DEPARTAMENTO
DE COMPUTACION

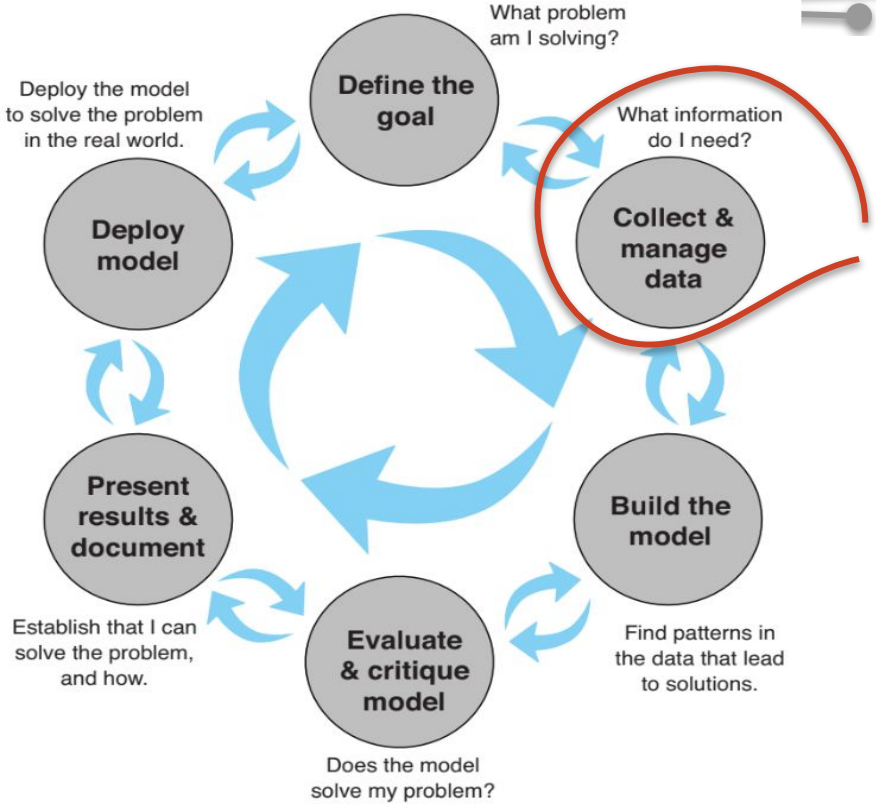
Facultad de Ciencias Exactas y Naturales - UBA

Laboratorio de datos

Reducción de la Dimensión

Primer Cuatrimestre 2024

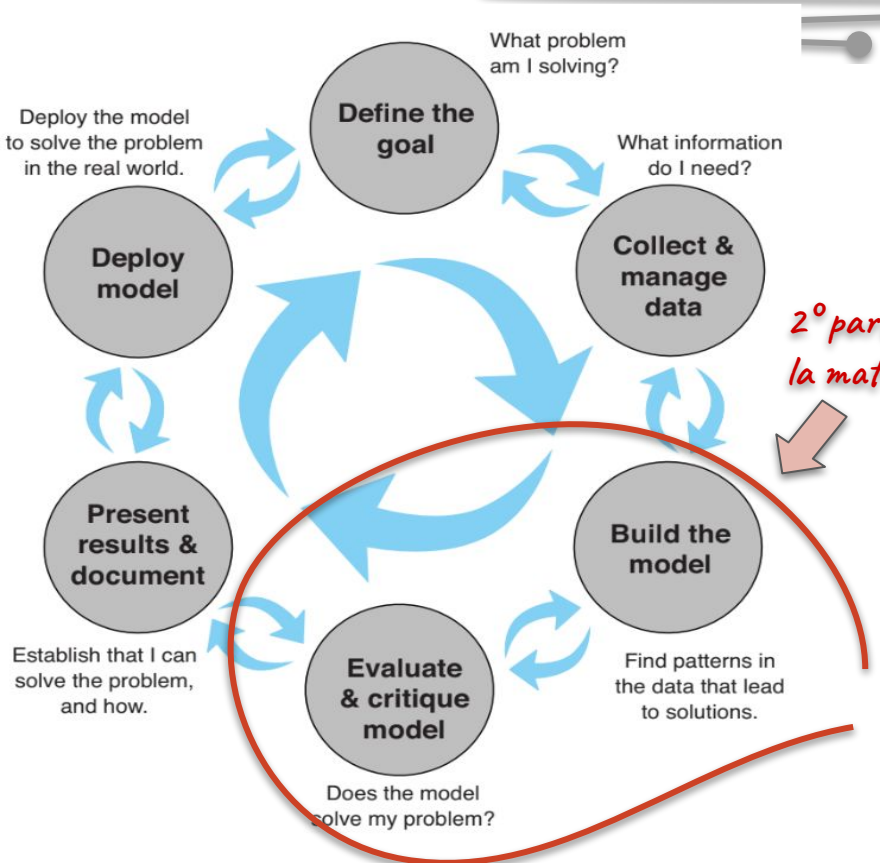
Recorrido de la materia (hasta ahora)



