Regresión

Data Mining

Ester Vidaña Vila

laSalle ENG
Universitat Ramon Llull

Regresión Lineal

Data Mining

Ester Vidaña Vila

laSalle ENG
Universitat Ramon Llull

Data science en el mundo real

- ¿Cómo lo hacen los **data scientist** para explotar los parámetros estadísticos en el mundo real?
- Se puede hacer:
  - **Regresión**
  - **Clasificación**
  - **Clustering**
- La idea es operar a través de un **modelo**
  - Presentación formal de una teoría
  - El modelo transforma los datos de entrada en respuestas: nos centraremos en **modelos lineales**
- Vamos a obsesionarnos con la **generalización**

Las navaja suiza de la estadística: Regresión

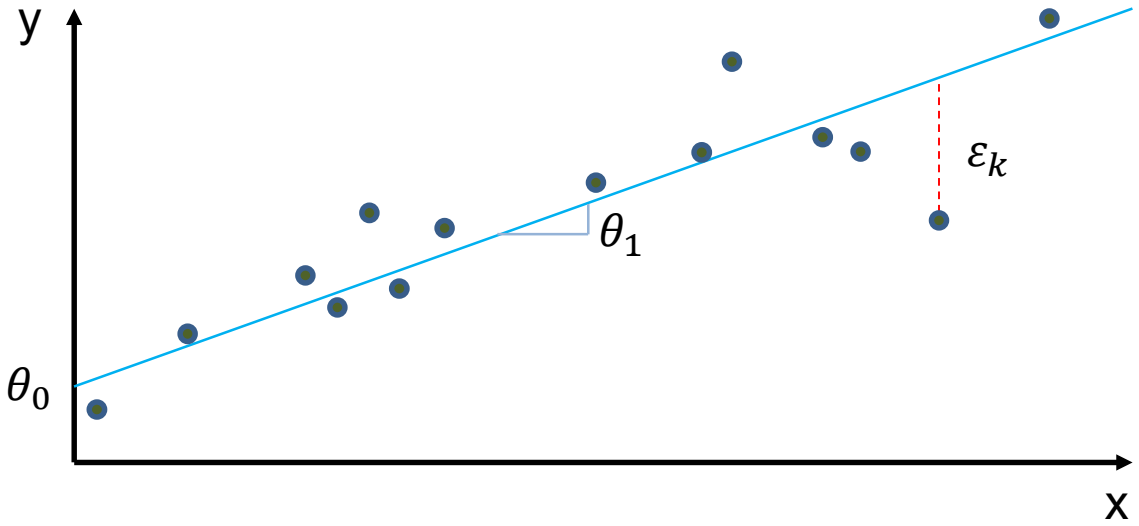
- Proceso para estimar la relación entre variables
- Tenemos
  - Una **variable dependiente** (o target) y
  - Una serie de **predictores** (o regresores, o variables independientes)
- En términos formales: $\vec{y} = f(x_1, x_2, \dots, x_n), \quad \vec{y} \in \mathbb{R}, \vec{x} \in \mathbb{R}^N$
- Nos ayuda a entender cómo el valor dependiente cambia cuando uno de sus predictores varía, mientras que el resto de predictores quedan fijos.

Regresión lineal (I)

- Un modelo de regression lineal consiste en una combinación lineal de parámetros

$$E[y|x_1, x_2, \ldots, x_n] = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n + \vec{\varepsilon}$$

donde θ_i son los parámetros que debemos estimar, x_j son los predictores y ε es el término de error



One dimension (that is: using a singe predictor x) linear regression model

Regresión lineal (II)

- Un modelo de regression lineal consiste en una combinación lineal de parámetros

$$E[y|x_1, x_2, \dots, x_n] = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n + \vec{\varepsilon}$$

