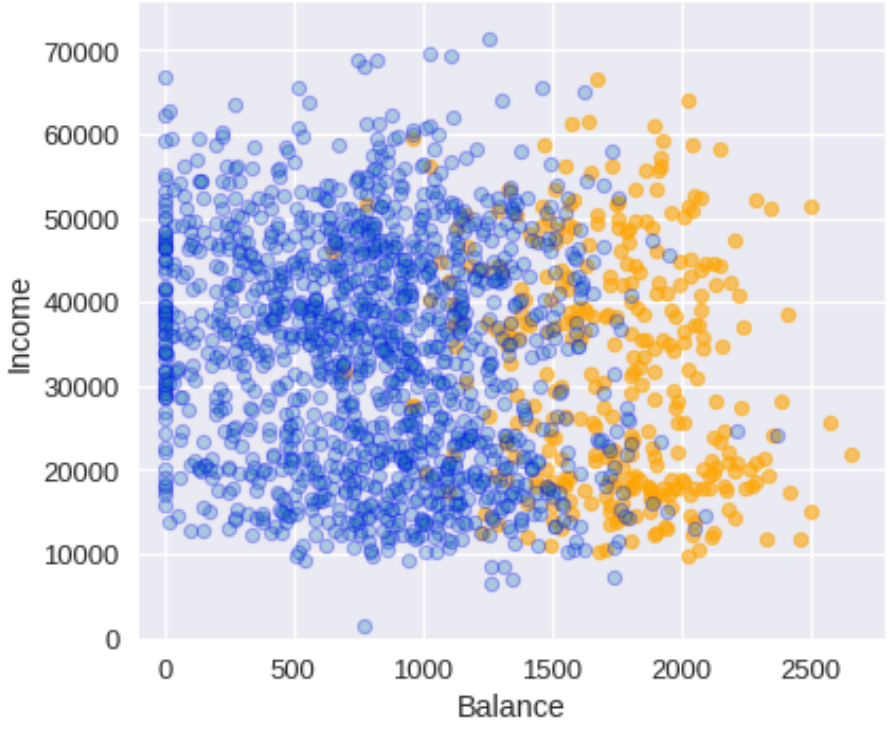
Cada imagen contiene un único dígito en escala de grises y está **etiquetada**, es decir se conoce cuál es el valor del dígito.

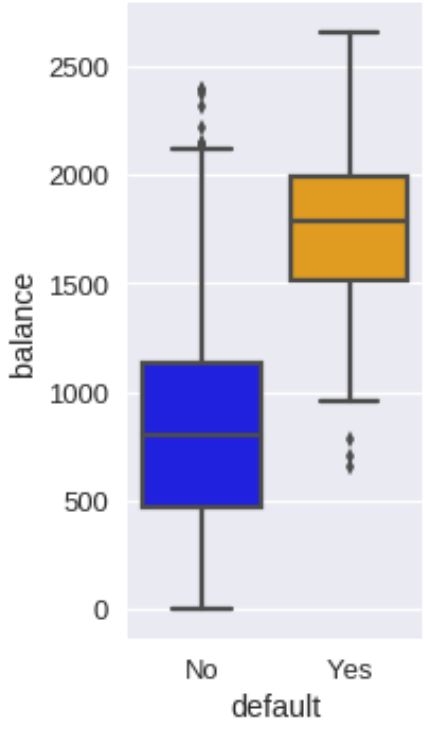
La imagen tiene 28 x 28 píxeles y está representada por **784 descriptores numéricos**, que son los valores de la intensidad de cada uno de los píxeles en un rango entre 0 y 255.

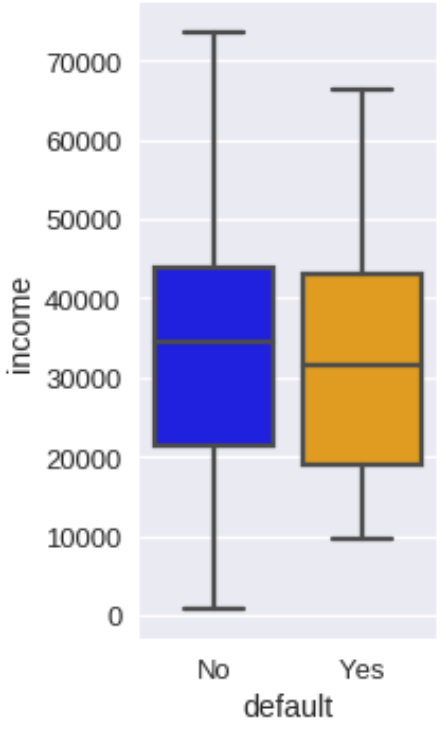


(Cuál es el dígito ?)

Ejemplo: Impago de la tarjeta de crédito







`seaborn.lmplot/plt.scatter`

`seaborn.boxplot`

X1 = Income: ingresos anuales de 10,000 individuos.

