

Minería de Datos

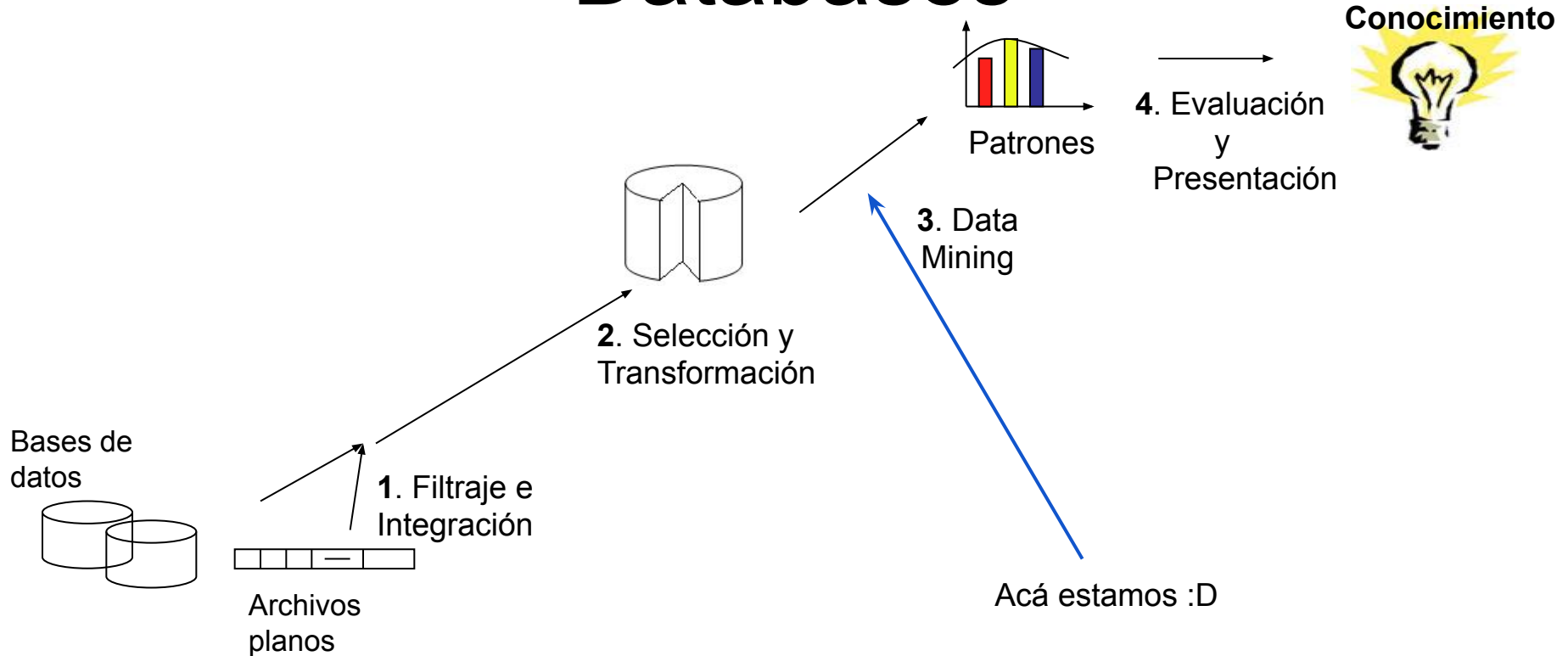
IIC2433

Modelos de Clasificación
Regresión Logística
Vicente Domínguez

¿Qué veremos esta clase?

- ¿Qué es el aprendizaje automático (machine learning)?
- Clasificación
- Modelo de Regresión Logística

Knowledge Discovery in Databases



Aprendizaje de máquina

(Machine Learning)

Darle a los computadores la habilidad de realizar una actividad, sin programarlos explícitamente.