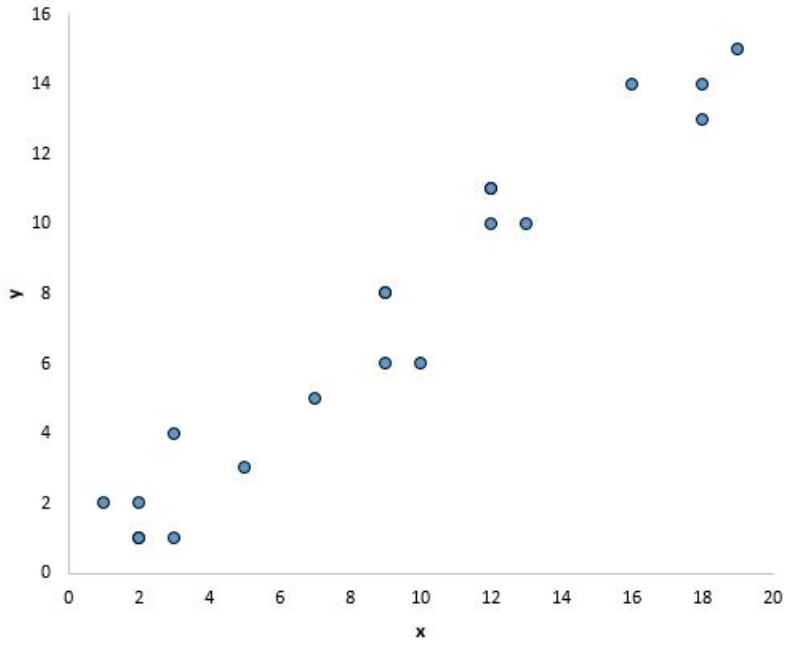
donde θ_i son los parámetros que debemos estimar, x_j son los predictores y ε es el término de error

- $\text{¿Cómo } encontramos \text{ los } parámetros\ \theta_i\text{?}$
  - Encontrar la línea que **minimice** el término de error ε
  - Ordinary least squares (*OLS*)

$$\vec{\theta} = (\vec{X}^T \cdot \vec{X})^{-1} \cdot \vec{X}^T \cdot \vec{Y}$$

Asunciones principal del modelo:

- Relación lineal: existe una relación lineal entre las variables independientes (x) y la variable dependiente (y).
  - Podemos comprobarlo con un scatterplot.

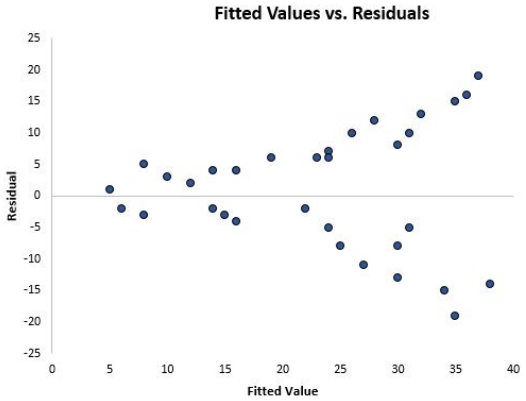


Asunciones principal del modelo:

- Relación lineal: existe una relación lineal entre las variables independientes (x) y la variable dependiente (y).
  - Podemos comprobarlo con un scatterplot.
- Colinealidad: Los predictores son linealmente independientes.

Asunciones principal del modelo:

- Relación lineal: existe una relación lineal entre las variables independientes (x) y la variable dependiente (y).
  - Podemos comprobarlo con un scatterplot.
- Colinealidad: Los predictores son linealmente independientes.
- Homoscedasticidad: La variancia del término de error es constante a través de las distintas observaciones.
  - Podemos hacer un plot que muestre el valor predicho vs. el término de error.



Asunciones principal del modelo:

- Relación lineal: existe una relación lineal entre las variables independientes (x) y la variable dependiente (y).
  - Podemos comprobarlo con un scatterplot.
- Colinealidad: Los predictores son linealmente independientes.
- Homoscedasticidad: La variancia del término de error es constante a través de las distintas observaciones.
  - Podemos hacer un plot que muestre el valor predicho vs. el término de error.
- Normalidad: El término de error tiene una **distribución normal**, con media cero y desviación estándar 1. Se puede comprobar haciendo un Q-Q plot.

Ver el futuro gracias a regresores lineales

- Estos modelos son extremadamente útiles ya que nos permiten hacer **predicciones:**
  - *Dada la energía que hemos consumido hoy, ¿cuánta consumiremos mañana?*
  - *¿Qué número de ventas esperamos para un producto determinado el siguiente trimestre fiscal?*
- ¡Estupendo! ¿Pero cuán fiables son nuestros modelos?
- Necesitamos una forma de estimar lo bien que hemos desarrollado nuestro modelo
- Hay diferentes formas de ver cuán bueno es un modelo.
- Por ejemplo, dos métodos comunes son:
  1. R^2
     - Nos dice cómo de bien se aproxima la **línea de regresión** a los valores de muestra
  2. Root mean squared error (RMSE)
     - Calcula la diferencia entre los valores reales y los valores predichos por el modelo

Ver el futuro gracias a regresores lineales

1. R^2
   - Nos dice cómo de bien se aproxima la **línea de regresión** a los valores de muestra.
   - Va de 0 a 1, métrica muy intuitiva.
   - Cuanto más próxima a 1 es la métrica, mejor.

