# Análisis de Clasificación

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

# Definición

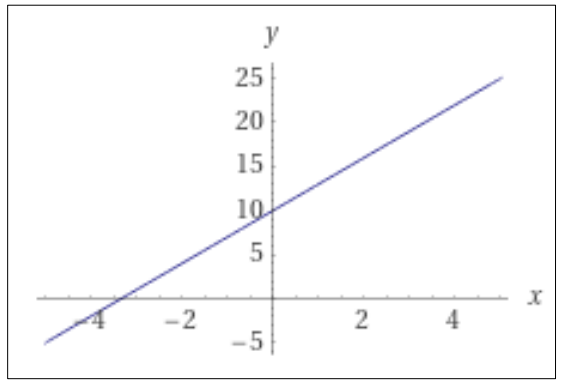
- Es un subconjunto del problema de regresión
- Se utiliza en problemas donde las variables son discretas
- Mayoría de enfoques resuelven problemas para variables binarias (aunque no todos)
- Cuando la variable tiene N categorías se suele representar como N modelos de clasificación binarios

# Recordatorio regresiones lineales

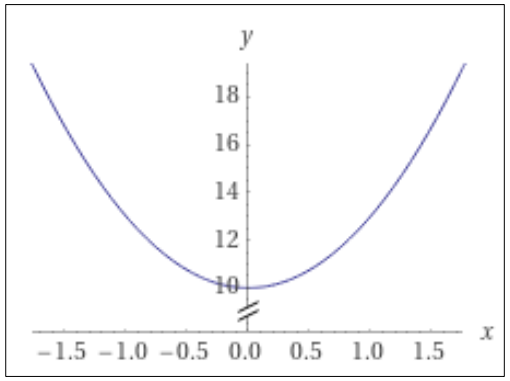
En un **modelo** de regresión lineal, la relación entre las variables es una función lineal de los parámetros.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad Y = \beta_0 + \beta_1 X^2 + \epsilon$$

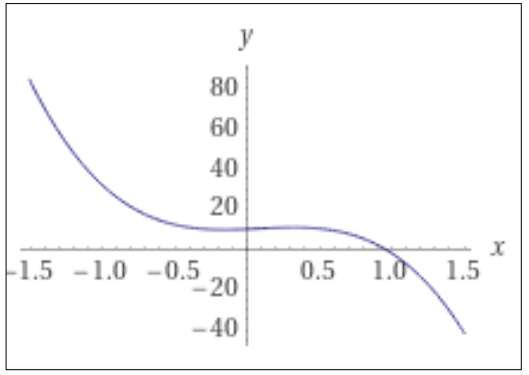
$$Y = \beta_0 + \beta_1 X + \beta_1 X^2 + \cdots + \beta_7 X^7 + \epsilon$$




$y = 10 + 3\,x$




$y = 10 + 3\,x^2$




$y = 10 + 3\,x + 5\,x^2 - 20\,x^3$

Son regresiones polinómicas de distinto grado y aunque el output de la función no es lineal, los **predictores** (parámetros) no dejan de ser una combinación lineal (o sea podemos ocupar el método de los mínimos cuadrados)

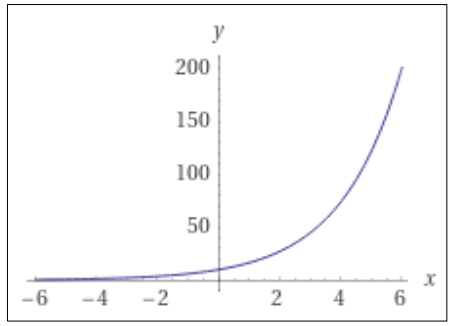
# Ejemplos de regresión NO lineal

En un **modelo** de regresión **no** lineal, la respuesta de la función no es lineal y los **predictores** (parámetros) no se combinan de forma lineal:

Una regresión **no** lineal puede ser modelada de diversas for**mas:**

Una regresión exponencial

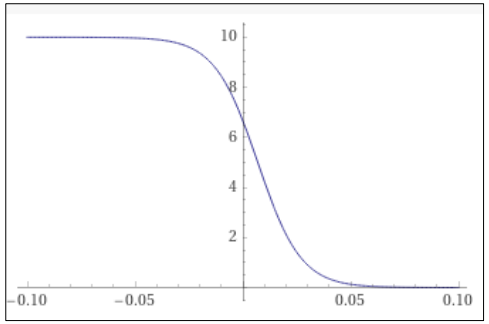
$$Y = \beta_0 \exp(\beta_1 X) + \epsilon$$




$$y = 10 \exp(0.5\, x)$$

Una regresión logística

$$Y = \frac{\beta_0}{1 + \beta_1 \exp(\beta_2 X)} + \epsilon$$



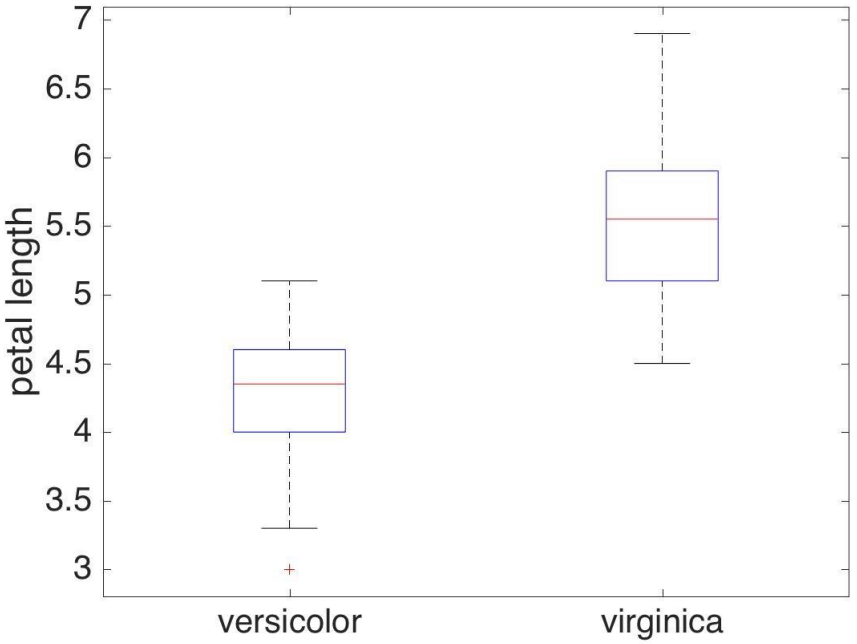

$$y = \frac{10}{1 + 0.5 \exp(100\, x)}$$

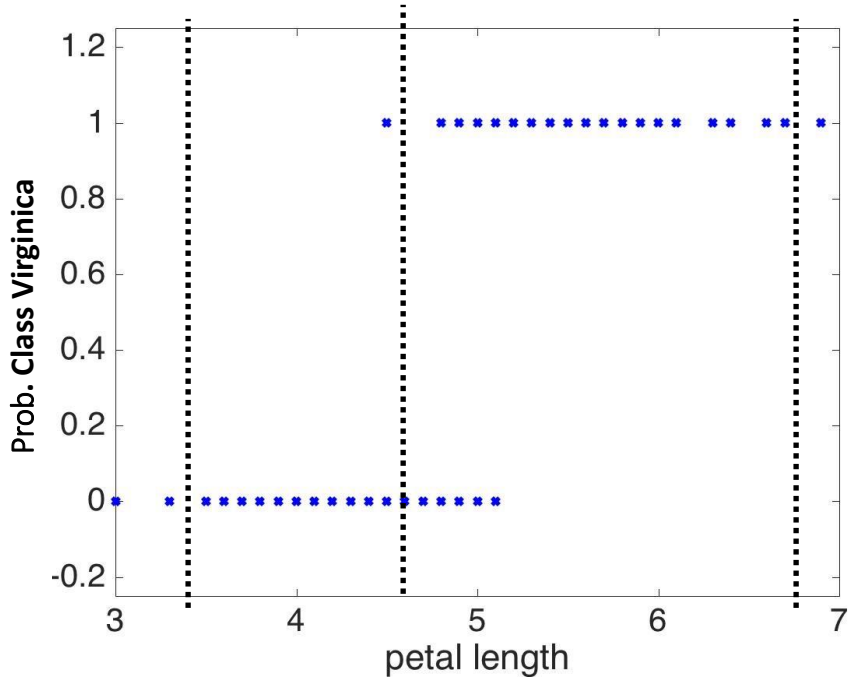
Una regresión de Poisson

Otros modelos…

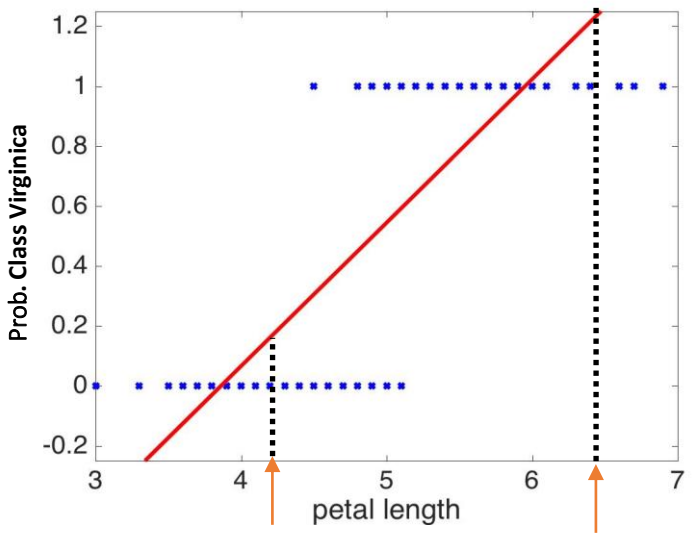
# Problema de clasificacion binario

Si tenemos 2 tipos de flores, cuya principal diferencias son el largo de sus pétalos, podríamos utilizar un modelo para clasificarlas



