Minería de Datos
IIC2433

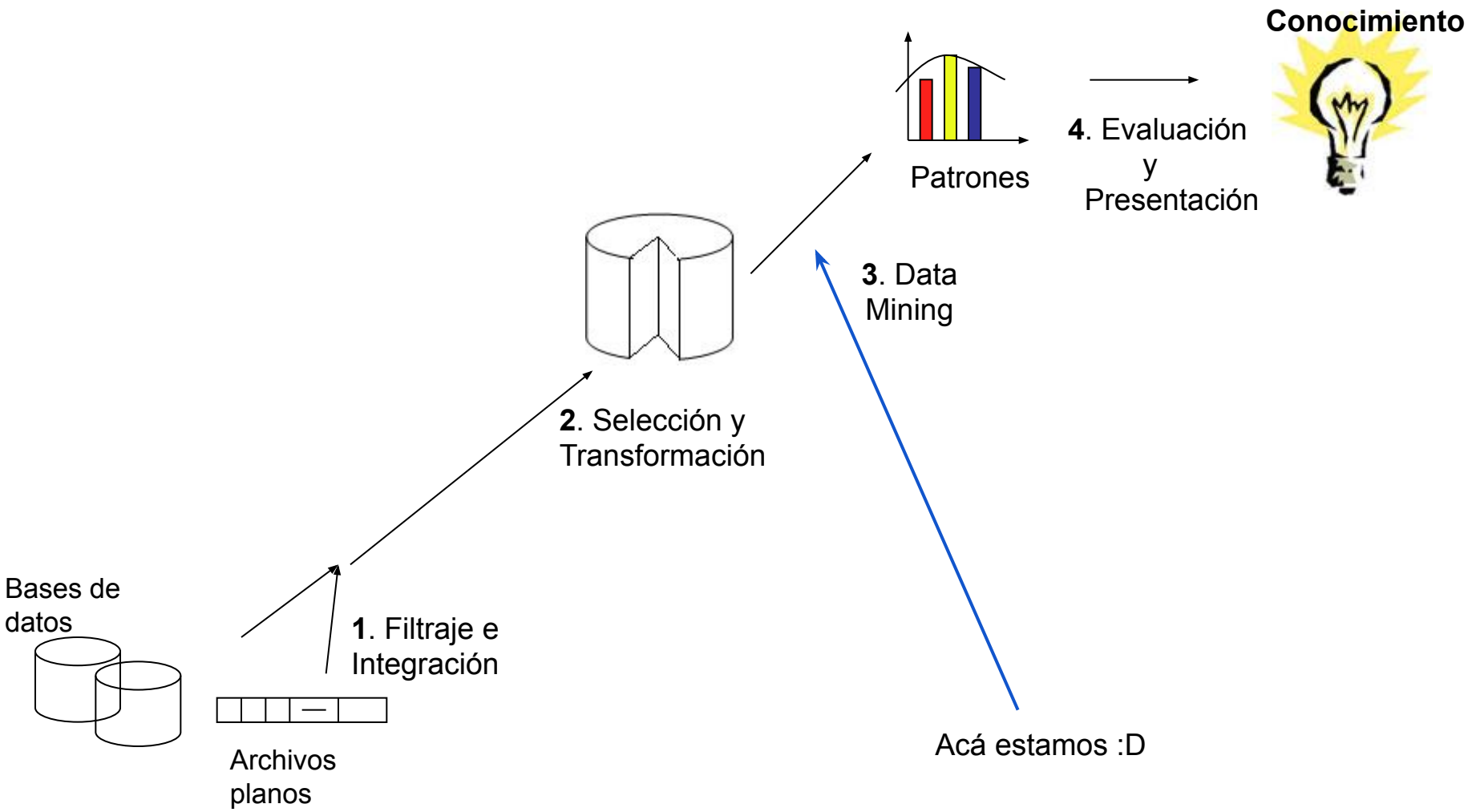
Modelos de Regresión

Vicente Domínguez

¿Qué veremos esta clase?

- Modelos de Regresión
- Cómo predecir una variable numérica

Knowledge Discovery in Databases



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

Métodos de aprendizaje

- Aprendizaje supervisado
  - Necesita conocimiento previo del problema y el valor a predecir
  - Se pueden usar valores numéricos o etiquetas
- Aprendizaje no supervisado
  - No se necesita conocimiento previo
  - Modelos buscan patrones dentro del conjunto de datos

Métodos de aprendizaje

- Aprendizaje supervisado **(necesita etiquetas)**
  - Clasificación
  - **Regresión**

- Aprendizaje no supervisado **(no necesita etiquetas)**
  - Clustering
  - Reglas de asociación
  - etc

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

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

Regresiones Lineales

Condiciones y supuestos

Para que un modelo de regresión lineal funcione correctamente, deben cumplirse ciertas condiciones.