X2 = Balance: saldo (de la deuda) de la tarjeta de crédito.

Y= Default: impago de la tarjeta de crédito {Yes, No}

Ejemplo: Impago de la tarjeta de crédito

	default	student	balance	income
9993	No	No	938.836241	56633.448744
9994	No	Yes	172.412987	14955.941689
9995	No	No	711.555020	52992.378914
9996	No	No	757.962918	19660.721768
9997	No	No	845.411989	58636.156984
9998	No	No	1569.009053	36669.112365
9999	No	Yes	200.922183	16862.952321

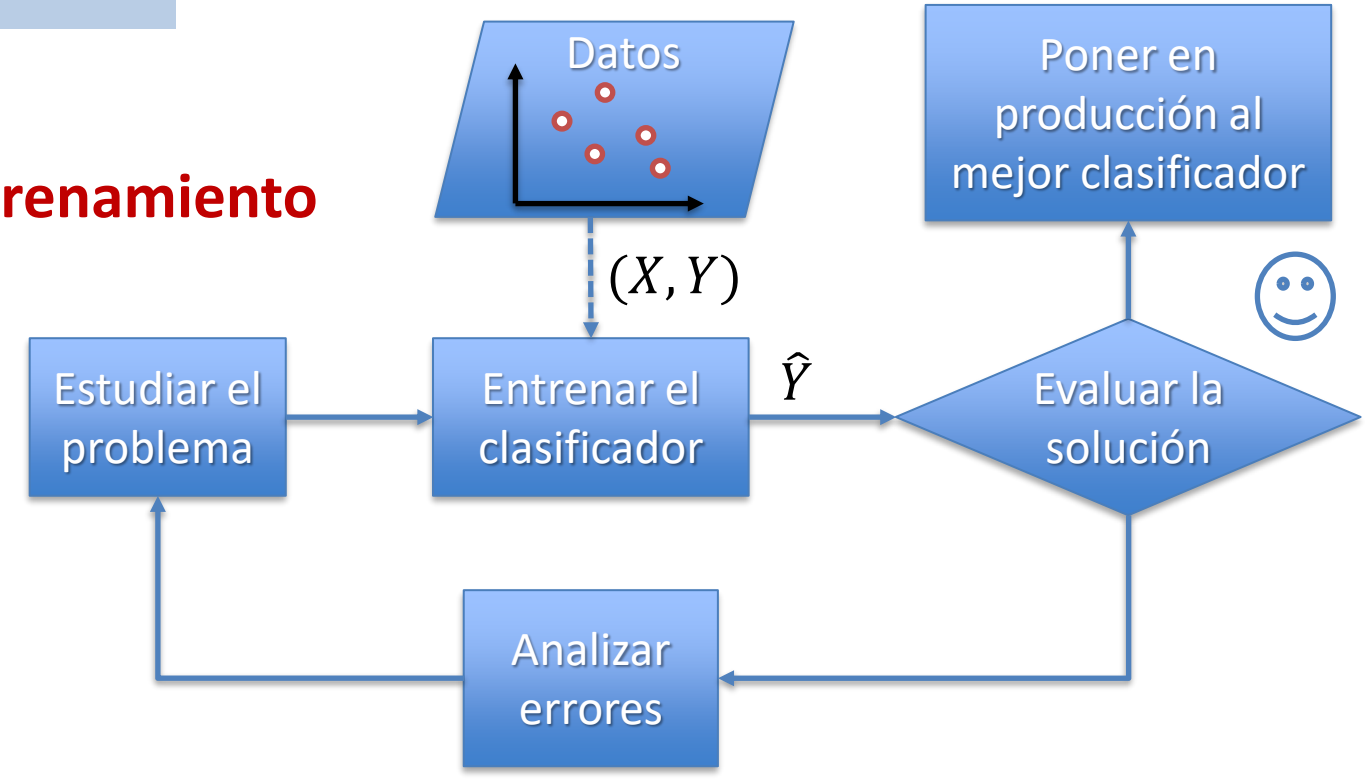
X1= Income: ingresos anuales de 10,000 individuos.
X2= Balance: saldo (de la deuda) de la tarjeta.
Y= Default: impago de la tarjeta de crédito {Yes, No}

```
1 df['default'].value_counts()
```

```
No      9667
Yes      333
Name: default, dtype: int64
```

Clasificación

Entrenamiento



Aplicación

Clasificación

Existen muchos tipos de clasificadores:

- Regresión logística
- Bayes
- Análisis discriminante lineal
- KNN
- Árboles de decisión, arboles aleatorios
- SVM
- Redes neuronales

…….. y otros

Cómo abordar la clasificación

Una forma natural para clasificar un objeto es mediante las **probabilidades** de que, dado un valor de los descriptores $(X = x)$, la salida pertenezca a una u otra de las categorías $(Y = k)$:

$$\Pr(Y = k|X = x)$$

Esto requiere estimar o modelizar la probabilidad condicional de la respuesta *Y* de cada clase dados los descriptores.

