Minería de Datos
IIC2433

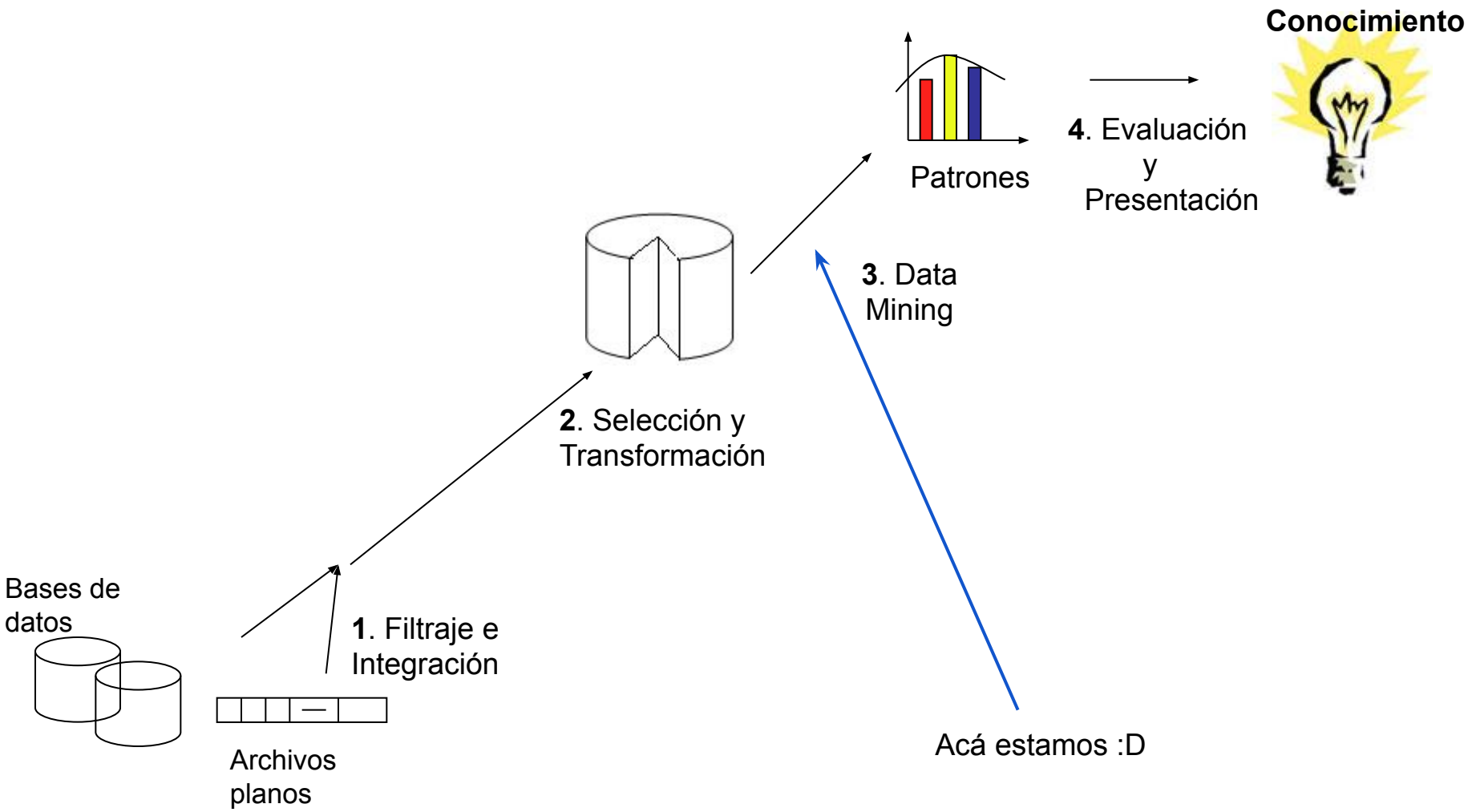
Modelos de Clasificación
Regresión Logística

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)

IBM®

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**


Perro


Perro


Gato


Gato


Perro


Gato

Aprendizaje supervisado

Clasificación

¿Qué es eso?

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

Calendario

Fecha semana	Martes	Jueves	Clase Martes - 1	Clase Martes - 2	Ayudantía Jueves / Control	Enunciados
10-ago.	11-ago.	13-ago.	Intro Administrativo		Clase - Data Warehouse - OLAP	
17-ago.	18-ago.	20-ago.	Web Scrapping	Actividad	Control Data WH	
24-ago.	25-ago.	27-ago.	Data Prep	Pandas	Ayudantía Pandas y librerías	
31-ago.	1-sept.	3-sept.	Association Rules	Association Rules	Ayudantía Association Rules	Tarea 1
7-sept.	8-sept.	10-sept.	PCA	Actividad	Control AR	
14-sept.	15-sept.	**17-sept.**	Regresiones	Actividad	**Feriado**	
21-sept.	**22-sept.**	**24-sept.**	**Semana Receso**	**Semana Receso**	**Semana Receso**	
28-sept.	29-sept.	1-oct.	Reg log	Actividad	Control PCA y Reg	
5-oct.	6-oct.	8-oct.	KNN	Árboles de Decision	Ayudantía Knn y Árbol de Decisión	Tarea 2