# Clasificación con regresión lineal



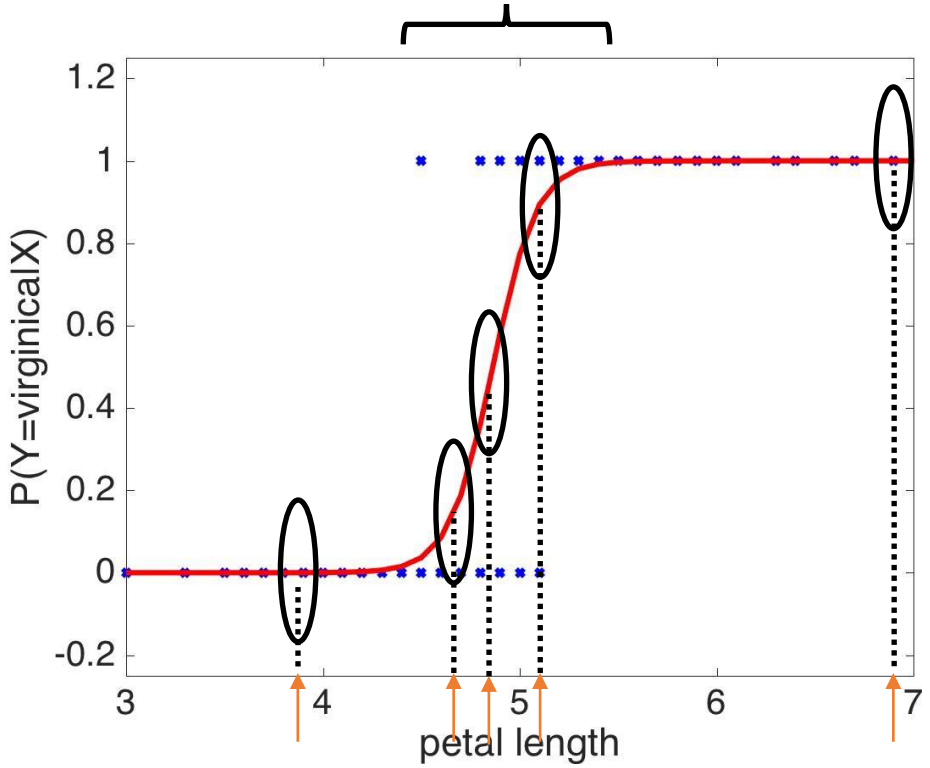

$$\hat{y}_i = -1.85 + 0.47x_i$$

Podríamos hacer un modelo de regresión lineal de grado 1 y usar la ecuación de la recta para ver la probidad que sea virginica, pero no se ajusta bien.

La función sale de los limites de 0 y 1, y adicionalmente no considera bien las proporciones de los datos en el centro.

Parte del problema es que los datos no se comportan de manera lineal, son binarios.

# Regresion logística

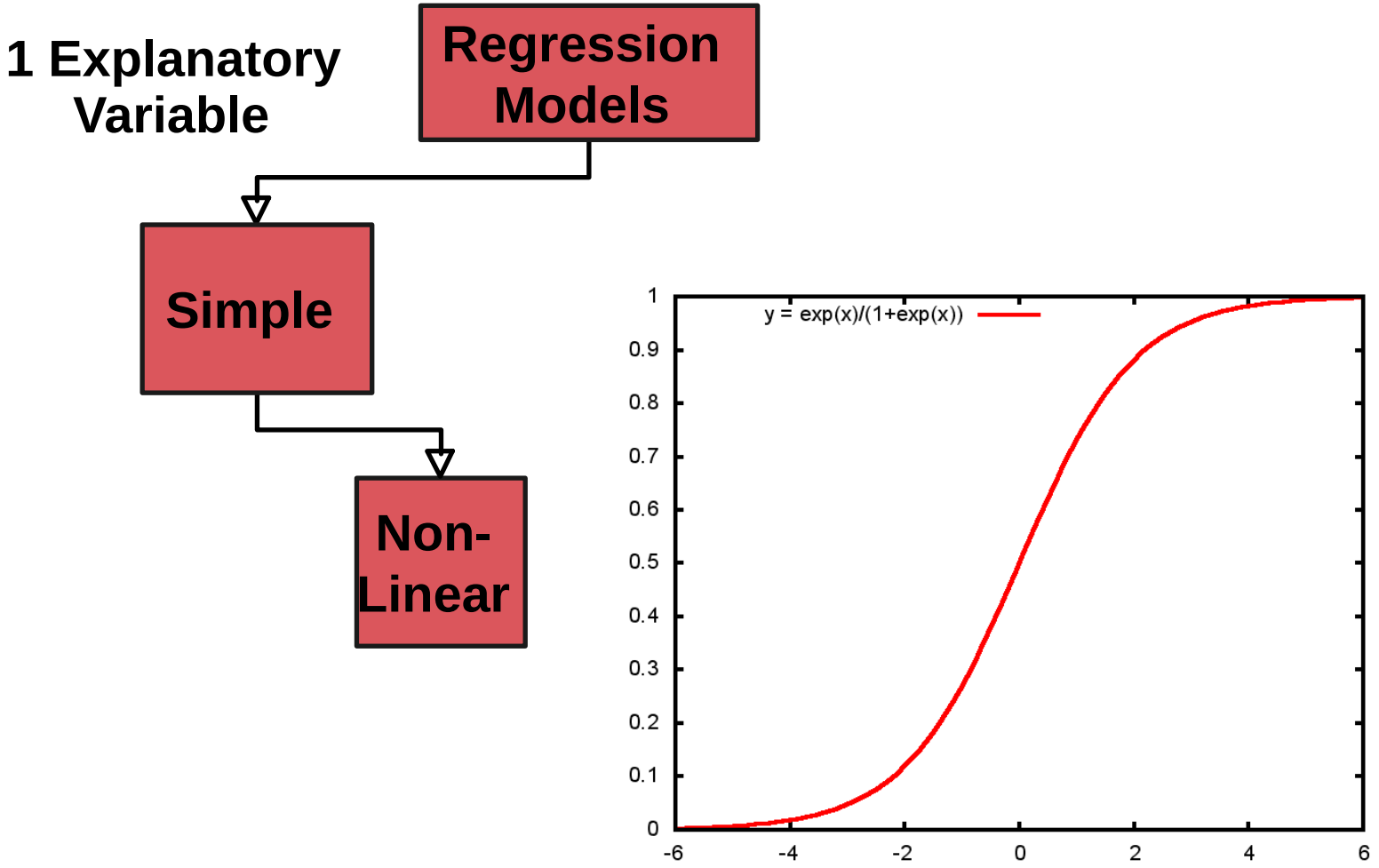



Si utilizamos una regresión NO lineal, por ejemplo una regresión logística, podemos encontrar una función que se adapte mucho mejor a una distribución dicotómica.

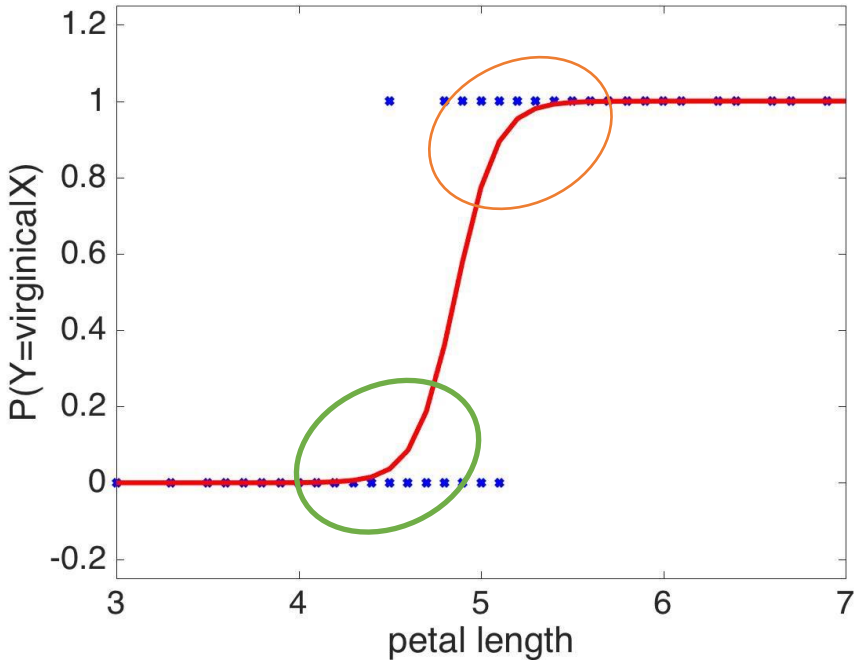
$$P(Y = 1|X = x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}$$

La función es dependiente de 2 constantes **beta 0** y **beta 1**

# Regresión logística simple



# Forma de la función logística



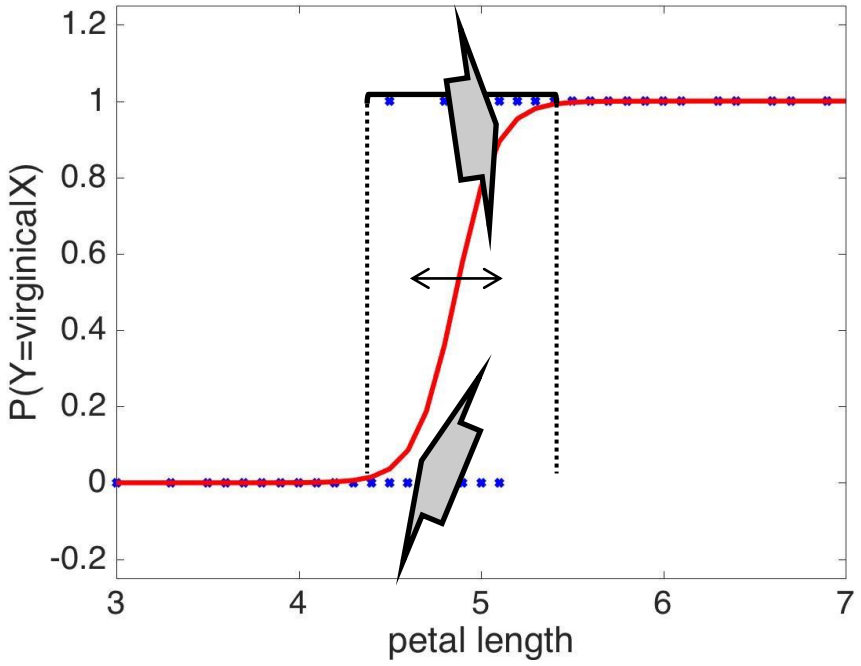

$$P(Y = 1|X = x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}$$

La función esta compuesta de dos exponenciales divididas. Una de ellas afecta el "despegue" desde el valor 0 y la otra el "aterrizaje" al valor 1.

Como las dos exponenciales dependen de las mismas constantes, podemos esperar que el comportamiento de "despegue" y "aterrizaje" sea simétrico.

Dependiendo de los beta, la función tendrá una forma distinta.