0,00 ---> No existe correlación lineal alguna entre las variables
0,10 ---> Correlación positiva débil
0,50 ---> Correlación positiva media
0,75 ---> Correlación positiva considerable
0,90 ---> Correlación positiva muy fuerte
1,00 ---> Correlación positiva perfecta

2. Root mean squared error (RMSE)
   - Calcula la diferencia entre los valores reales y los valores predichos por el modelo.
   - Tiene las mismas unidades que la variable respuesta.
   - Cuanto más pequeño, más ajustado (o bueno) es el modelo.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Ejemplo: tamaño de casa vs. precio de casa

Price	Size
145000	1240
68000	370
115000	1130
…	…
127200	1340

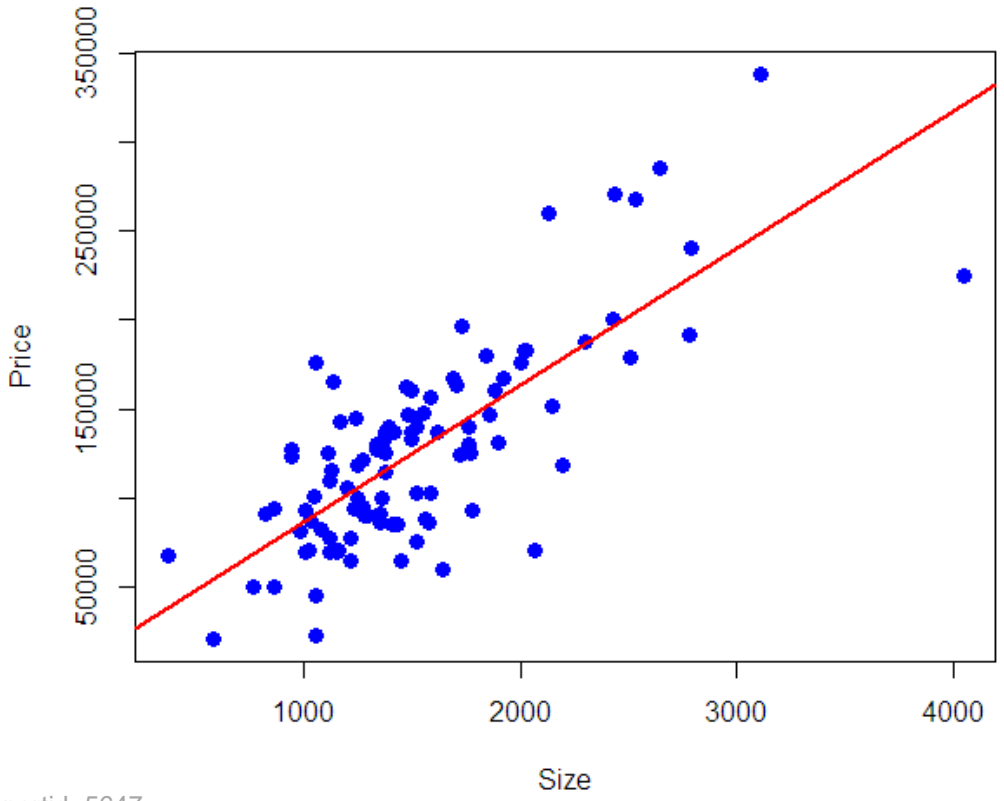
Model summary

Coefficients:

	Estimate
(Intercept)	9161.159
Size	77.008

R-squared: 0.5752
RMSE: 36361.05

$$Price = \theta_0 + \theta_1 \cdot Size + \vec{\varepsilon}$$



Data set courtesy of https://www.statcrunch.com/5.0/viewreport.php?reportid=5647

Regresión Logística

Data Mining

Ester Vidaña Vila

laSalle ENG
Universitat Ramon Llull

Clasificación a base de regresión

- Proceso general relacionado con la **clasificación**
  - Identificar a qué categoría de un set pertenecen nuevas observaciones a partir de los datos de entrada.
  - Es un **mapeo** de un conjunto de datos de ejemplo a un conjunto de categorías predeterminadas (*class labels*)
- Se puede ver como un caso especial de regresión
  - En este caso, la variable de salida es **discreta**