1° parte de la materia

- ✓ Lenguaje de programación (Python)
- ✓ Modelado conceptual de los datos (DER)
- ✓ Representación de los datos (modelo relacional)
- ✓ Formas de consultar los datos (AR/SQL)
- ✓ Recomendaciones para el diseño (Normalización)
- ✓ Calidad de datos
- ✓ Leyes acerca de la Protección de Datos

Recorrido de la materia (hasta ahora)



2° parte de la materia

- ✓ Visualización y Exploración de los datos
- ✓ Intro a Modelado: Clasificación y Regresión
- ✓ Clasificación: Árboles de decisión
- ✓ Evaluación y selección de modelos
- ✓ Regresión y KNN
- ✓ Modelos Lineales en Regresión
- ✓ No supervisado: Clustering

Herramientas de aprendizaje no supervisado

Clustering - Agrupamiento

Métodos para encontrar subgrupos homogéneos dentro del conjunto entero de los datos.

Reducción de dimensionalidad

Métodos para proyectar los datos -en general de dimensiones altas- en un espacio de menor dimensión, que haga posible su manipulación (o visualización) pero preserve las características del conjunto original.

Suele usarse también como paso previo al clustering.

Reducción de la dimensión

Objetivos

- Visualización
- Interpretación de los datos
- Regularización de los datos
- Simplificación de los modelos a utilizar

Reducción de la dimensión

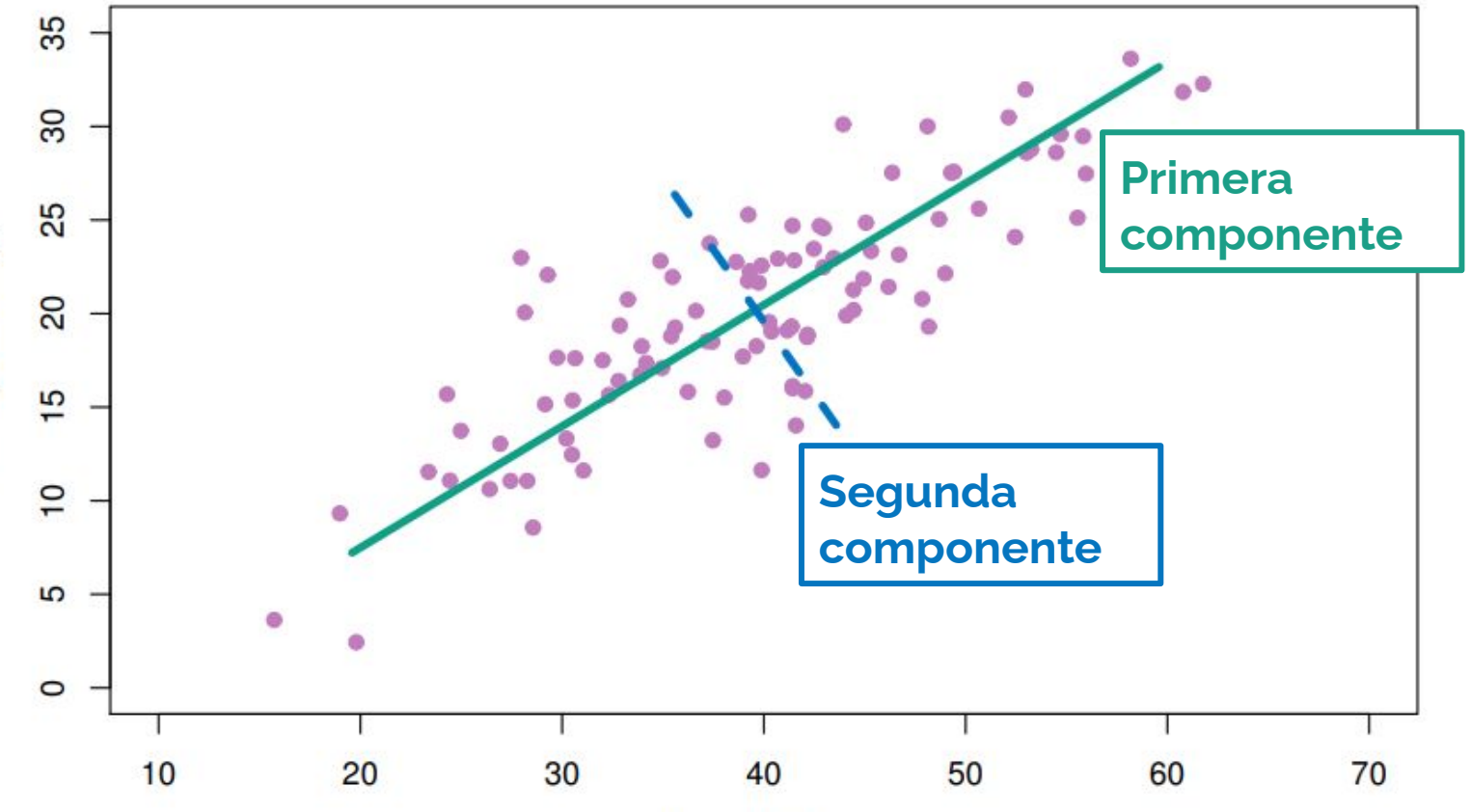
Técnicas (hay más)