¿Cómo se hace ?

Clasificador ideal de Bayes

Supongamos que existen K clases $\{label_1, label_2, \dots, label_K\}$, y que se codificadas como $\{1, 2, \dots, K\}$. Sean

$$p_k(x) = \Pr(Y = k|X = x), \qquad k = 1,2 \dots, K.$$

las probabilidades condicionadas en x (particularmente, la probabilidad de que $Y = k$, dado el vector predictor x). Entonces, el clasificador óptimo Bayesiano en x, es definido por:

$$f(X = x) = j \; si \; p_j(x) = \max\{p_1(x), p_2(x), \dots, p_K(x)\}$$

Se elige la clase que tiene mayor probabilidad para el valor del descriptor calculado.

Puede demostrarse que este clasificador minimiza el error medio en los aciertos.

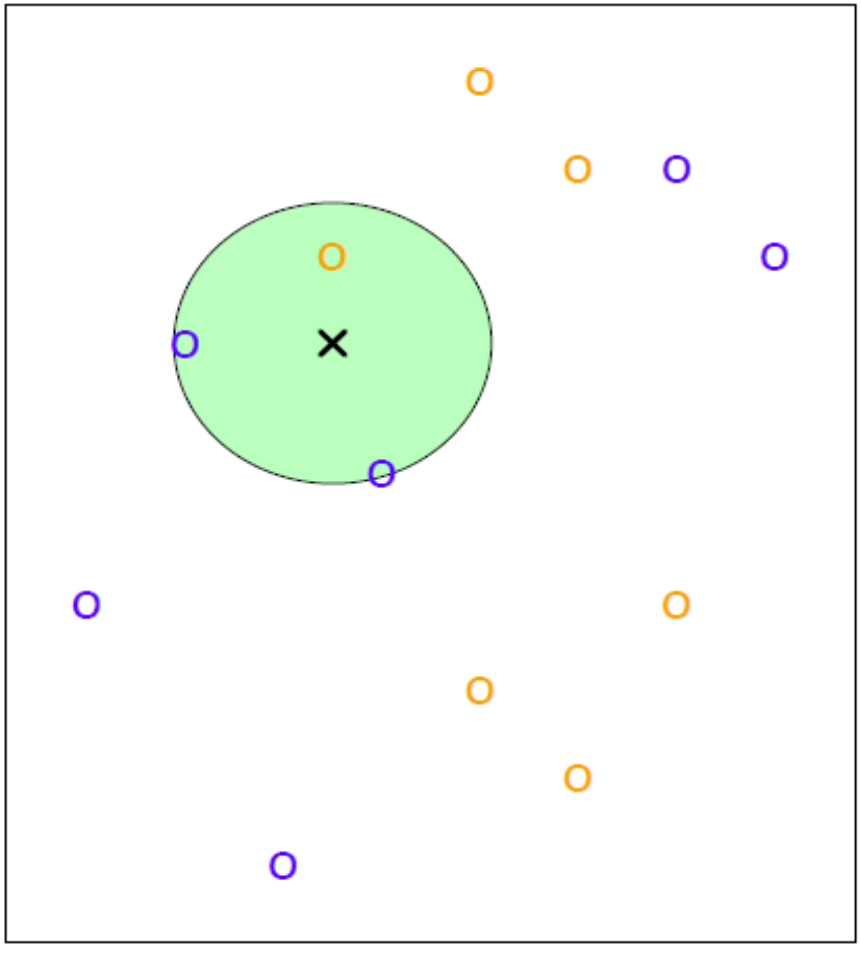
K-Nearest Neighbors (KNN)