# Parámetros



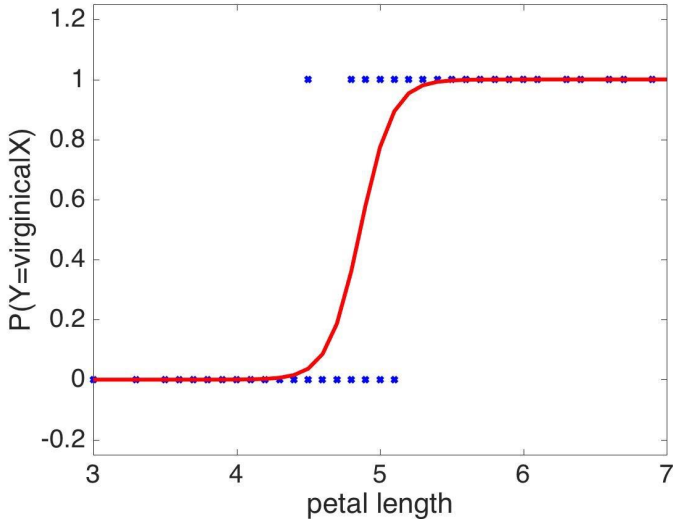

Beta 0 **no** es acompañado de la variable independiente. El valor de beta 0, define el centro de la función en el eje X

$$P(Y = 1|X = x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}$$

Beta 1 multiplica a la variable independiente, por lo cual tiene una relación con “petal length”. Beta 1 define que tan pronunciado es el “despegue” y “aterrizaje”.

# Aprendizaje

Nuestro objetivo es **encontrar los mejores Beta posible** haciendo que el **likelihood** del modelo se maximice. Dicho de otra forma, usar los puntos conocidos para ajustar de la mejor forma posible la función de probabilidad P(Y|X).




$$P(Y_i|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \pi_i$$

Consideramos que la clase a predecir se comporta de forma independiente Bernoulli.

$$L(M) = \prod_{i=1}^{n} P(Y_i|X_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

$$\text{Max } \ln(L(M|\beta_0, \beta_1)) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + exp(\beta_0 + \beta_1 x_i))$$

**Para encontrar Beta 0 y Beta 1, debemos estimar los parámetros; no existe una solución analítica.**

# Interpretación

El modelo de regresión logística viene dado por

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

que es equivalente a

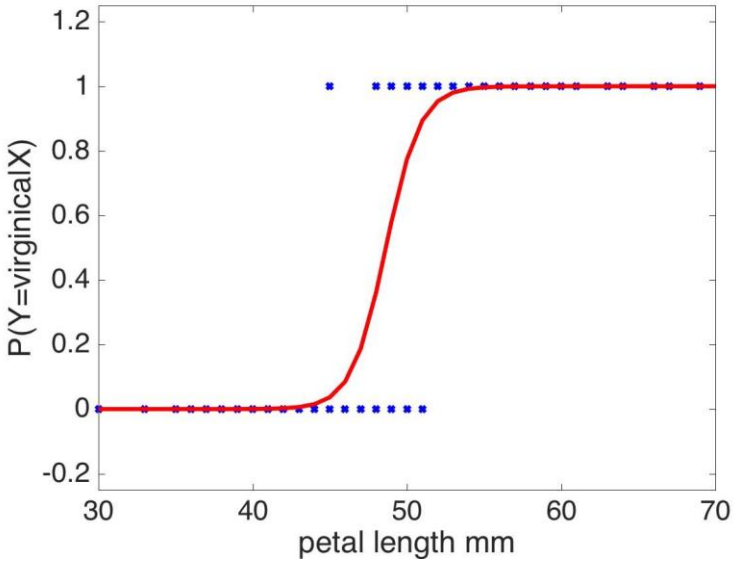
$$\ln\left(\frac{P(Y|X)}{1 - P(Y|X)}\right) = \beta_0 + \beta_1 X$$

Transformación logística

$$\frac{P(Y|X)}{1 - P(Y|X)} = \frac{\frac{\#successY}{n}}{1 - \frac{\#successY}{n}} = \frac{\#successY}{n - \#successY} = odds$$

Las probabilidades (odds) son la relación de resultados favorables con resultados desfavorables.

# Ejemplo




Beta0 = -43,78

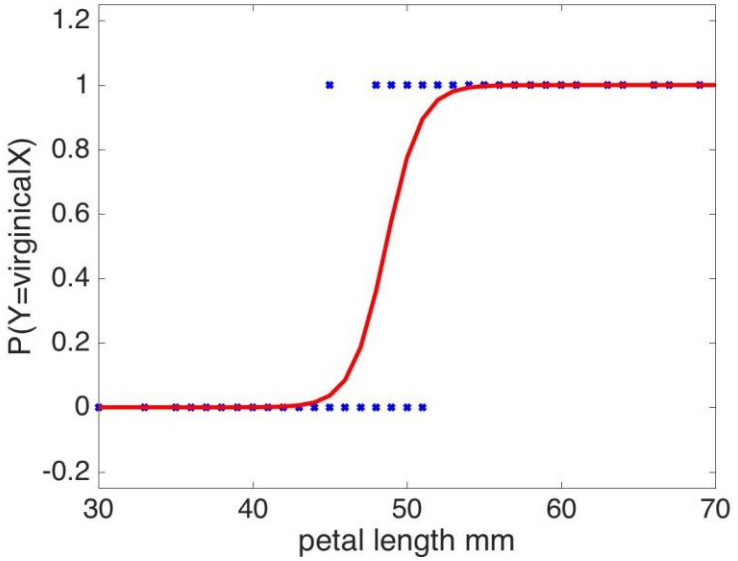
Beta1 = 0,90

Beta1 nos indica como la variable “petal length” afecta al modelo. Como es un número positivo, quiere decir que a medida que aumenta el “petal length” la probidad que un punto sea de clase virginica aumenta

$$P(Y = 1|X = x_i) = \frac{\exp(-43.78 + 0.90x_i)}{1 + \exp(-43.78 + 0.90x_i)}$$

Beta0 indica el peso base. Como es un numero negativo, existe una peso base de clasificar todo como **no virginica**.

# Ejemplo




Beta0 = -43,78

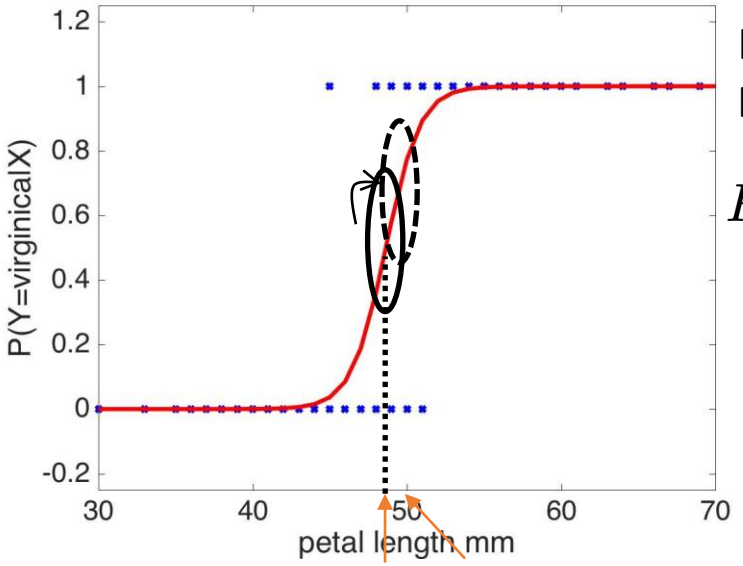
**Beta1 = 0,90**

$$P(Y = 1|X = x_i) = \frac{\exp(-43.78 + 0.90x_i)}{1 + \exp(-43.78 + 0.90x_i)}$$

$$\exp(\hat{\beta}_1) = \exp(0.90) \approx 2.46$$

Este numero se denomina el multiplicador de **odds**. Por cada unidad que avancemos en la variable “petal length” (mm) este numero será el amplificador de odds.

# Ejemplo




Beta0 = -43,78

**Beta1 = 0,90**

$$P(Y=1|X=x_i)=\frac{\exp(-43.78+0.90x_i)}{1+\exp(-43.78+0.90x_i)}$$

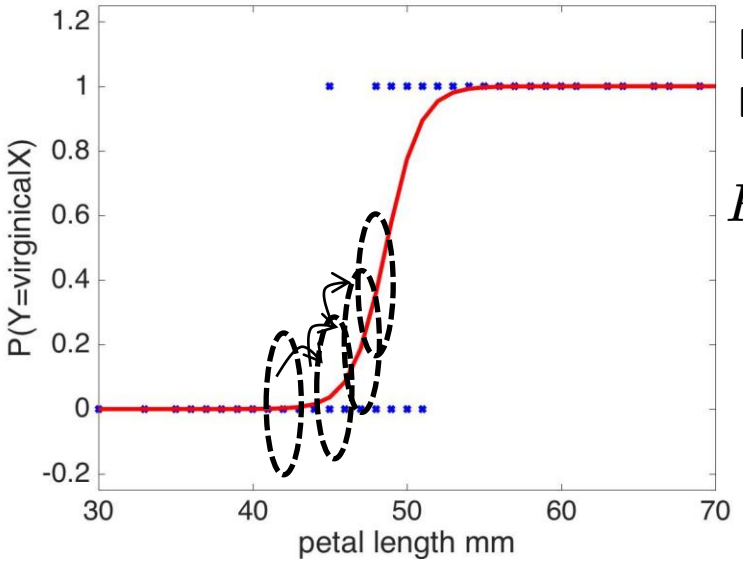
$$\exp(\hat{\beta}_1)=\exp(0.90)\approx 2.46$$

$$P(Y=1|X=49mm)\approx 0.58 \Rightarrow \text{odds}=\frac{0.58}{1-0.58}\approx 1.39$$

$$P(Y=1|X=50mm)\approx 0.77 \Rightarrow \text{odds}=\frac{0.77}{1-0.77}\approx 3.41\approx 1.39*2.46$$

Un nuevo punto **a 1 unidad de distancia del otro**, tendrá un **odds** igual al de su vecino multiplicado por el **multiplicador** de odds

# Ejemplo




Beta0 = -43,78

Beta1 = 0,90

$$P(Y=1|X=x_i)=\frac{\exp(-43.78+0.90x_i)}{1+\exp(-43.78+0.90x_i)}$$

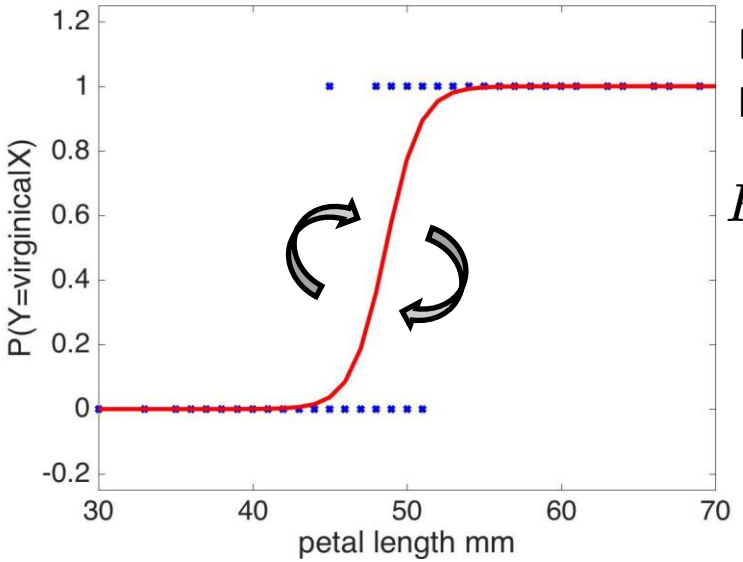
$$\exp(\hat{\beta}_1)=\exp(0.90)\approx 2.46$$

$$P(Y=1|X=49mm)\approx 0.58 \Rightarrow \text{odds}=\frac{0.58}{1-0.58}\approx 1.39$$

$$P(Y=1|X=50mm)\approx 0.77 \Rightarrow \text{odds}=\frac{0.77}{1-0.77}\approx 3.41\approx 1.39*2.46$$