- PCA: Análisis de Componentes Principales
- MDS: Multidimensional Scaling
- ISOmap: Isometric Feature Mapping
- t-SNE: t-Stochastic Neighbor Embedding

PCA - Principal Component Analysis

A partir de las variables originales, se construyen **combinaciones lineales**. Se buscan las direcciones que maximizan la variabilidad de los datos.

Se basa en la idea de que los datos, si bien se encuentran en cierto espacio n-dimensional, están mayormente dentro de un **subespacio** de menor dimensión.


Primera componente
Segunda componente
0
5
10
15
20
25
30
35
10
20
30
40
50
60
70

Si tenemos p variables, la primera componente principal (PC1) será una combinación lineal de la forma:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$$

donde los coeficientes están normalizados, es decir:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

y se elige de manera de maximizar la varianza. Dada una muestra i-ésima en particular, su proyección sobre la componente PC1 será:

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{p1} x_{ip}$$

Los coeficientes de PC1 definen la dirección sobre la cual los datos varían más.

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

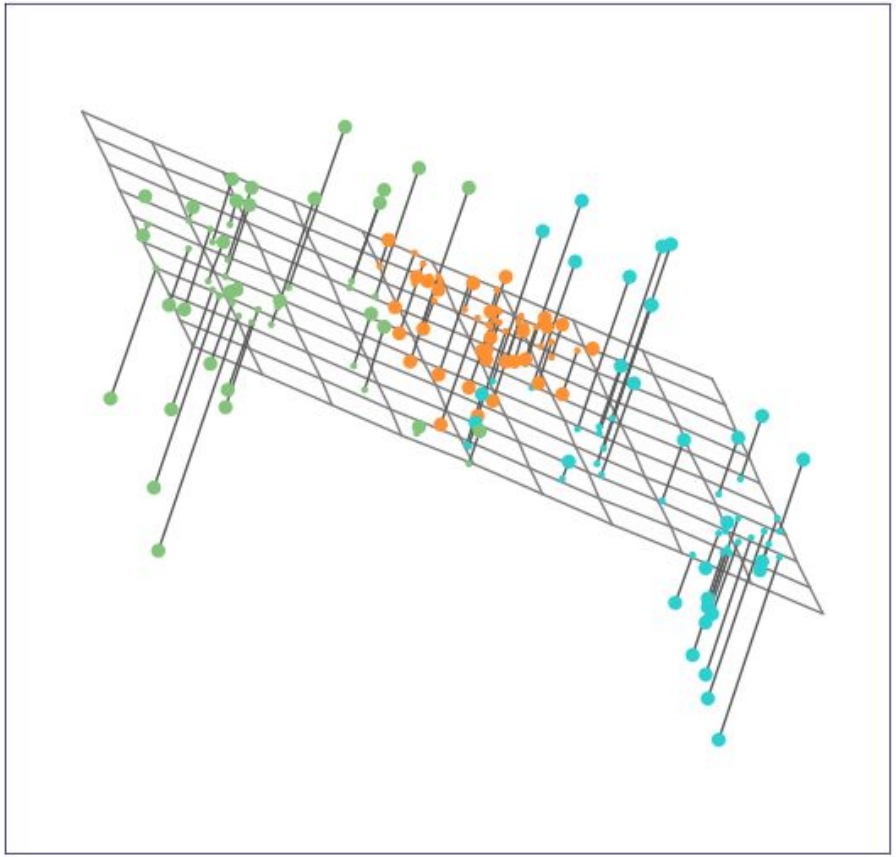
La segunda componente, PC2, es la dirección de **mayor varianza**, dentro de las direcciones **ortogonales** a PC1.