Clasificación con regresión logística

- Una respuesta **binaria** relacionada con una serie de predictores.
  - Así pues, la salida de la clase se traduce en dígitos binarios 0/1.
    - *NO_PASS_EXAM* / *PASS_EXAM*…
- La regresión logística estima la **probabilidad** de que una característica esté presente dados los valores del resto de variables.
  - Es lo que se llama un *threshold classifier*
- Formalmente:

$$\Pr(\vec{y}|\vec{x}) = \log\frac{\Pr(\vec{x})}{1 - \Pr(\vec{x})} = \theta_0 + \vec{\theta}\cdot\vec{x}$$

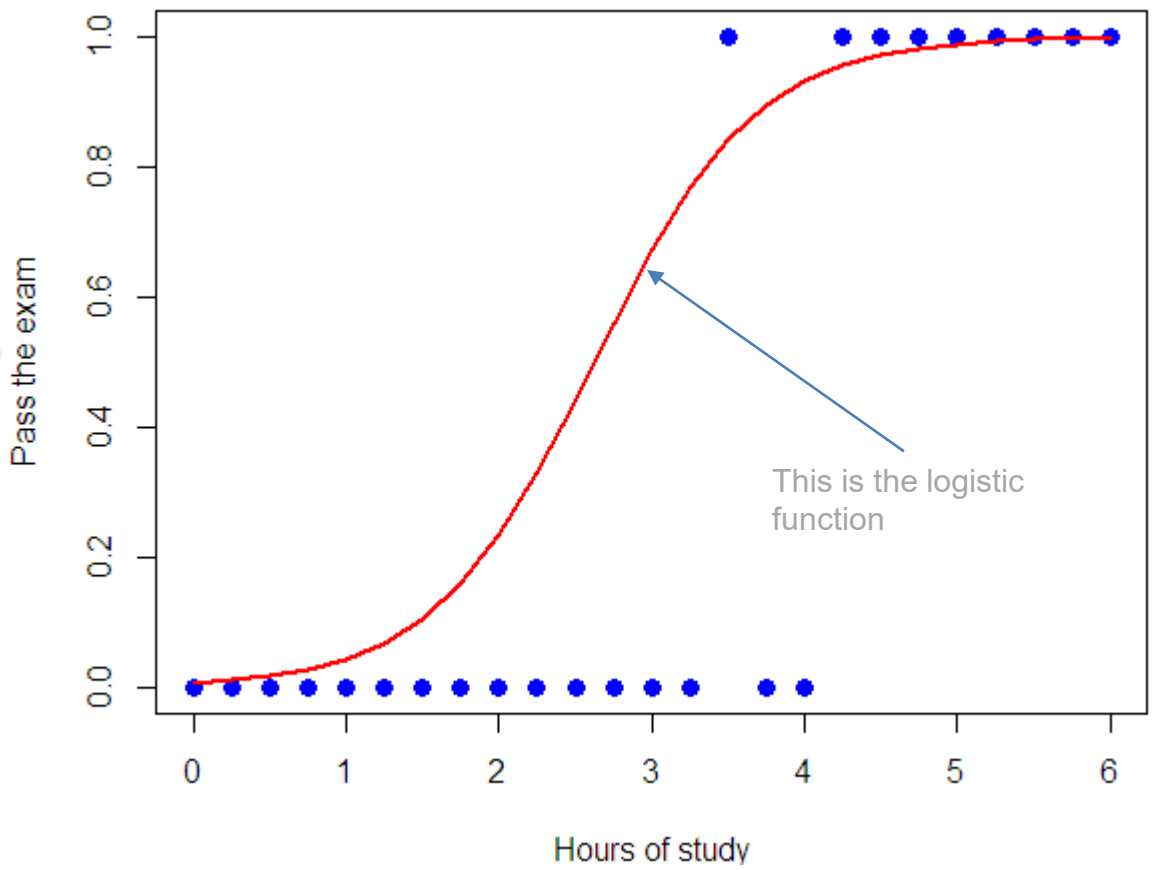
$$\Pr(\vec{x};\vec{\theta}) = \left[1 + e^{-\theta_0 - \vec{\theta}\cdot\vec{x}}\right]^{-1}$$

Esta es la función logística

Ejemplo: Hours of study vs pass the exam

Hours of Study	Pass the Exam
0,00	0
0,25	0
0,50	0
0,75	0
1,00	0
1,25	0
1,50	0
1,75	0
2,00	0
2,25	0
2,50	0
2,75	0
3,00	0
3,25	0
3,50	1
3,75	0
4,00	0
4,25	1
4,50	1
4,75	1
5,00	1
5,25	1
5,50	1
5,75	1
6,00	1

Probabilidad!