O sea, mientras mas grande un multiplicador de odds, la variable asociada a dicho Beta, tiene mas peso en afectar la probabilidad que una entidad **pertenezca a una clase o a otra.**

# Ejemplo




Beta0 = -43,78

Beta1 = 0,90

$$P(Y=1|X=x_i)=\frac{\exp(-43.78+0.90x_i)}{1+\exp(-43.78+0.90x_i)}$$

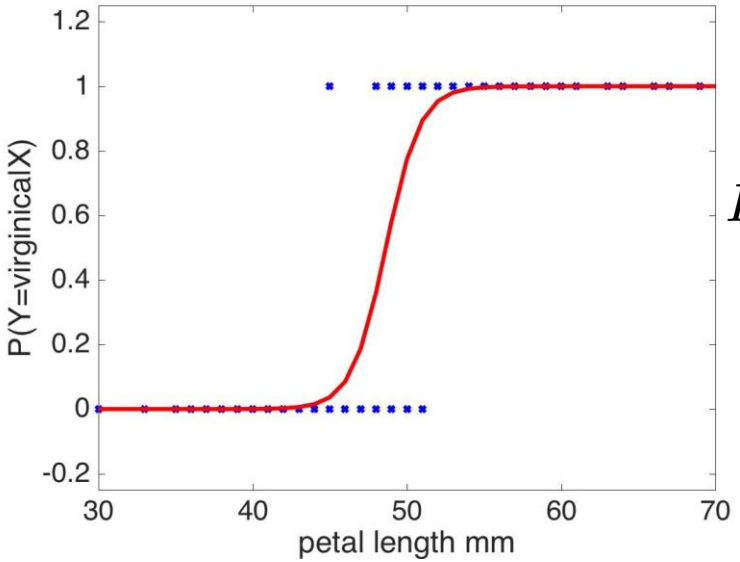
$$\exp(\hat{\beta}_1)=\exp(0.90)\approx 2.46$$

$$P(Y=1|X=49mm)\approx 0.58 \Rightarrow \text{odds}=\frac{0.58}{1-0.58}\approx 1.39$$

$$P(Y=1|X=50mm)\approx 0.77 \Rightarrow \text{odds}=\frac{0.77}{1-0.77}\approx 3.41\approx 1.39*2.46$$

Otra forma de verlo, es que las variables asociadas a multiplicadores de odds grandes, provocan que la curva de la función sea mas pronunciada. Al movernos muy poco en X (variable), variamos mucho Y (probabilidad)

# Ejemplo




Beta0 = -43,78

Beta1 = 0,90

$$P(Y = 1|X = x_i) = \frac{\exp(-43.78 + 0.90x_i)}{1 + \exp(-43.78 + 0.90x_i)}$$

$$\exp(\hat{\beta}_1) = \exp(0.90) \approx 2.46$$

Betas con valores alejados de 0 nos indican que la variable asociada, a dicho Beta, ayuda a separar muy bien las clases en términos probabilísticos.

# Regresión logística múltiple

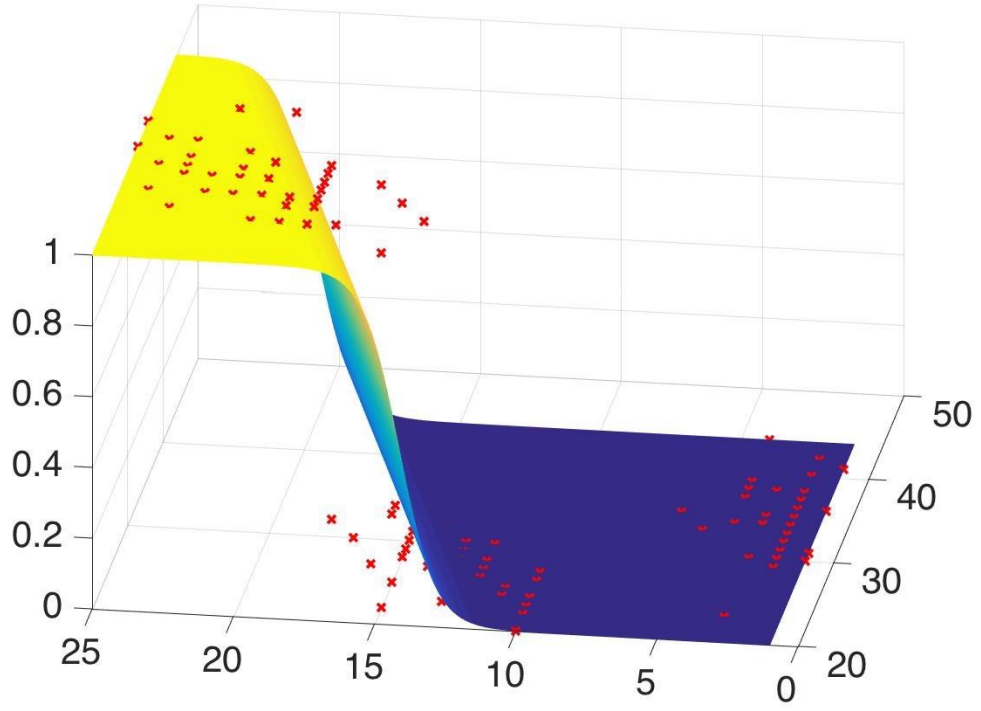
$$P(Y|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}$$

En un regresión de múltiples dimensiones, nuestra forma se agranda. Se añade un Beta por cada variable.

Todo el resto funciona usando la misma lógica, solo que ahora el algoritmo debe buscar no solo Beta0 y Beta1; sino **todos los betas de manera simultanea para maximizar el likehood**.

$$\exp(\hat{\beta}_1)$$

El multiplicador de Beta_i será el ratio de cambio en la probabilidad, que aporta esa dimensión, al problema por cada unidad de desplazamiento, suponiendo que todas las otras variables se mantienen constantes.



# Análisis de Clasificación

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

# Evaluación de modelos supervisados

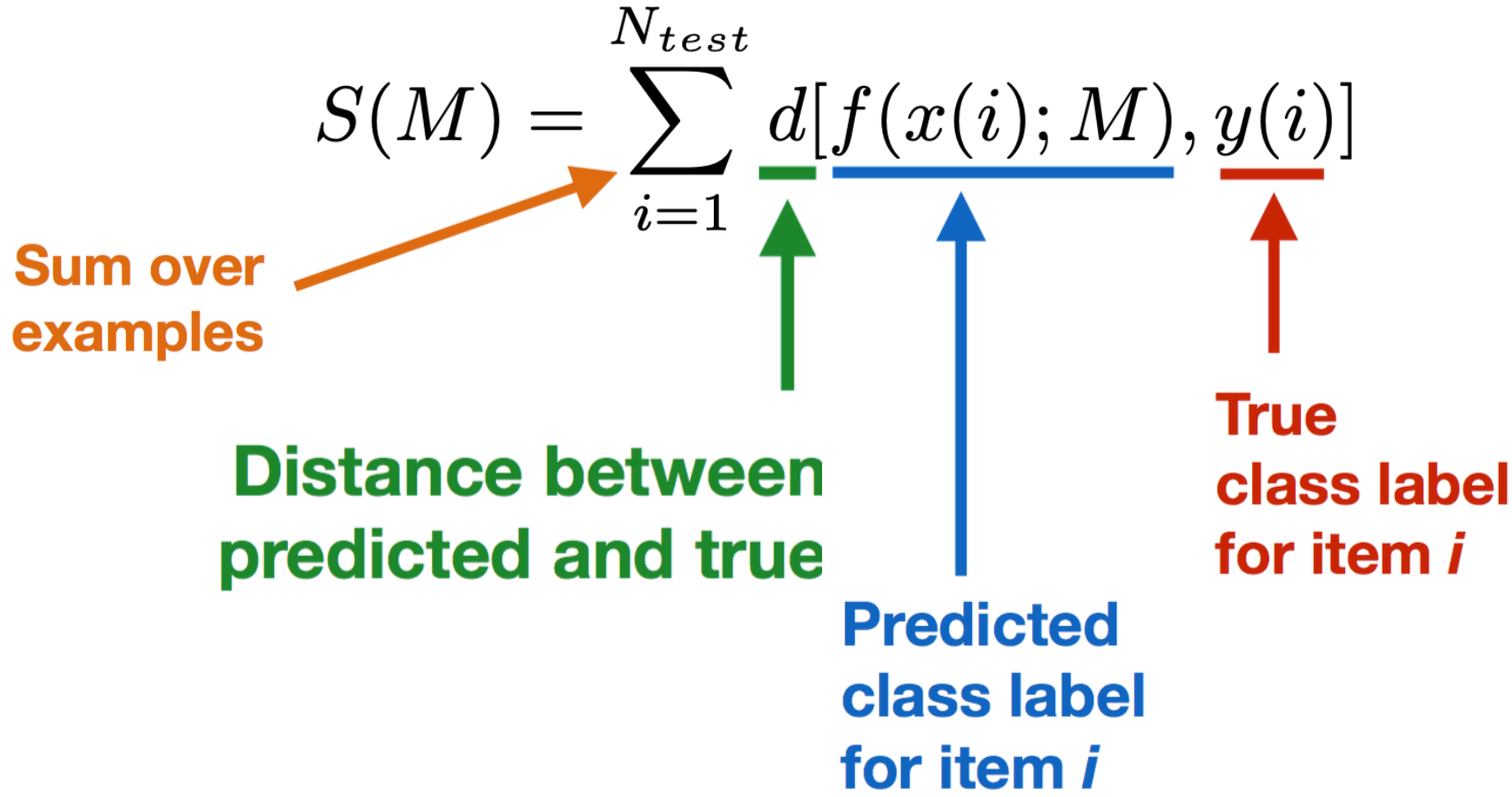
Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

# Métricas de evaluación

Evaluación del rendimiento para modelos predictivos:



## Métricas comunes de desempeño

- Zero-one loss:

$$S_{0/1}(M) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} I[f(x(i); M), y(i)]$$

$$\text{where } I(a,b) = \begin{cases} 1 & a \neq b \\ 0 & \text{otherwise} \end{cases}$$