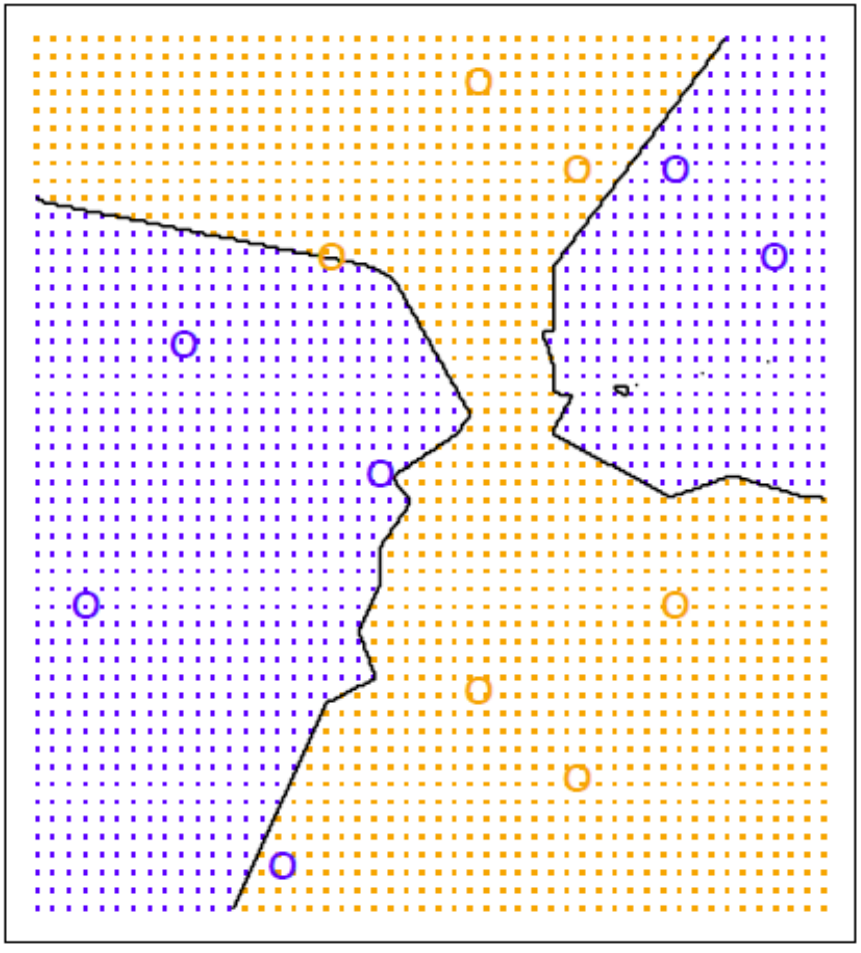
- El clasificador de Bayes es ideal (imposible de calcular) y sirve como un estándar de oro para comparar otros métodos.
- Un método para *estimar* la probabilidad condicional de *Y* dado *X*, es KNN.
- Dados un posible entero *K*, y una observación de *prueba* x_0:

$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

Estima la probabilidad condicional para la clase *j* como la fracción de puntos en la vecindad $\mathcal{N}_0$ que pertenecen a la clase *j*.

Ejemplo de KNN con k = 3





¿Se puede usar regresión lineal para clasificar?

Caso de la tarjeta de crédito

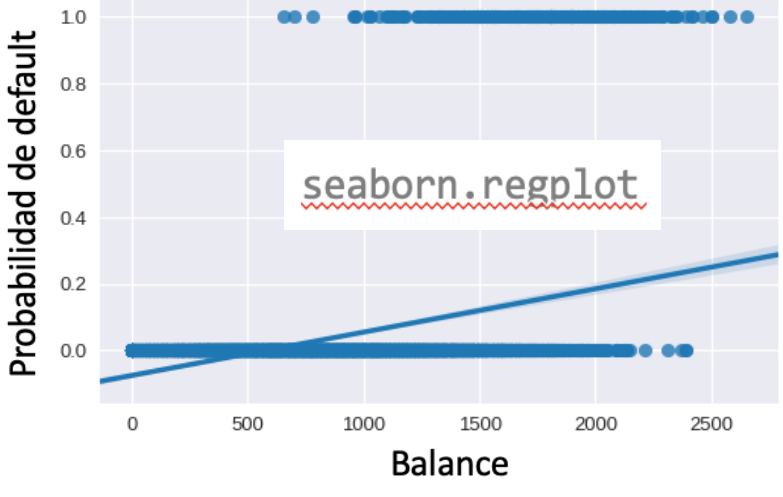
Supongamos que la variable de salida Default se codifica como:

$$Y = \begin{cases} 0 & si \ \ No \\ 1 & si \ \ Yes \end{cases}$$

Consideremos la probabilidad para la clase Default: $p(X) = \Pr(Y = 1|X)$

Un modelo lineal sería $p(X) = \theta_0 + \theta_1 X$

Ajuste mínimos cuadrados

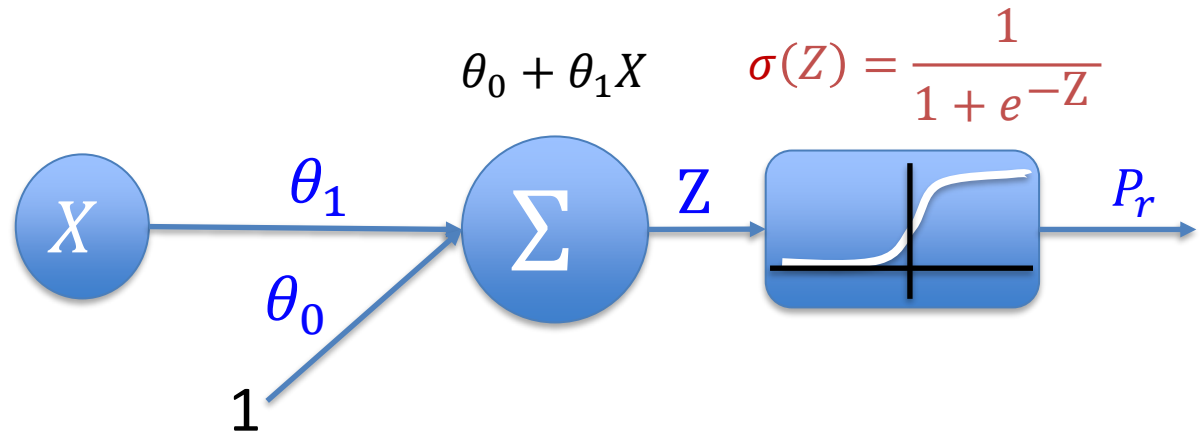


- Probabilidades negativas para balances pequeños
- Probabilidades > 1 para balances muy grandes