- Homocedasticidad
- Independencia
- Normalidad

Aparte también hay medidas que nos permiten evaluar que tan bien ajusta el modelo:

- R^2 o varianza explicada
- RSS o suma de errores residuales

Ordinary Least Squares

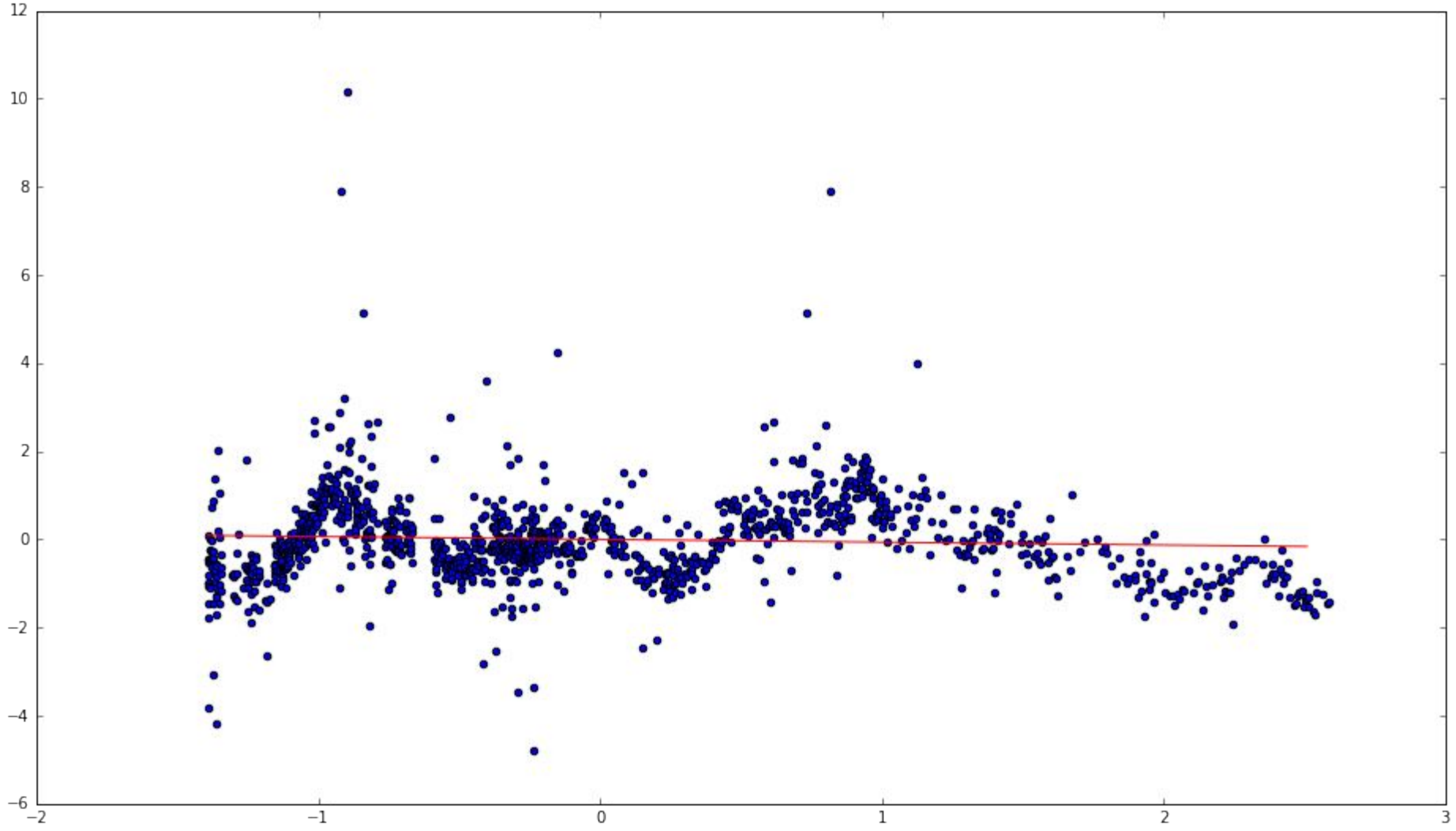
Busca minimizar el error cuadrático residual RSS

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

$$\hat{\beta} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} = \frac{\operatorname{Cov}[x,y]}{\operatorname{Var}[x]}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\, \overline{x}$$

Modelo de regresión simple



Regresiones no lineales

- Siempre se puede hacer un cambio de *Kernel* sobre el conjunto de datos.
- Es decir, se puede hacer una transformación con combinaciones lineales, de funciones no lineales sobre el conjunto de datos

$$f(X) \rightarrow f(\Phi(X))$$

Donde $\Phi(X)$ es una transformación sobre X, por ejemplo

$$\Phi(X) \rightarrow [X, X^2, X^4]$$

Regresiones no lineales

- Ajustar una regresión lineal no significa **ajustar una recta**.
- Si ajustamos una línea sobre un espacio curvo nos debería ajustar una curva.
- En particular también podemos ajustar cualquier función no lineal como *Kernel*, por ejemplo una Gaussiana

$$f(X) = e^{-\frac{1}{2}\left(\frac{X-\mu_i}{\sigma}\right)^2}$$

Regresiones no lineales