- Squared loss:

$$S_{sq}(M) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} [f(x(i); M) - y(i)]^2$$

# Matriz de confusión

Se concentra en la capacidad predictiva de un modelo, en lugar de la rapidez con la que se tarda en clasificar o crear modelos, escalabilidad, etc.

| **Confusion matrix** | | **Predicted Class** | |
|---|---|---|---|
| | | No | Yes |
| **Actual Class** | No | True Negative | False Positive |
| | Yes | False Negative | True Positive |

Métricas asociadas para la matriz de confusión:

$$\text{Recall} = \frac{TP}{TP + FN}$$

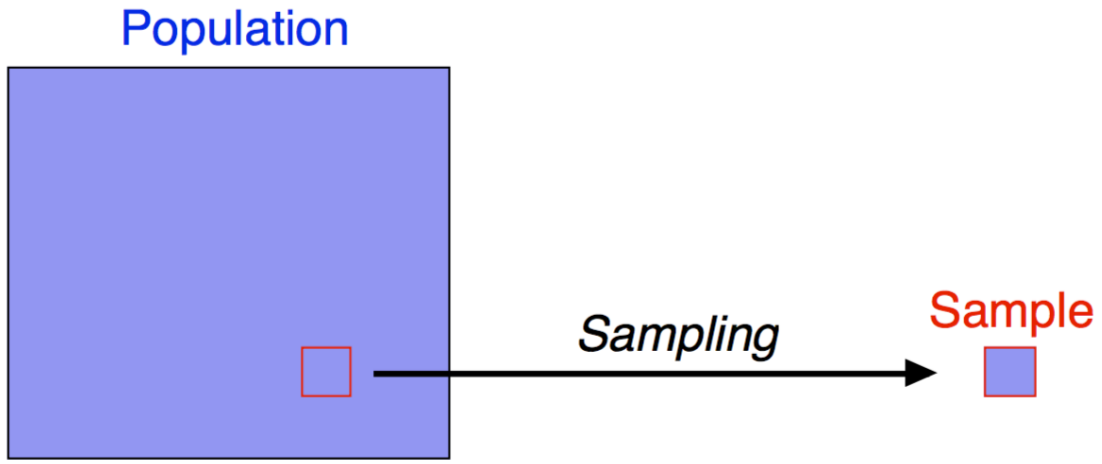
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{F1-score} = \frac{2 * TP}{2 * TP + FP + FN}$$

# Muestreo

En minería de datos a menudo trabajamos con una muestra de datos de la población de interés.

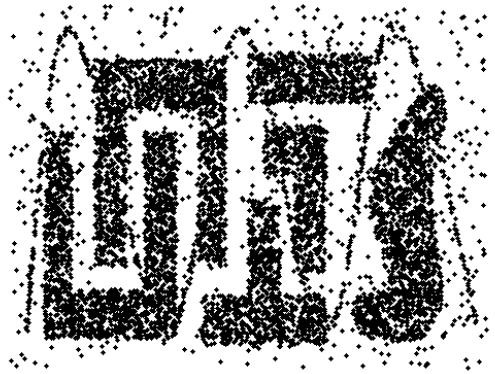


# Tipos de muestreo

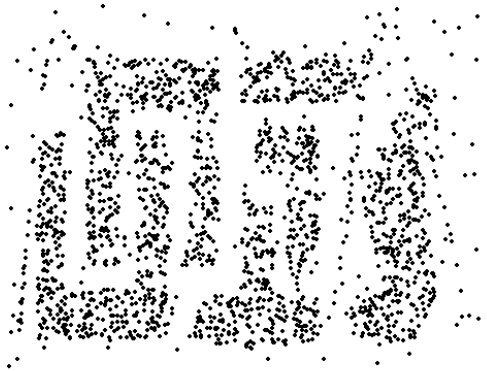
**Muestreo aleatorio simple**: Hay una probabilidad igual de seleccionar cualquier elemento en particular
**Muestreo sin reemplazo**: a medida que se selecciona cada elemento, se elimina de la población
**Muestreo con reemplazo**: los artículos no se eliminan de la población, ya que se seleccionan para la muestra; el mismo artículo se puede recoger más de una vez
**Muestreo estratificado**: Dividir los datos en varias particiones; a continuación, extraer muestras aleatorias de cada partición

# Tamaño del muestreo

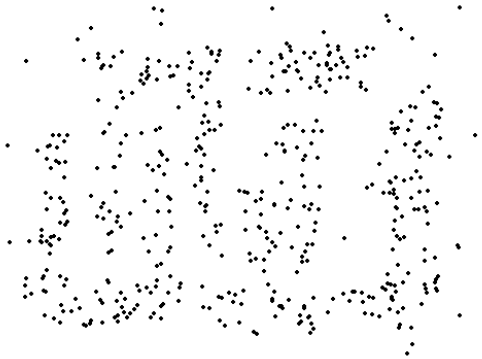
¿Cómo afecta el tamaño de la muestra
al aprendizaje?



**8000 points**



**2000 Points**

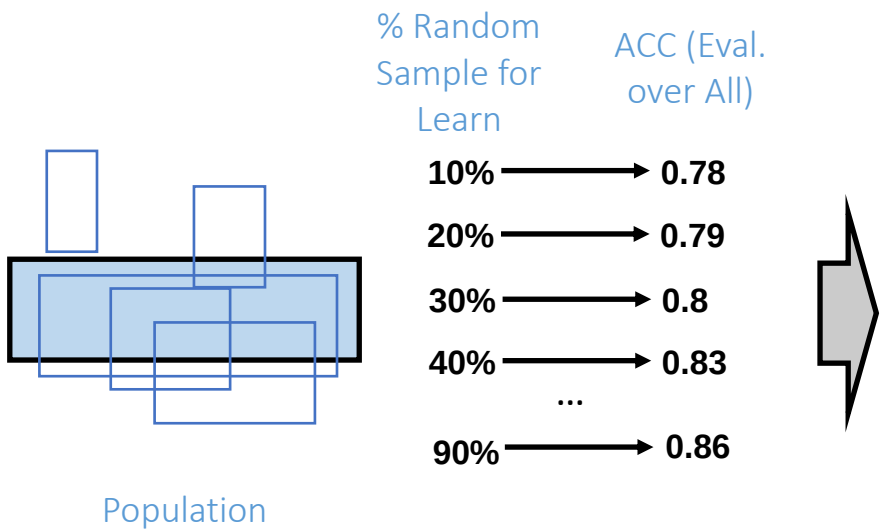


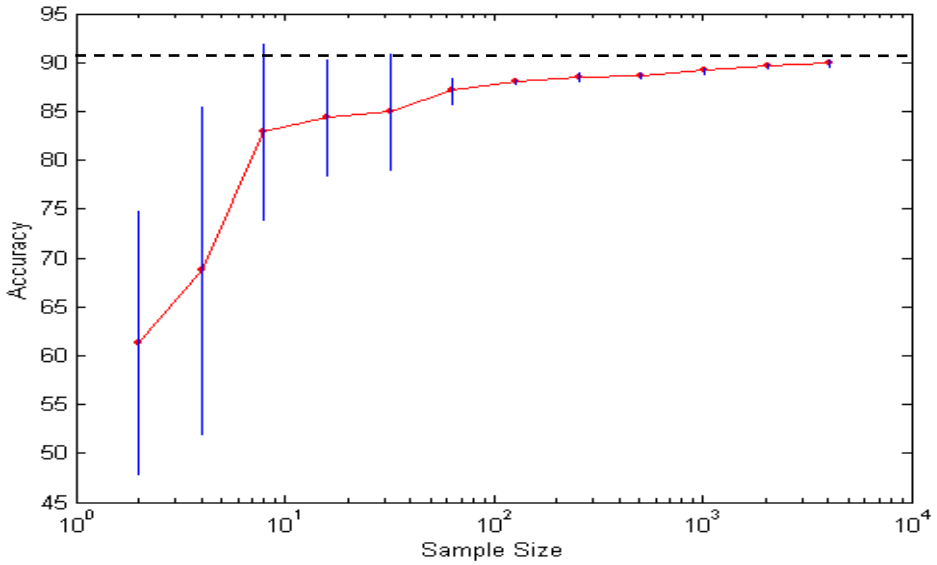
**500 Points**

# Error según tamaño de muestra

Curva de aprendizaje: muestra cómo cambia la precisión con diferentes tamaños de muestra.

**Desde el dataset S, donde | S|=n**
**Para i=[10, 20, ... ,100]**
**Muestra aleatoria de i% de S para construir la muestra S'**
**Entrena modelo con S'**
**Evalua modelo con S**
**Visualiza tamaño del conjunto de entrenamiento vs precisión**