Threshold classifier

- La regresión logística muestra un valor de probabilidad para la salida.
  - Threshold classifier!
- *Ejemplo:*

  IF $\Pr(\vec{x};\vec{\theta}) \geq$ **0.5** ***THEN*** *class = PASS_EXAM,*

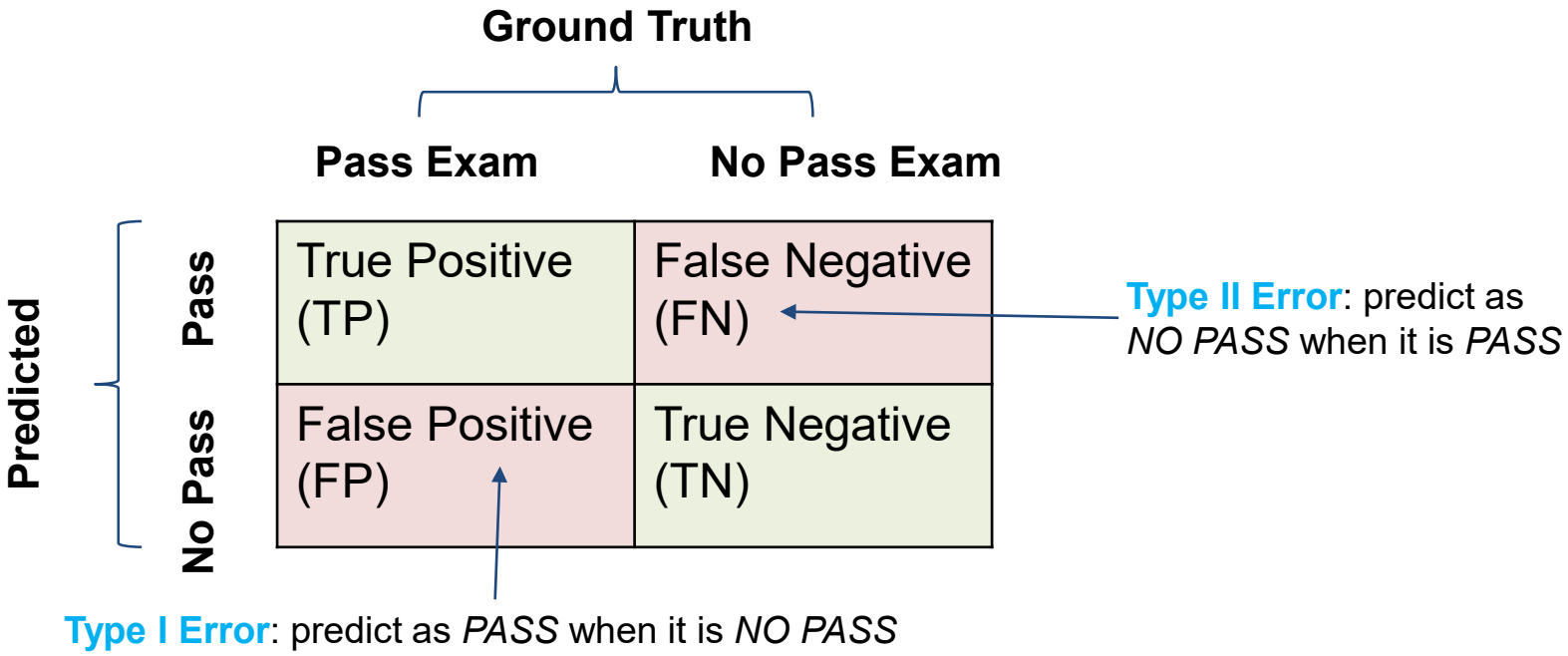
  Class = NO_PASS_EXAM ***OTHERWISE***
- Podemos modificar el threshold (0.5 en el ejemplo de arriba) para obtener una clasificación más precisa
- *Ejemplo:*

  IF $\Pr(\vec{x};\vec{\theta}) \geq$ **0.9** ***THEN*** *class = PASS_EXAM,*

  Class = NO_PASS_EXAM ***OTHERWISE***

Métricas de evaluación del modelo: matriz de confusión

- **Confusion matrix**
  - Nos indica cómo de bien se ha comportado el clasificador.
  - Vemos la relación entre los casos verdaderos (ground truth) y las predicciones.