Idea inviable

Regresión logística

$$P_r(Y = 1|X) = p(X) = \frac{e^{\theta_0+\theta_1 X}}{1 + e^{\theta_0+\theta_1 X}}$$



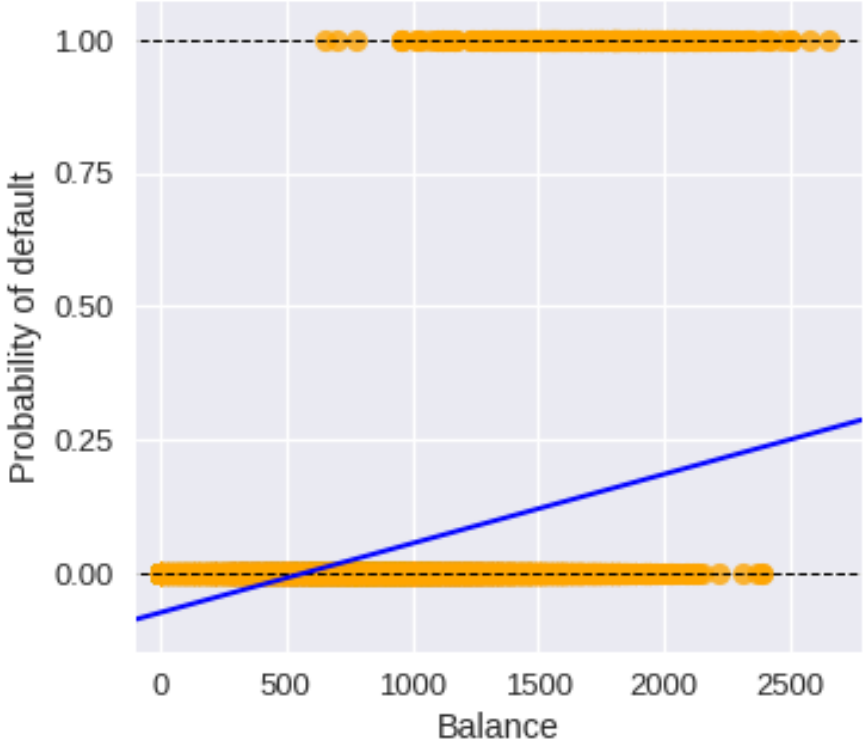
log-odds
logit

$$log\left(\frac{p(X)}{1-p(X)}\right) = \theta_0 + \theta_1 X$$

odss

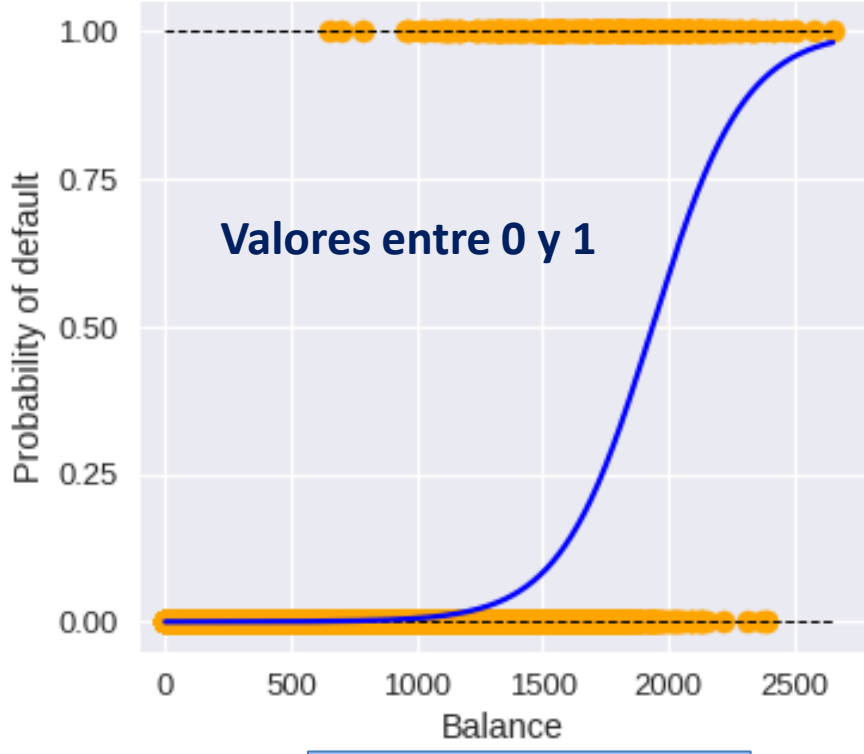
Regresión Lineal versus Regresión Logística