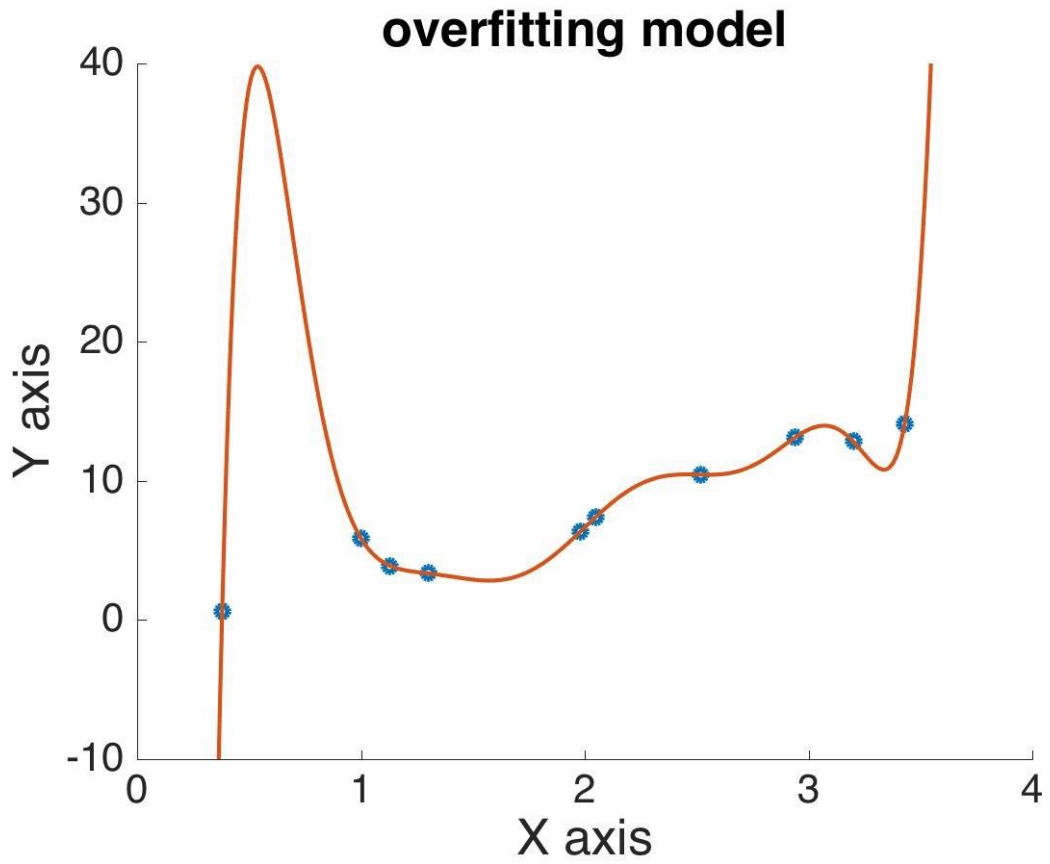


# Sobreajuste

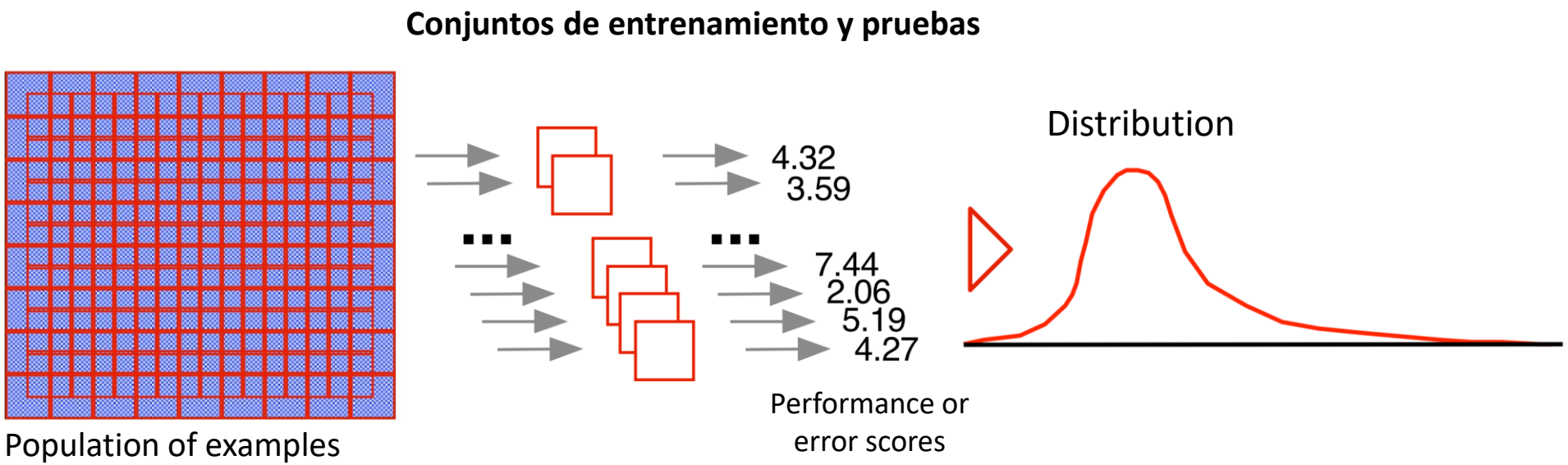
En el sobreajuste, un modelo estadístico describe el error aleatorio o el ruido en lugar de la relación subyacente.

El sobreajuste se produce cuando un modelo es excesivamente complejo, como tener demasiados parámetros en relación con el número de observaciones.



# Muestreo de errores

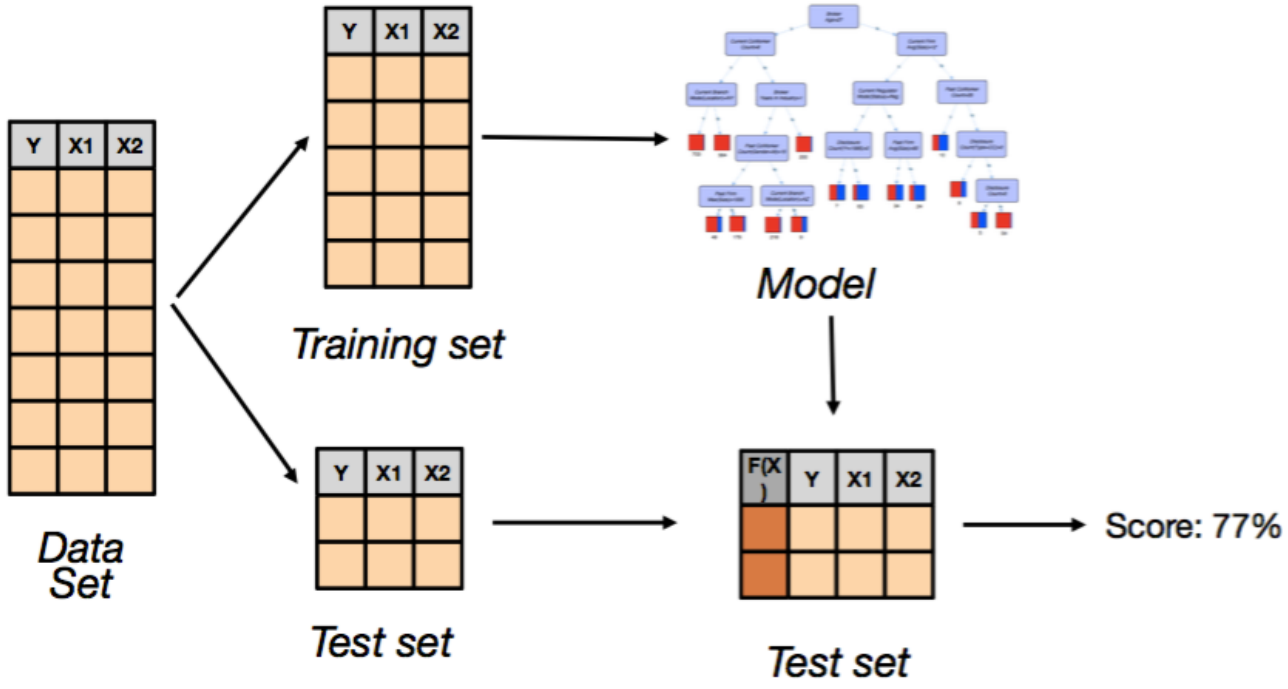
Para estimar el error de un modelo también podemos utilizar el muestreo para estimar la distribución del error.




- Distribución => hay una media y una variación de la distribución de errores.