*La minería de datos y el **aprendizaje de máquina** se traslapan y no tienen límites claros

Programación tradicional (explícita)

Kasparov vs. Deep Blue (1997)



Aprendizaje de máquina

Lee Sedol vs. AlphaGo (2016)



Google DeepMind

Aprendizaje de máquina

Tipos de tareas

- Aprendizaje supervisado
 - Clasificación
 - Regresión
- Aprendizaje no supervisado
 - Clustering
 - Aprendizaje por refuerzo
 - etc

Aprendizaje supervisado

Clasificación

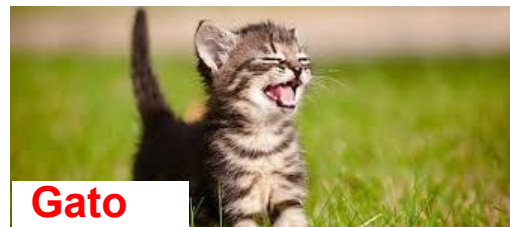
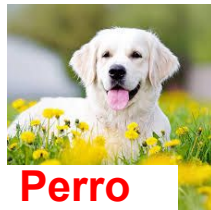
Tarea para el computador:

Decir si en una foto hay un perro o un gato

Aprendizaje supervisado

Clasificación

Conjunto de entrenamiento **etiquetado**



Aprendizaje supervisado

Clasificación

¿Qué es eso?



Perro

Aprendizaje no supervisado

Clustering

Tarea para el computador:

Identificar grupos de elementos similares

Aprendizaje no supervisado

Clustering

Conjunto de datos **no etiquetados**



Aprendizaje de máquina

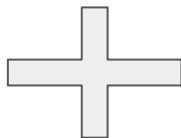
Tipos de tareas

- Aprendizaje supervisado **(necesita etiquetas)**
 - Clasificación
 - Regresión
- Aprendizaje no supervisado **(no necesita etiquetas)**
 - Clustering
 - Aprendizaje por refuerzo
 - etc

Clasificación

1. *Etapa de entrenamiento*

Conjunto de
entrenamiento



Algoritmo de
clasificación



**Clasificador
entrenado**

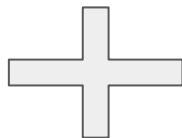
Dato	Etiqueta
	Gato
	Perro
	Perro
	Gato
	Gato
	Perro

- Árboles de decisión
- Naïve Bayes
- KNN
- SVM
- etc...

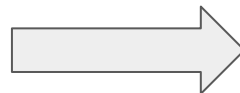
Clasificación

2. *Etapas de clasificación*

Datos no
etiquetados



**Clasificador
entrenado**



Clases
(Etiquetas)

que obtuvimos en la
etapa 1

Dato	Etiqueta
	?
	?
	?
	?
	?
	?

Dato	Etiqueta
	Perro
	Gato
	Perro
	Gato
	Gato
	Perro

Calendario

Martes	Jueves	Clase Martes - 1	Clase Martes - 2	Ayudantía Jueves
8-mar	10-mar	Intro Administrativo, El problema de aprender	Manejo de datos	Ayudantía Pandas y librerías
15-mar	17-mar	Regresión Lineal	Regresión Lineal	
22-mar	24-mar	Regresión Logística	Regresión Logística	Ayudantía Tarea
29-mar	31-mar	Clasificadores - Evaluación	KNN	
5-abr	7-abr	Árbol de decisión	Árbol de decisión	
12-abr	14-abr	Ensamble	Random Forest, Gradient Boosted Trees	FERIADO (JUEVES SANTO)

Regresiones Lineales

- Técnica estadística donde se trata de ajustar parámetros de una función lineal sobre un conjunto de datos.
- Se busca predecir el valor de una variable dependiente cuantitativa (predicha) utilizando variables independientes (predictores)
- Finalmente, queremos determinar cómo afecta nuestra variable independiente a la dependiente

$$Y = \alpha + \beta X$$

¿Cómo podemos utilizar una regresión lineal como un clasificador?

- ¿Hay alguna propiedad o modelamiento que debemos hacer en ella?
- ¿Alguna idea?
- ¿Qué valores debería tener Y ?

$$Y = \alpha + \beta X$$

Regresión Logística

- Se puede ajustar una regresión para cada clase
- Luego, cambiamos el valor de Y de cada instancia por:
 - $Y = 1$ si pertenece a la clase
 - $Y = 0$ si no pertenece
- Ahora, si me llega un valor nuevo:
 - Calculo el valor predicho por cada regresión.
 - El valor más alto obtenido por una regresión me dirá la clase que predeciré de dicha instancia

Regresión Logística

- ¿Basta sólo esto?
- ¿Está acotado el output de la regresión lineal a valores entre 0 y 1?
- ¿Cómo distribuye el valor del Y predicho?