$$p(x) = P_r(Y = 1|X)$$



Regresión lineal

$$p(X) = \theta_0 + \theta_1 X$$



Regresión logística

$$p(X) = \frac{e^{\theta_0+\theta_1 X}}{1 + e^{\theta_0+\theta_1 X}}$$

Interpretación de β_1

$$\underbrace{\frac{p(X)}{1-p(X)}}_{\text{Odds}} = e^{\theta_0 + \theta_1 X} \qquad \underbrace{log\left(\frac{p(X)}{1-p(X)}\right)}_{\text{Logit}} = \theta_0 + \theta_1 X$$

- No es fácil interpretar lo que significa θ_1 con la regresión logística, porque estamos prediciendo $P(Y)$ y no *Y*.
- Si $\theta_1 = 0$, significa que no hay relación entre *Y* y *X*.
- Si $\theta_1 > 0$, significa que cuando X aumenta, también lo hace la probabilidad de que Y = 1.
- Si $\theta_1 < 0$, significa que cuando X aumenta, la probabilidad de que Y = 1 disminuye.
- Pero, ¿cuánto más grande o más pequeño? La probabilidad depende del valor particular de *X*.

Máxima verosimilitud (*likelihood*) para estimar parámetros

Se trata de **estimar** los parámetros (θ_0, θ_1)

tales que la probabilidad predicha para la clase Y (default) sea lo más próxima a 1 cuando el individuo observado tenga de verdad default y lo más próxima a 0 en caso contrario.

Con esta idea, definimos la **función de verosimilitud**

$$\mathcal{L}(\theta_0, \theta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

Esta verosimilitud da la probabilidad de los unos y ceros observados en los datos como sucesos independientes.

Máxima verosimilitud (*likelihood*) para estimar parámetros

$$\mathcal{L}(\theta_0, \theta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

- Se puede encontrar θ_0 y θ_1 **maximizando** la verosimilitud de los datos observados.
- Se convierte en un **problema de optimización** de una función objetivo.
- Existen diversos métodos iterativos de optimización, por ejemplo: el método de Newton, o el método del gradiente descendiente.

¿Cómo aplicar el gradiente descendiente?

$$\mathrm{J}(\boldsymbol{\theta}) = -\log L = -\sum_{i:y_i=1} \log p(x_i) - \sum_{i:y_i=1} \log(1 - p(x_i))$$

$$\mathrm{J}(\boldsymbol{\theta}) = -y_i \sum_i \log p(x_i) - (1 - y_i) \sum_i \log(1 - p(x_i))$$

$$\theta_j \leftarrow \theta_j - \eta \frac{\partial}{\partial \theta_j} \mathrm{J}(\boldsymbol{\theta})$$

Implementación de la regresión logística: sklearn

```
1  from sklearn.linear_model import LogisticRegression
2
3  model = LogisticRegression(solver='newton-cg')
4
5  X_train = df['balance'].values.reshape(-1,1)
6
7  y = df['default']
8
9  model.fit(X_train,y)
10
11 print("classes: {}\ncoefficientes: {}\nintercept: {}".format(
12     model.classes_,model.coef_, model.intercept_))
13
```

```
classes: ['No' 'Yes']
coefficientes: [[ 0.00549892]]
intercept: [-10.65133019]
```

Predicción con la regresión logística

¿Cuál es la estimación para la probabilidad de default (impago) para alguien con un balance de $1000?

$$\hat{p}(X) = \frac{e^{\hat{\theta}_0+\hat{\theta}_1 X}}{1+e^{\hat{\theta}_0+\hat{\theta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

```
1  model.predict_proba(1000)
```

```
array([[ 0.99424785,  0.00575215]])
```

No Yes

Predicción con la regresión logística

¿Y con un balance de $2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

```
1  model.predict_proba(2000)
```

```
array([[ 0.41423066,  0.58576934]])
```

No **Yes**

Descriptores cualitativos en la regresión logística

	default	student	balance	income
0	No	No	729.526495	44361.625074
1	No	Yes	817.180407	12106.134700
2	No	No	1073.549164	31767.138947

Descriptor cualitativo = student (Yes, No)

Para ajustar el modelo de regresión logística se introduce la variable:
1=student, 0=No student

```
1  results = smf.logit('default2 ~ student', data=df).fit()
2  results.summary2().tables[1]
```

		Coef.
θ_0	**Intercept**	-3.504128
θ_1	**student[T.Yes]**	0.404887

Descriptores cualitativos en la regresión logística

	Coef.
Intercept	-3.504128
student[T.Yes]	0.404887

Coeficiente positivo

Los estudiantes tienden a tener una mayor probabilidad de impago que los que no son estudiantes.