# Conjunto de prueba y entrenamiento

Dividir conjunto de datos S en un conjunto de entrenamiento y otro de prueba

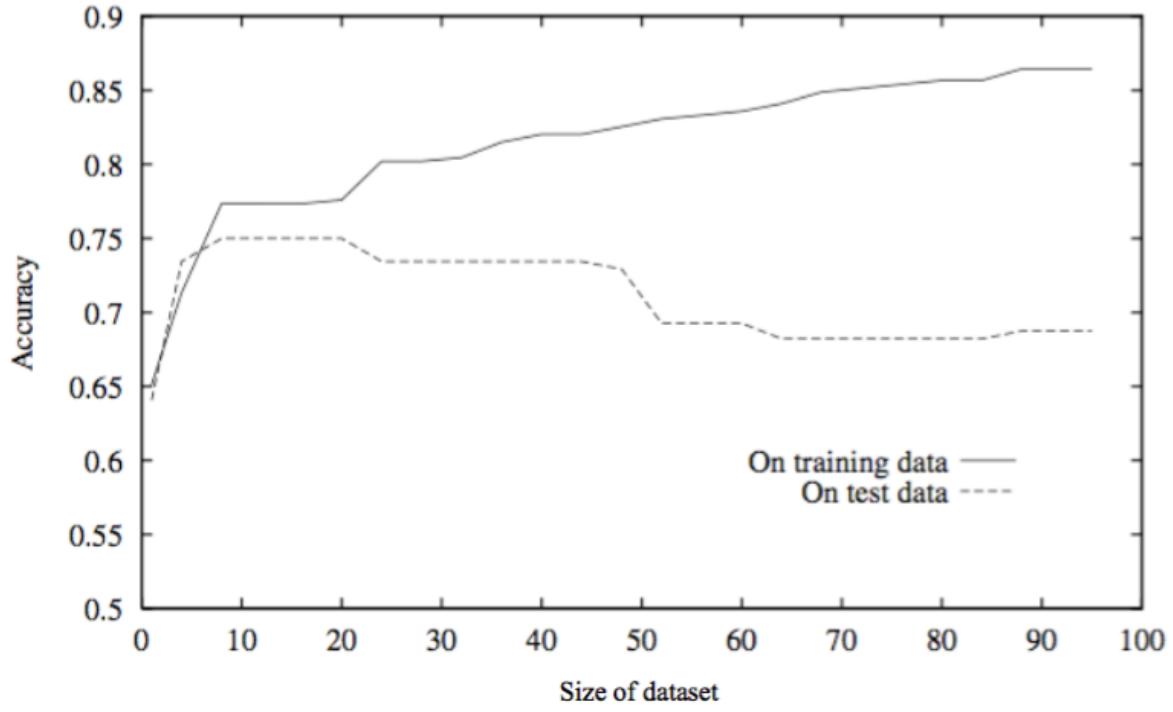



- Estimación variará debido al tamaño y composición del conjunto de pruebas

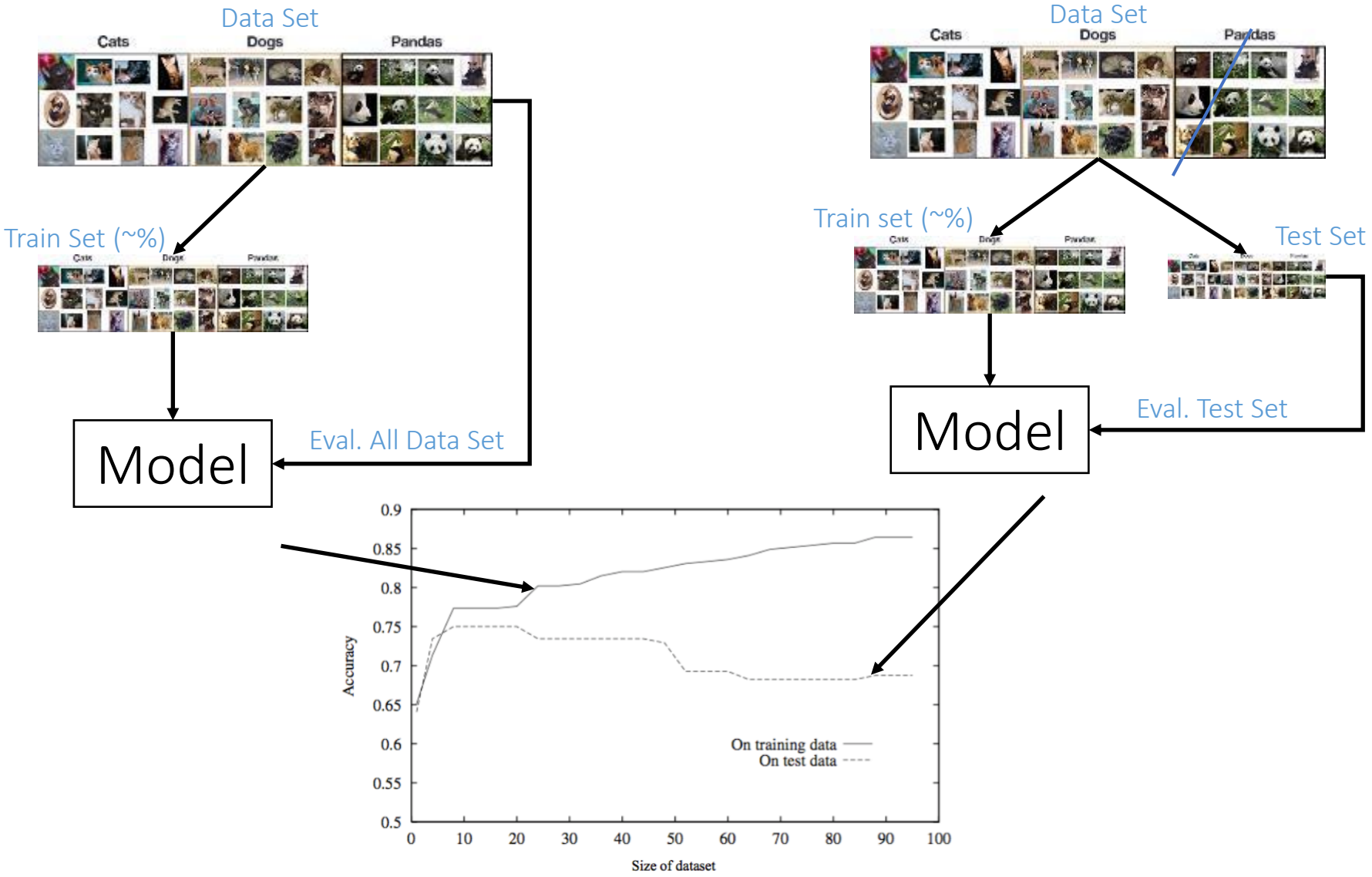
# Proceso de entrenamiento y testeo

Del conjunto de datos S, divida el conjunto de datos en S train y S test
Para i=[10, 20, ... ,100]

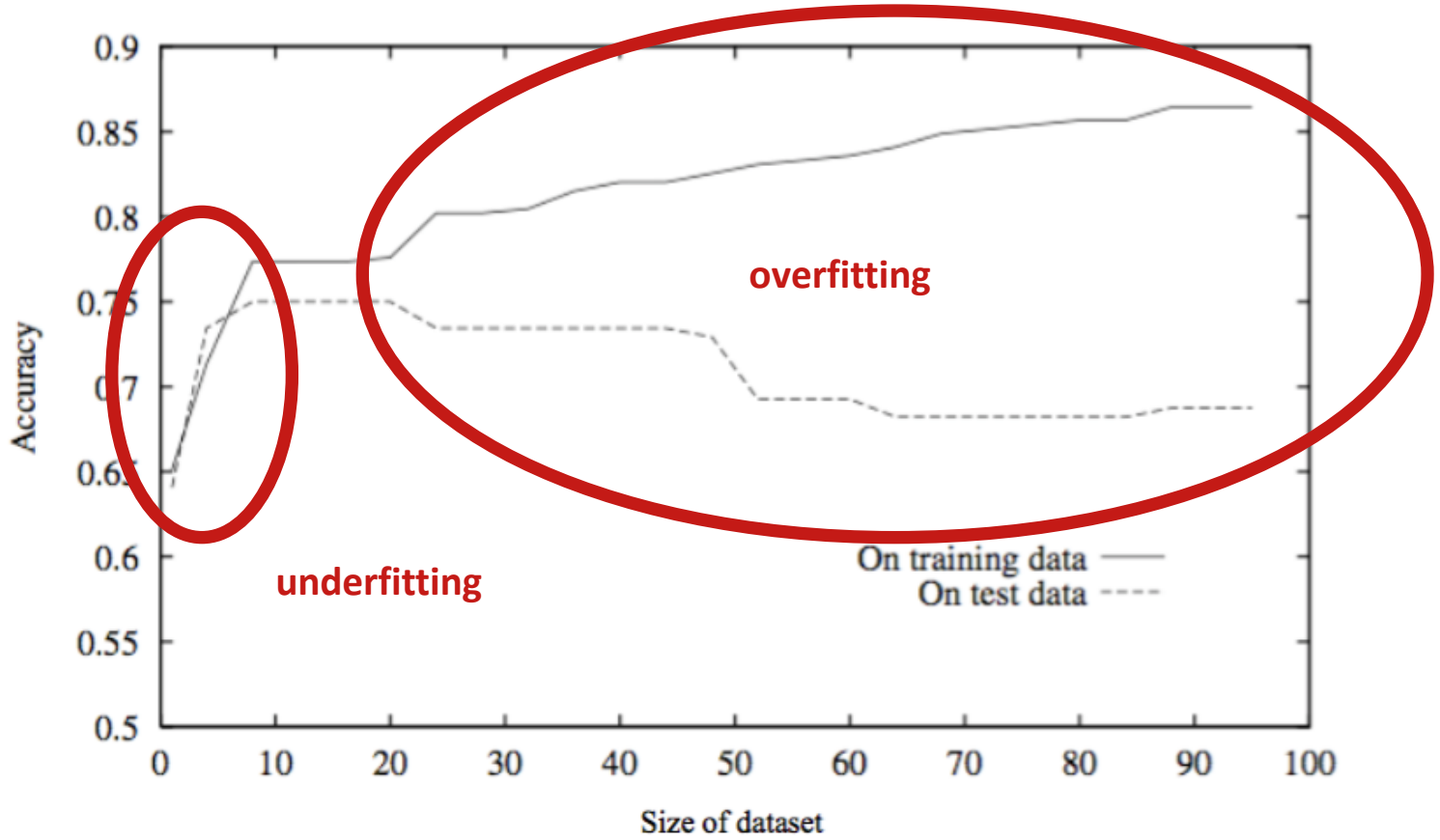
Muestrear aleatoriamente el i% para construir la muestra S train
Entrenar modelo con S train
Evaluar modelo en S test



# Ejemplo


Data Set
Cats
Dogs
Pandas
Train Set (~%)
Model
Eval. All Data Set
Data Set
Cats
Dogs
Pandas
Train set (~%)
Test Set
Model
Eval. Test Set
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
0
10
20
30
40
50
60
70
80
90
100
Size of dataset
On training data
On test data

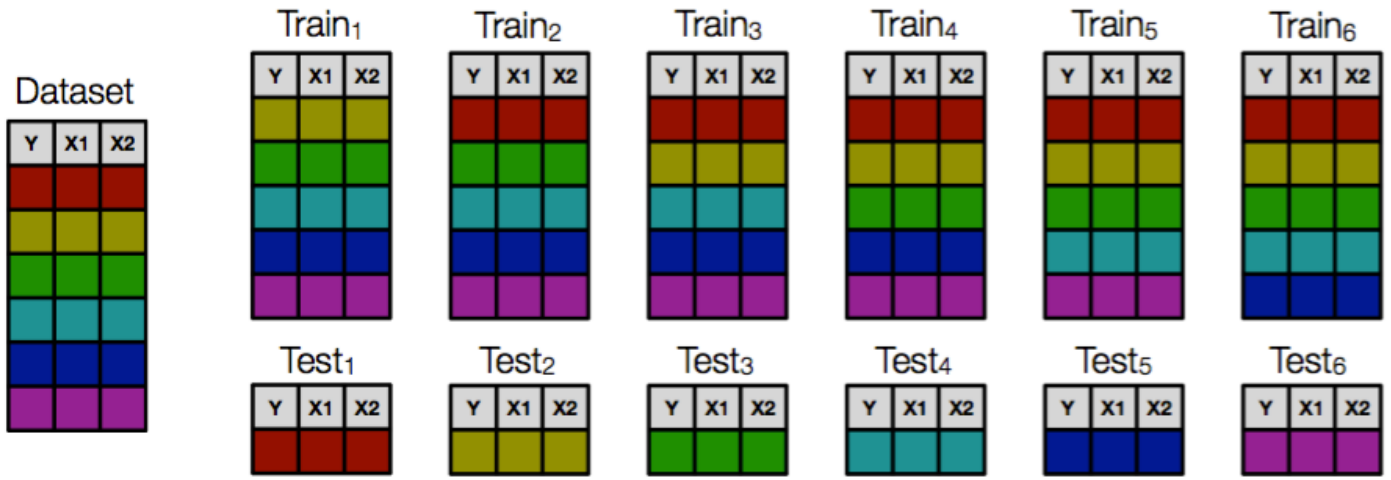
# Sobre ajuste y sub ajuste



# Validación cruzada (k-fold)

La validación cruzada combina (promedios de) las medidas de ajuste (error de predicción) para obtener una estimación más precisa del desempeño de la predicción del modelo

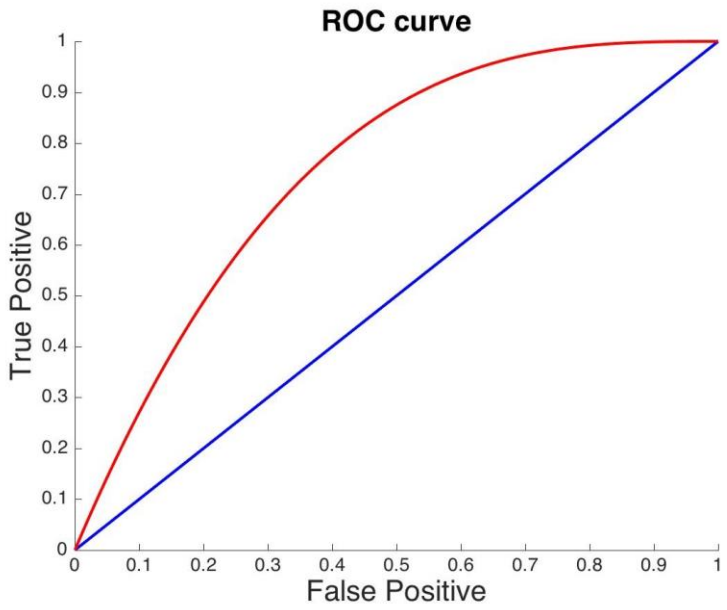
Particionar aleatoriamente los datos de entrenamiento en k pliegues
Para i=1 a k
Entrenar modelo con partición de entrenamiento i
Evaluar modelo con partición de prueba i
Promediar resultados de k pliegues



# ROC curve

Desarrollada en la década de 1950 en teoría de detección de señales para analizar señales ruidosas
Caracteriza la compensación entre golpes positivos y falsas alarmas
La curva ROC traza la tasa de Verdaderos positivos (TP) en el eje y contra la tasa Falsos Positivos en el eje x para diferentes valores.