Regresión Logística

Las siguientes slides están basadas en las del profesor Mauricio Arriagada

Regresión Logística

- Modelamos la salida del clasificador deseado como una función de probabilidad

$\text{class}(X) = 1$



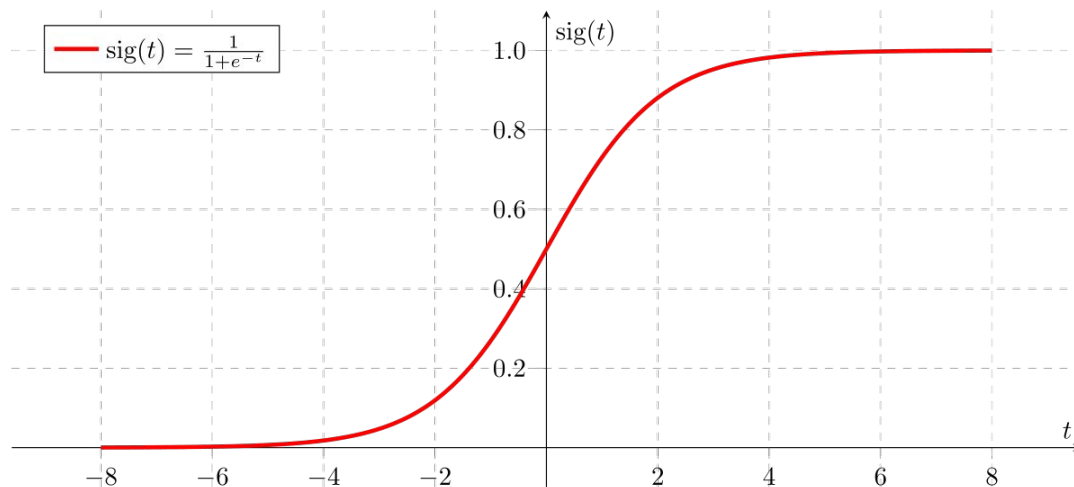
$P(\text{class} = 1 \mid X)$

$\text{Class}(X) = 0$

$P(\text{class} = 0 \mid X)$

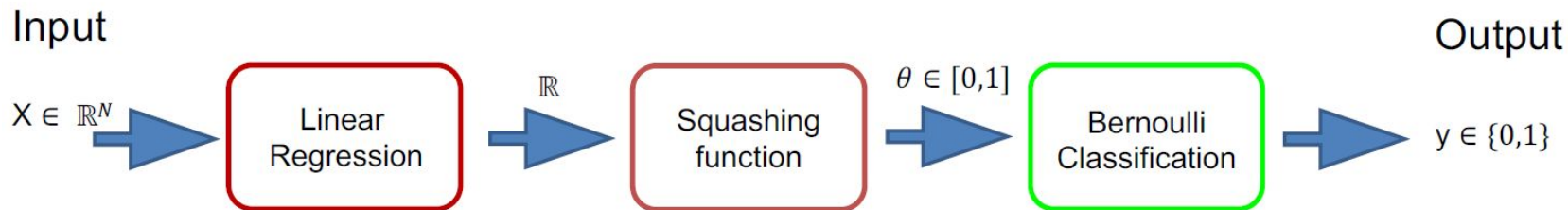
Regresión Logística

- Si reducimos la salida de una regresión lineal al intervalo $[0,1]$, podríamos usar esa salida como $P(Y = y)$



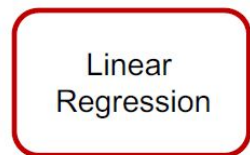
Regresión Logística

- Si reducimos la salida de una regresión lineal al intervalo $[0,1]$, podríamos usar esa salida como $P(Y = y)$

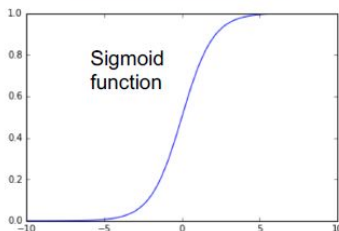
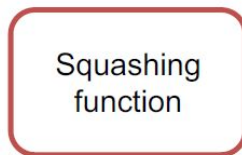