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Regresión logística con múltiples variables

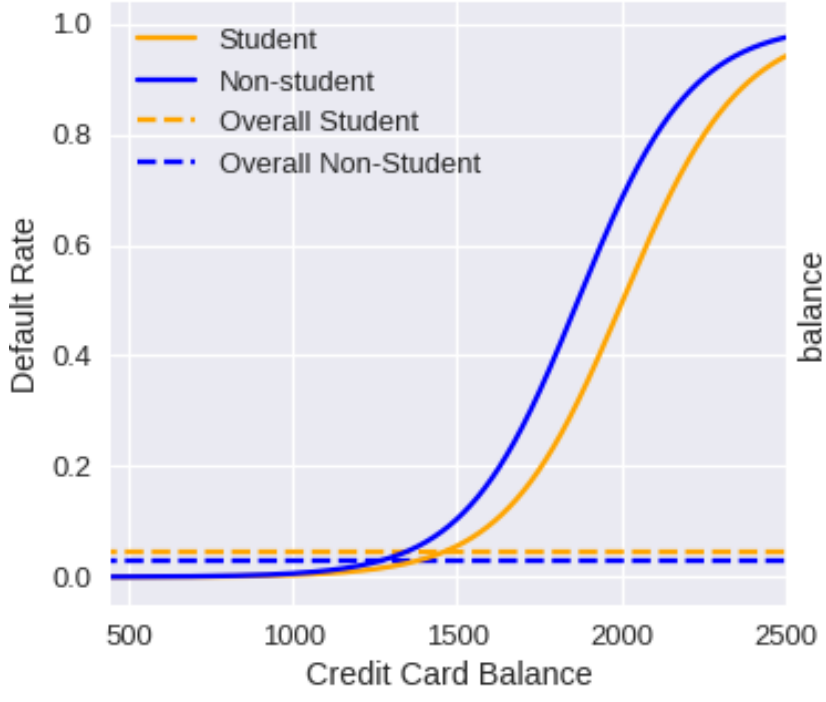
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$$

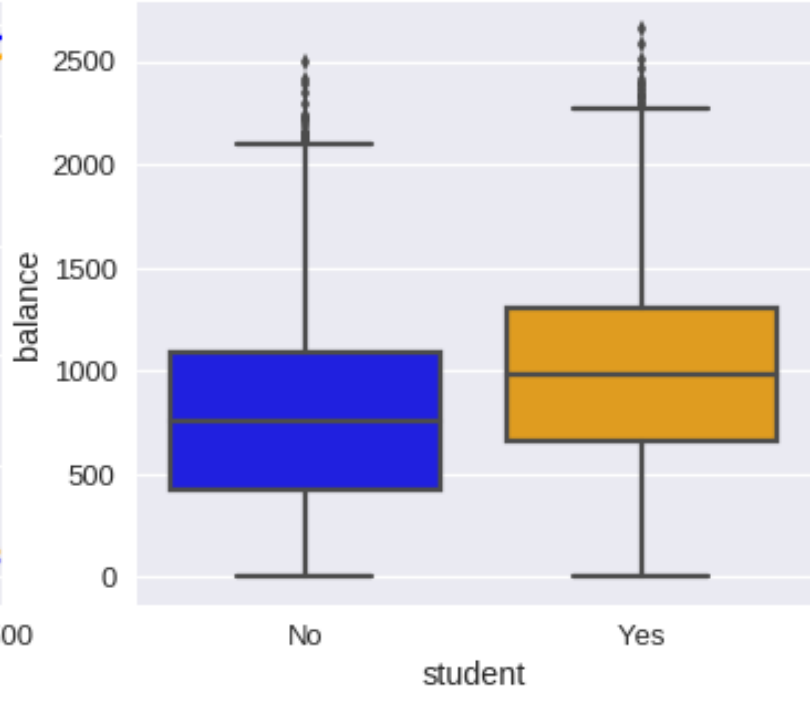
$$p(X) = \frac{1}{1+e^{-\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}}$$

	Coef.
Intercept	-10.869045
student[T.Yes]	-0.646776
balance	0.005737
income	0.000003

¿Por qué ahora el coeficiente para el estudiante es negativo, mientras antes fue positivo?

Confusión (confounding)





- Los estudiantes tienden a tener balances más altos que los no estudiantes, por lo que su tasa de incumplimiento (media) es más alta.
- Pero para un valor fijo del balance, los estudiantes incumplen menos.

Generalmente, una regresión o una clasificación con varios descriptores relevantes ofrece más riqueza que con un único descriptor.

Regresión logística con más de dos clases

En muchos casos, se necesita clasificar en varias (más de dos) clases.