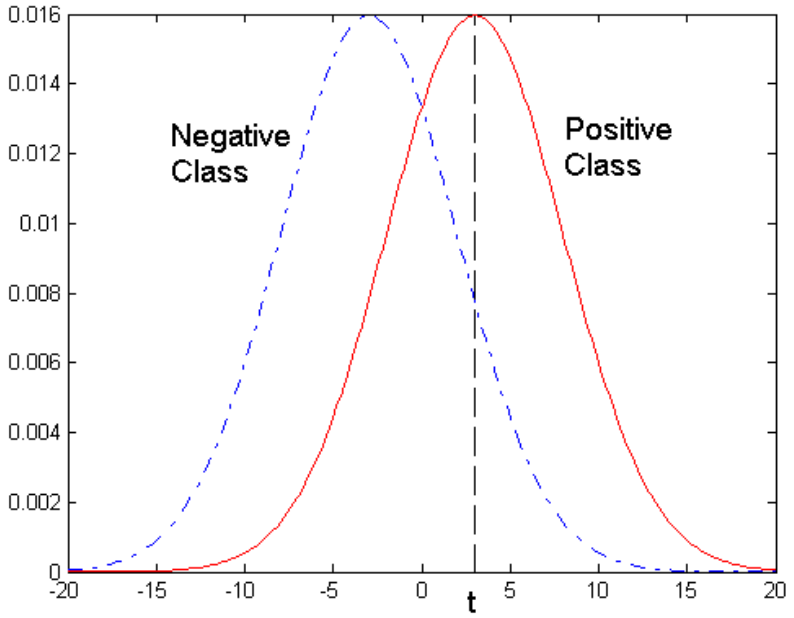
- Desempeño de cada clasificador es representado como un punto en la curva ROC.
- Al cambiar el umbral de algoritmo, distribución de muestras o matriz de costes, cambia la ubicación del punto, que genera la curva final

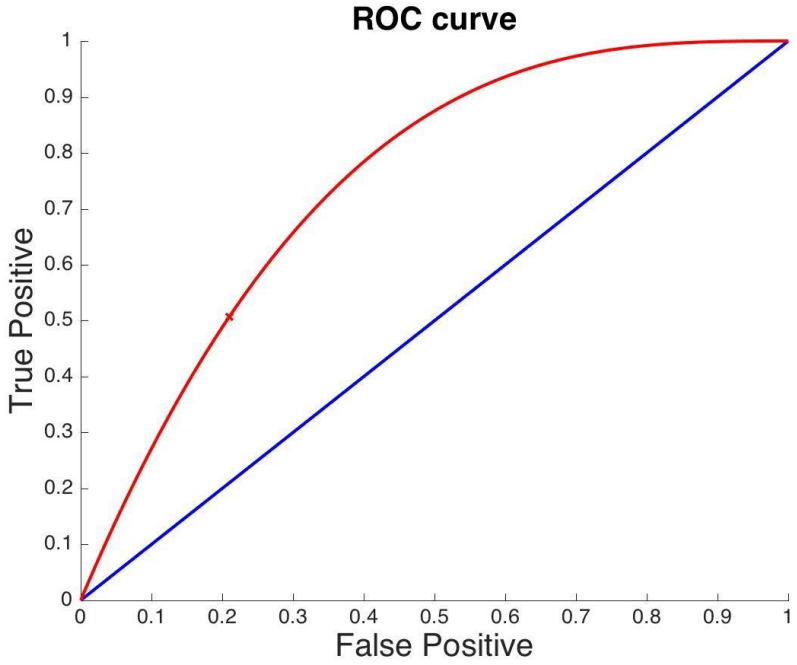


# Construcción de la curva ROC

Conjunto de datos que contiene 2 clases (positivas y negativas).

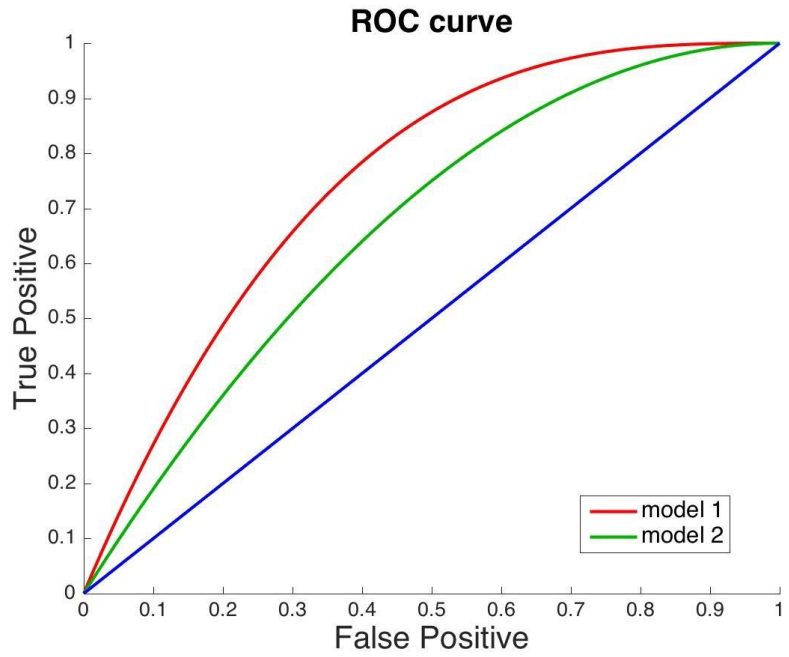
Cualquier punto ubicado en x > t se clasifica como positivo

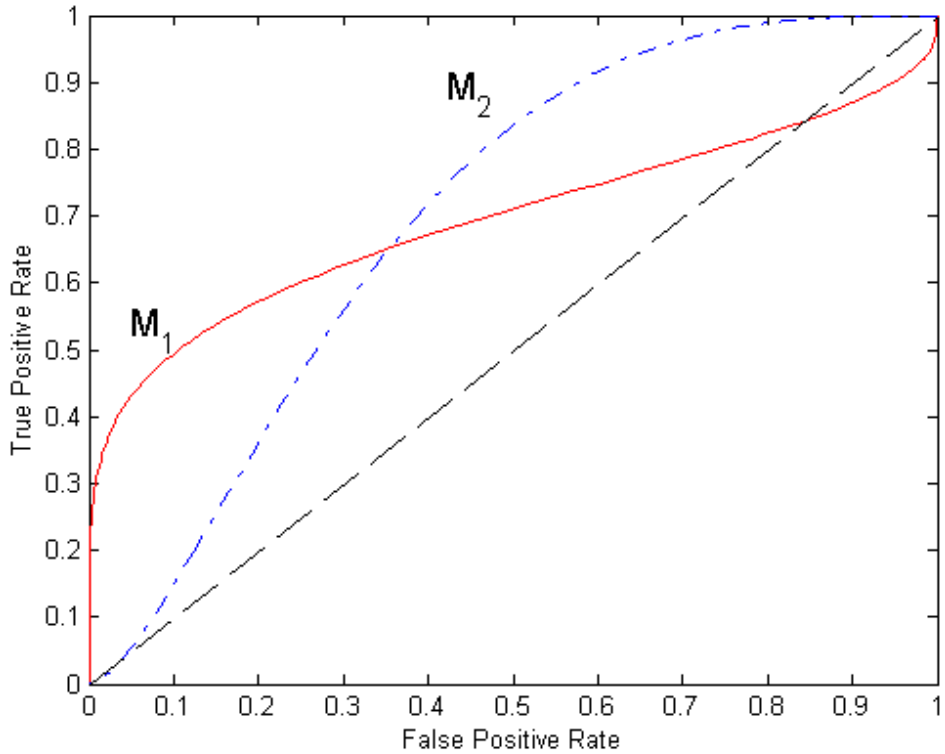



- En el umbral t TP=0.5 y FP=0.21

# Podemos comparar visualmente modelos usando la curva ROC

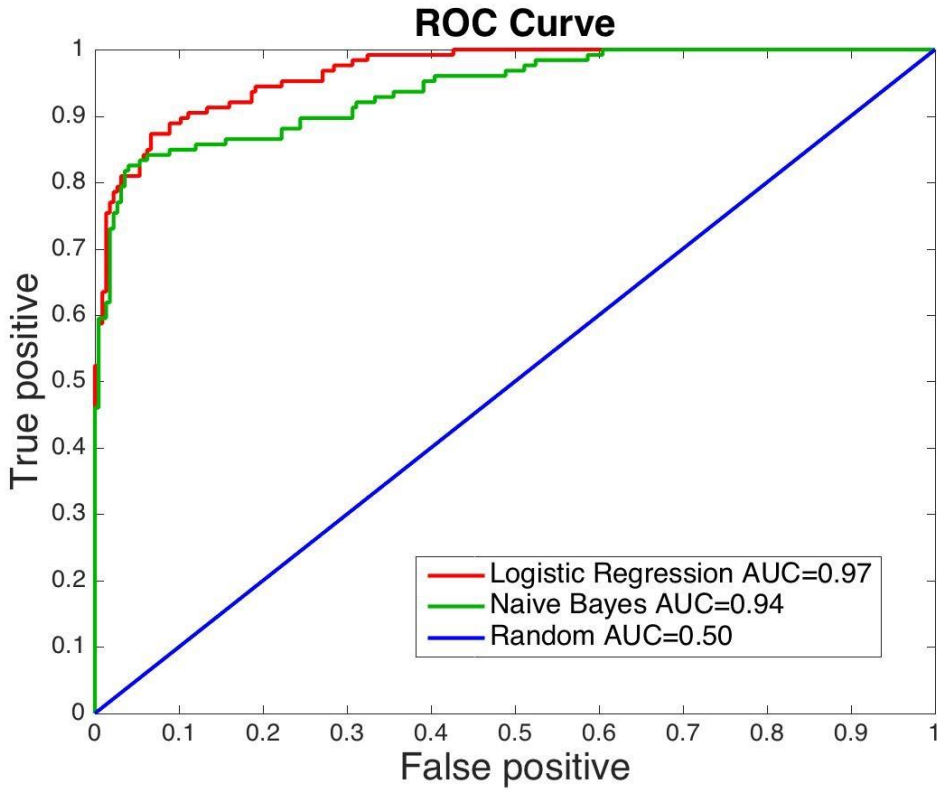


# AUC

Área bajo curva (AUC): es el área debajo de la curva ROC y sintetiza el rendimiento del modelo.

También podemos comparar modelos basándonos en el AUC



# Evaluación de modelos supervisados

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2