Regresión Logística

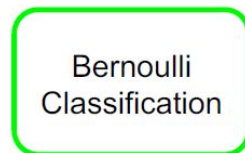
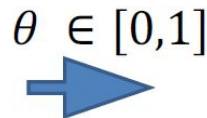
- Si reducimos la salida de una regresión lineal al intervalo $[0,1]$, podríamos usar esa salida como $P(Y = y)$



$$\beta_0 + \sum_{j=1}^d \beta_j x_j$$



$$f_{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$



$$P(Y = y|\theta) = \theta^y (1 - \theta)^{1-y}$$

Regresión Logística

- Otro punto de vista: estamos haciendo una regresión sobre las probabilidades (log odds)
- Log odds: log de la proporción de obtener un "éxito" (codificado como 1) sobre obtener "fracaso" (codificado como 0)
- Entonces, lo que realmente estamos haciendo es una regresión en log odds:

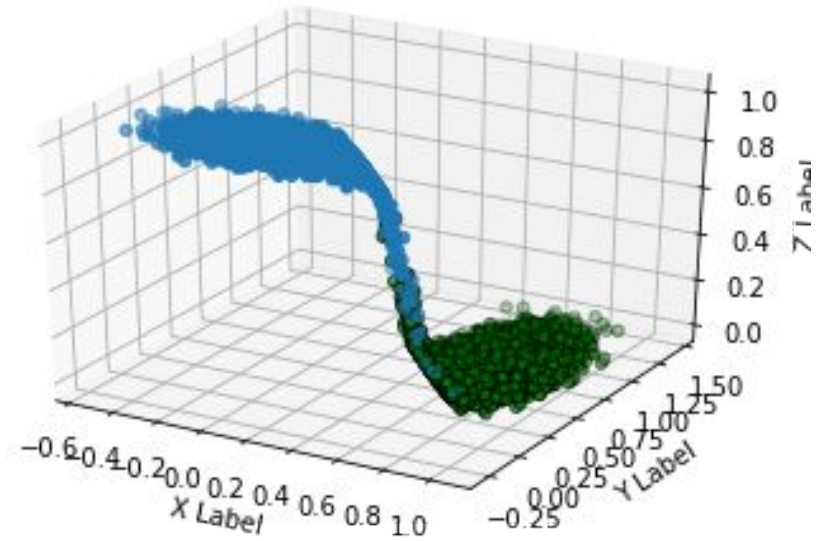
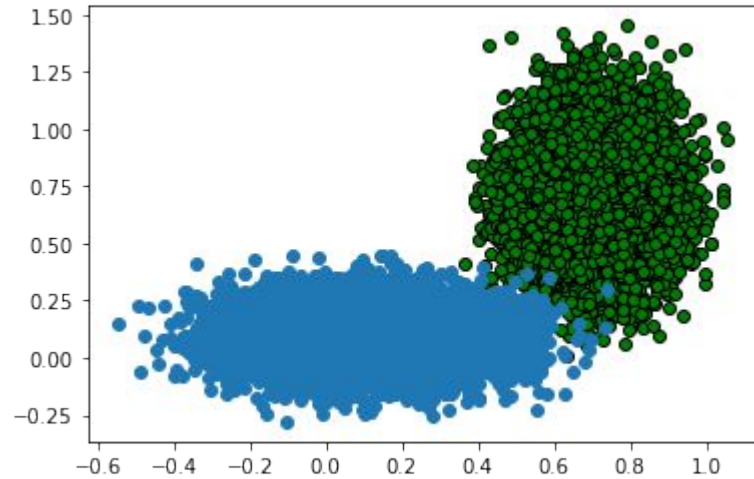
$$\log \frac{\theta}{(1 - \theta)} = \beta_0 + \sum_{j=1}^d \beta_j x_j$$

(Logit)

$$\theta = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^d \beta_j x_j)}}$$

(Sigmoid)

Regresión Logística



Regresión Logística

- Finalmente, lo que se busca es un W (pesos) tal que se optimice

$$\max P(Y|\theta)$$

$$\theta = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^d \beta_j x_j)}}$$

Descenso del gradiente

- A pesar de que con las regresiones lineales podemos encontrar los parámetros de forma analítica, eso muchas veces no es el caso.
- La mayoría de las veces se utiliza un método de descenso de gradiente.