El modelo de regresión logística se puede extender a varias clases, modelizando la probabilidad de cada clase (k) en la forma

Regresión multinomial

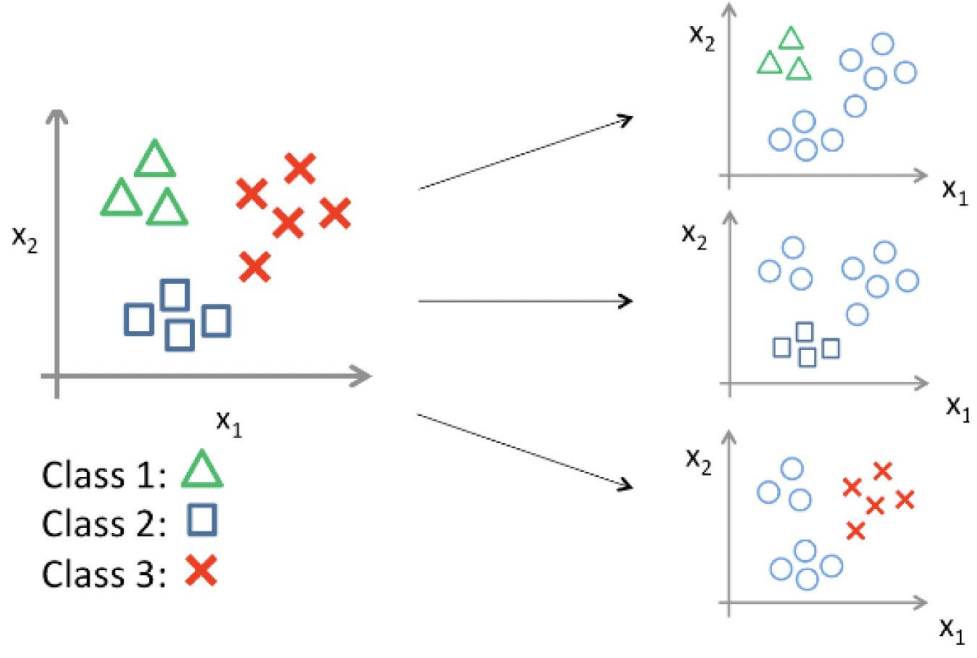
$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\cdots+\beta_{p\ell}X_p}}$$

Sklearn lo implementa mediante el parámetro:

`multi_class = "multinomial"`

Regresión logística con más de dos clases

Otra posibilidad es plantear una clasificación binaria para cada clase frente al resto de clases ***("one-versus-all")***



- Se escoge para cada observación la **probabilidad máxima**.
- Es fácilmente interpretable.
- Sklearn lo implementa mediante el parámetro:
  - ✓ `multi_class = "ovr"`

Medidas del rendimiento de un clasificador

Matriz de confusión

Ejemplo de clasificación de 60000 imágenes MIST

Clasificación

Clases Verdaderas	CLASIFICACIÓN NEGATIVA	CLASIFICACIÓN POSITIVA
CLASE NEGATIVA	Verdadero Negativo VN = 53272	Falso Positivo FP = 1307
CLASE POSITIVA	Falso Negativo FN = 1077	Verdadero Positivo VP = 4344

Medidas del rendimiento de un clasificador

Exactitud (*accuracy*)

Es la proporción de aciertos:

$$exactitud = \frac{VN + VP}{total\ de\ objetos} = \frac{53272 + 4344}{60000} = 0.96 = 96\%$$

Precisión (valor predictivo positivo)

Es la exactitud de las predicciones positivas. Es decir, la proporción de verdaderos positivos (aciertos) sobre el total de objetos que han sido clasificados como positivos, esto es

$$Precisión = \frac{VP}{VP + FP} = \frac{4344}{4344 + 1307} = 0.77 = 77\%$$

Sensibilidad (Proporción de verdaderos positivos - PVP) o Recall

Es la proporción de objetos positivos que son clasificados correctamente. Es decir, la proporción de aciertos dentro del conjunto total de objetos que son realmente positivos:

$$Sensibilidad = PVP = \frac{VP}{VP + FN} = \frac{4344}{4344 + 1077} = 0.79 = 79\%$$

Medidas del rendimiento de un clasificador

Especificidad (Proporción de verdaderos negativos - PVN)

Es la proporción de objetos negativos que son clasificados correctamente. Es decir, la proporción de aciertos dentro del conjunto total de objetos que son realmente negativos:

$$Especificidad = PVN = \frac{VN}{VN + FP} = \frac{53272}{53272 + 1307} = 0.98 = 98\%$$

Proporción de falsos positivos (PFP)

Es la proporción de objetos negativos que son clasificados incorrectamente, es decir como si fueran positivos. Este valor es complementario con la especificidad y se calcula en la forma

$$PFP = 1 - Especificidad = \frac{FP}{VN + FP} = \frac{1307}{53272 + 1307} = 0.02 = 2\%$$