Métricas de evaluación del modelo

- **A partir de la matriz de confusion, podemos obtener distintas métricas:**

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}}$$

$$= \text{probability of a positive test given that the patient has the disease}$$

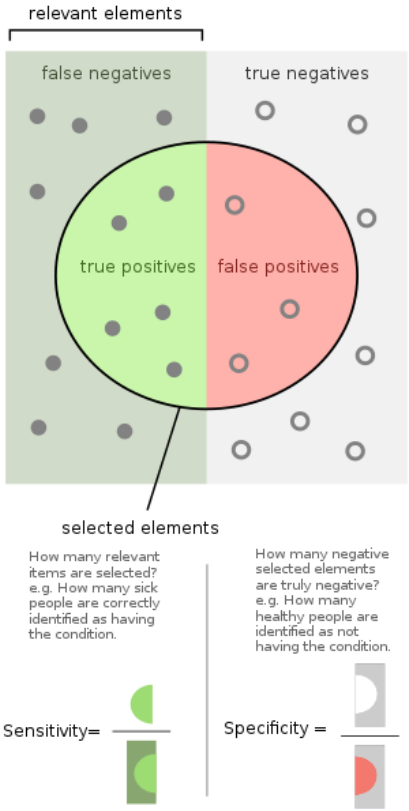
También llamado True Positive Rate o Recall

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}}$$

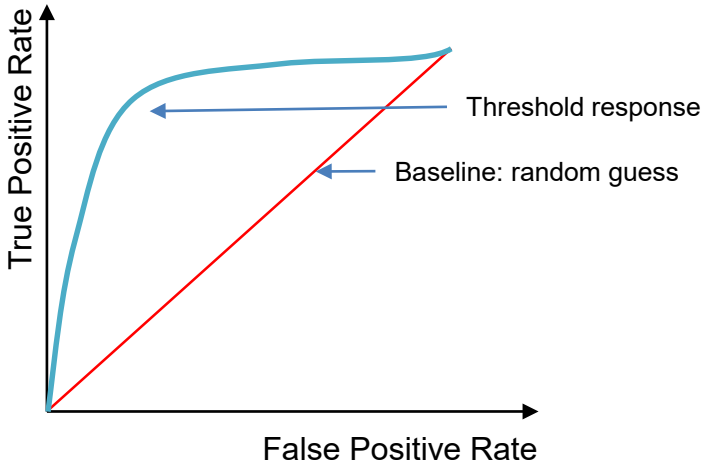
$$= \text{probability of a negative test given that the patient is well}$$

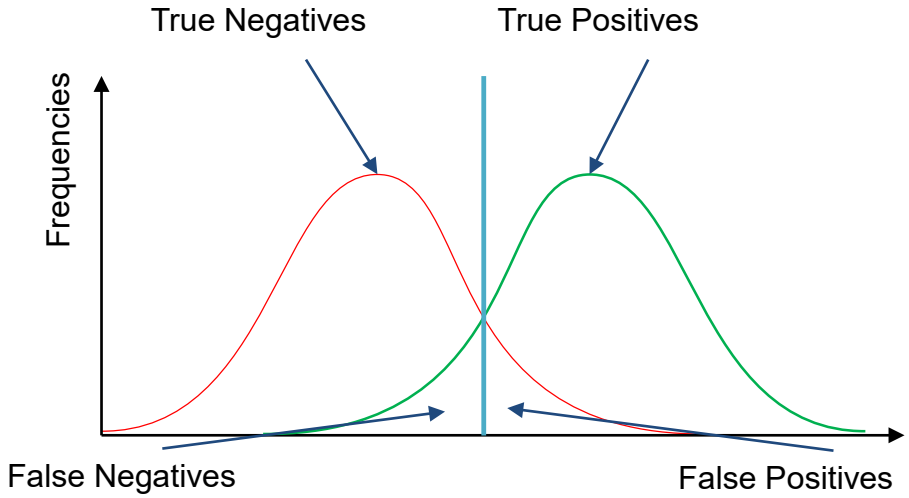
También llamado True Negative Rate



Receiver Operating Characteristic (ROC)

- A partir de la matriz de confusión, podemos generar la curva **Receiver Operating Characteristic** (ROC) para automatizar el proceso de encontrar el mejor threshold.
  - Es un plot que ilustra el rendimiento del clasificador binario a medida que se va variando el threshold.