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 debamos 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

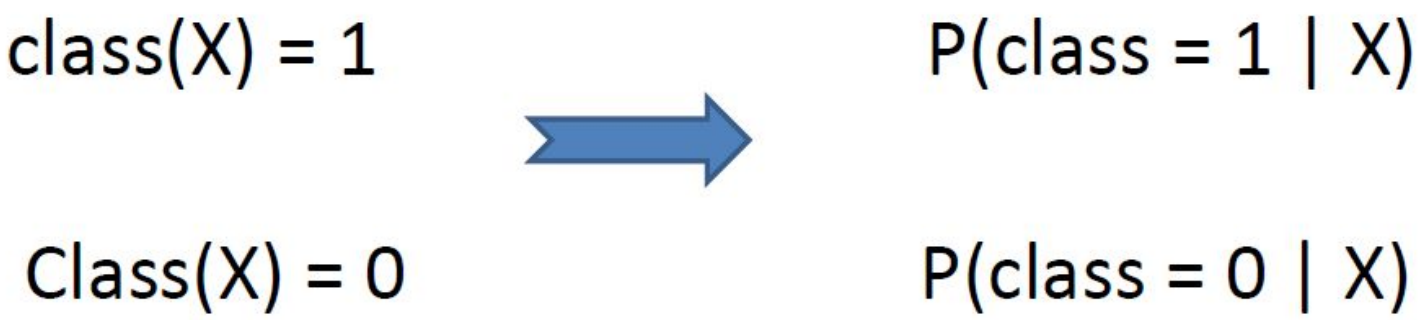
- ¿Basta sólo esto?
- ¿Está acotado el output de la regresión lineal a valores entre 0 y 1?
- ¿Cómo distribuye el valor del Y predicho?