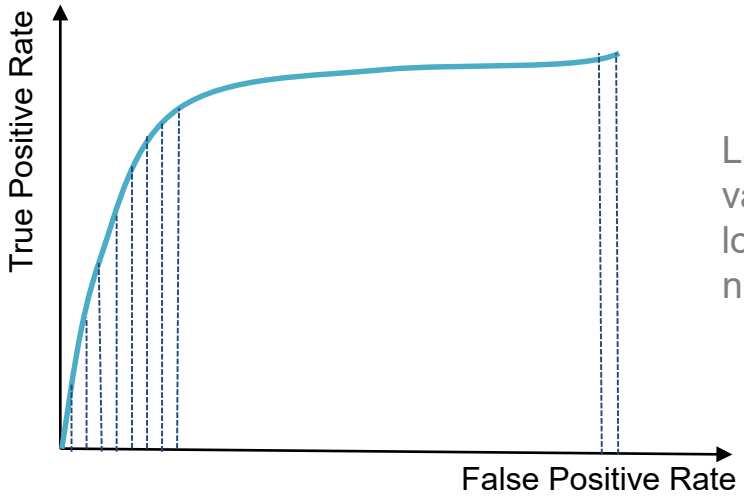




El problema es el **overlap entre predicciones**!

Receiver Operating Characteristic (ROC)

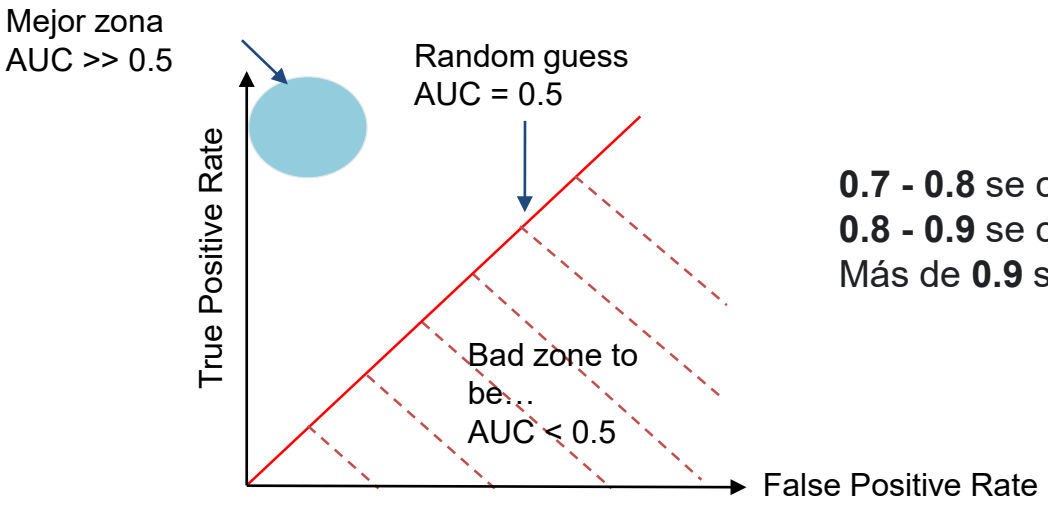
- Queremos un balance entre el True Positive Rate (TPR) y el False Positive Rate (FPR)
  - Como depende del threshold, podemos hacer un barrido para encontrar "el mejor threshold posible", por ejemplo maximizando el TPR.
  - Si definimos un número de thresholds ***muy grande*** obtenemos una curva



La curva es el resultado de variar el threshold y mostrar los valores de TPR/FPR de nuestro clasificador

Receiver Operating Characteristic (ROC)

- El área bajo la curva ROC nos dice cómo de bien se ha comportado nuestro clasificador:
  - Lo bien que discrimina entre clases
- Se llama **Area Under the Curve** (AUC)
  - Muestra la probabilidad de que el clasificador clasifique correctamente una muestra



0.7 - 0.8 se considera acceptable
0.8 - 0.9 se considera excelente
Más de **0.9** se considera excepcional