Regresión Logística

Las siguientes slides están basadas en las del profesor Mauricio Arriagada

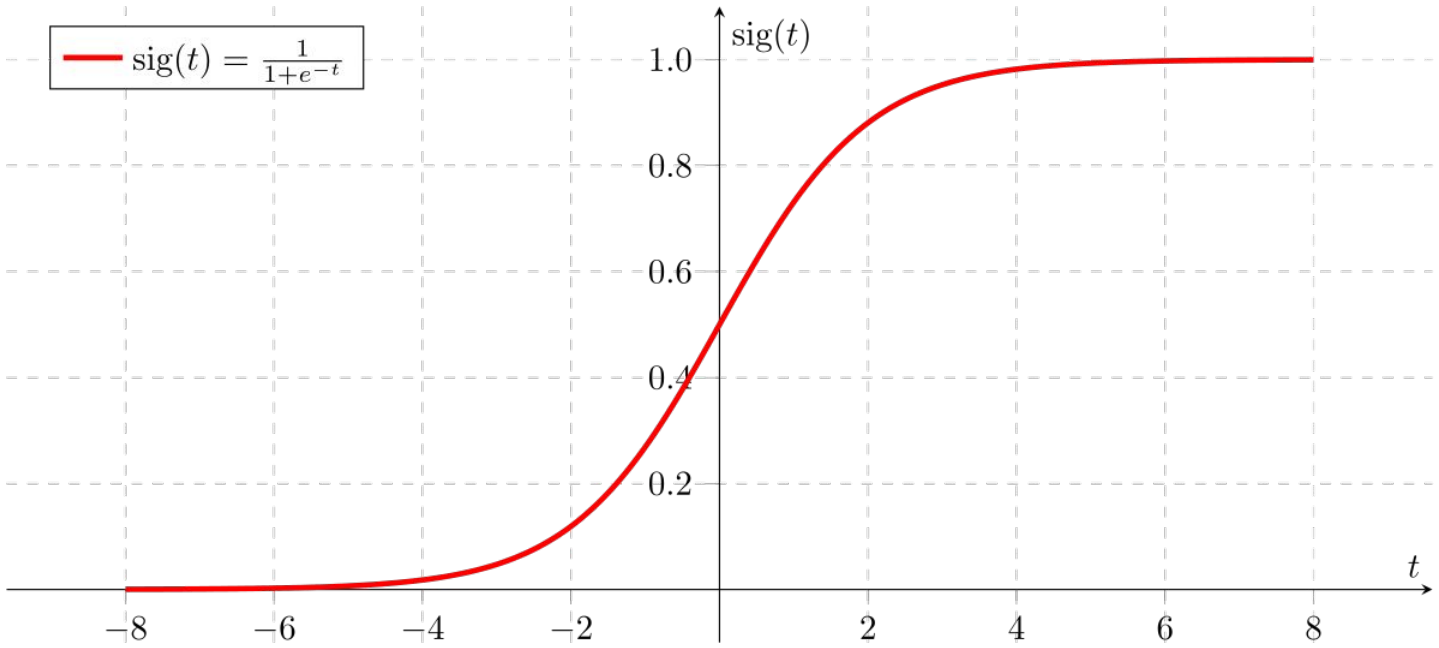
Regresión Logística

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



Regresión Logística

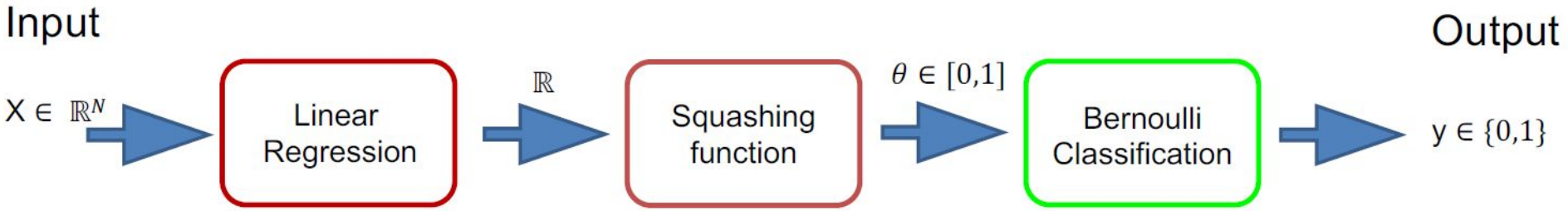
- Si reducimos la salida de una regresión lineal al intervalo [0,1], podríamos usar esa salida como P (Y = y)



https://miro.medium.com/max/1400/1*RqXFpiNGwdiKBWyLJc_E7g.png

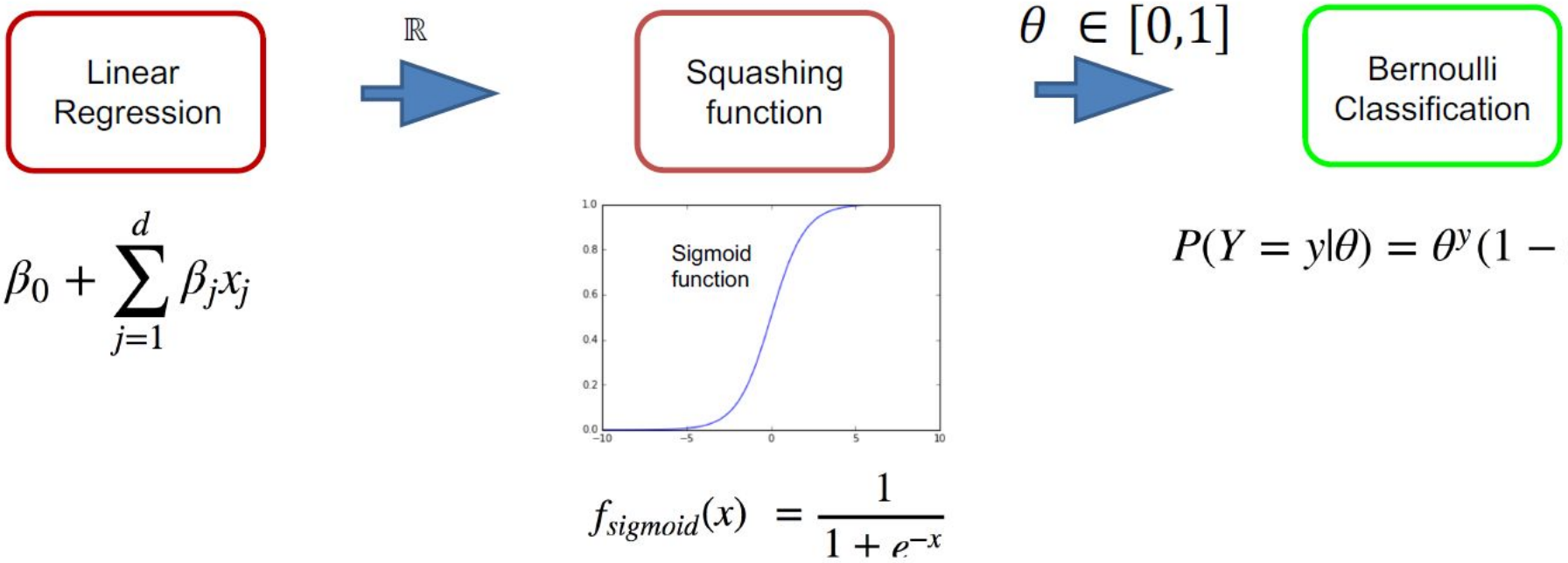
Regresión Logística

- 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

- 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)

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

(Sigmoide)

Regresión Logística

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

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

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