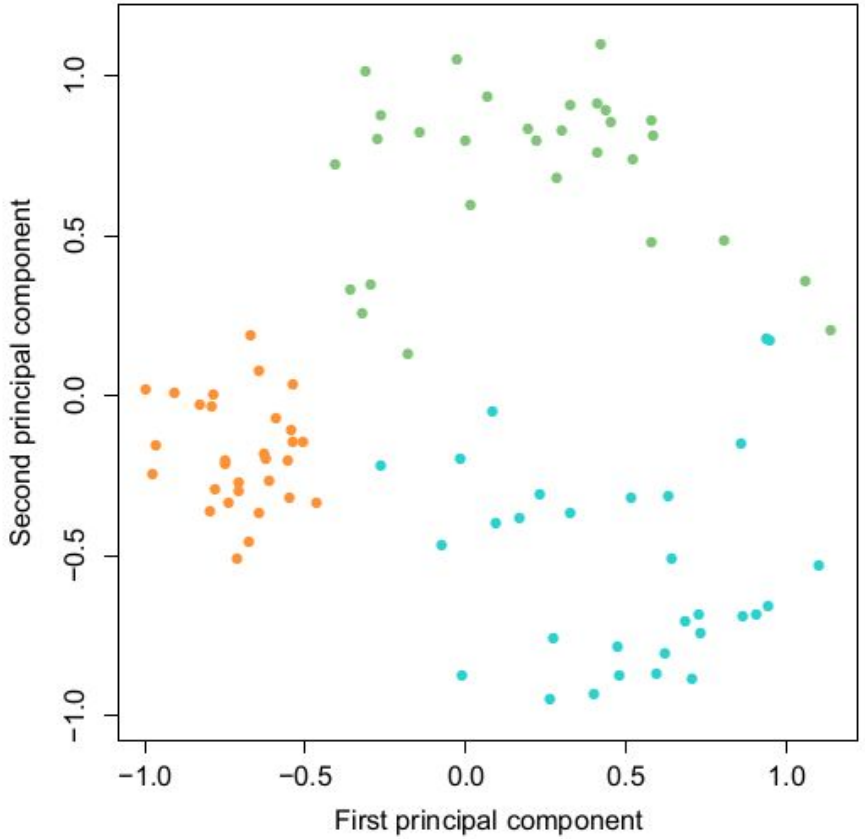
Así, hasta la última componente componente (tantas como variables).

Las direcciones de las componentes principales generan un subespacio que se acerca a los datos.

Por ejemplo, <PC1, PC2> representa el plano que está más cerca de los puntos (en términos de la distancia euclídea).

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p}\sum_{i=1}^{n} \left(x_{ij} - \sum_{m=1}^{M} a_{im}b_{jm} \right)^2 \right\}$$




Second principal component
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
First principal component

Repaso - Varianza muestral

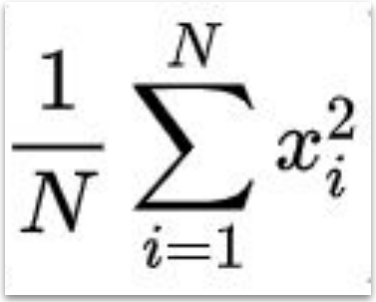
Vimos que la varianza muestral, para una muestra de la variable x, con n instancias x_i, se calcula como:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(o a veces dividiendo por n-1).

Repaso - Varianza muestral

Vamos a suponer que el promedio cero. De modo que la varianza muestral va a ser:



Queremos construir una dirección, que llamaremos Z, que maximice la varianza muestral. La vamos a escribir como combinación de las variables X.

$$Z = \Phi_1 X_1 + \Phi_2 X_2 + \ldots + \Phi_n X_n$$

(donde los coeficientes deben estar normalizados, es decir, sus cuadrados deben sumar 1).

Vamos a querer elegir los coeficientes de manera que Z tenga la máxima varianza muestral posible. ¿Cómo son las “muestras” en Z?

Siendo Z la combinación lineal:

$$Z = \Phi_1 X_1 + \Phi_2 X_2 + \ldots + \Phi_n X_n$$

entonces, para cada i, la muestra i-ésima de Z es:

$$Z_i = \Phi_1 X_{i1} + \Phi_2 X_{i2} + \ldots + \Phi_n X_{in}$$

En términos matriciales, X tiene n columnas, una por variable (feature) y tiene una fila por cada muestra i.

Entonces la varianza muestral de Z, será:

$$(1/n)\sum Z_i^2 = (1/n)[(\mathbf{\Phi}_1 X_{i1})^2 + (\mathbf{\Phi}_2 X_{i2})^2 + \ldots + (\mathbf{\Phi}_n X_{in})^2]$$

Varianza explicada

¿Cuánta información se preserva? ¿Cuánta se pierde?
¿Cómo lo calculamos?

Podemos considerar la proporción de varianza explicada, PVE, es decir cuánta varianza explican las componentes, sobre la varianza total.

Varianza total:

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

Varianza explicada por PCm:

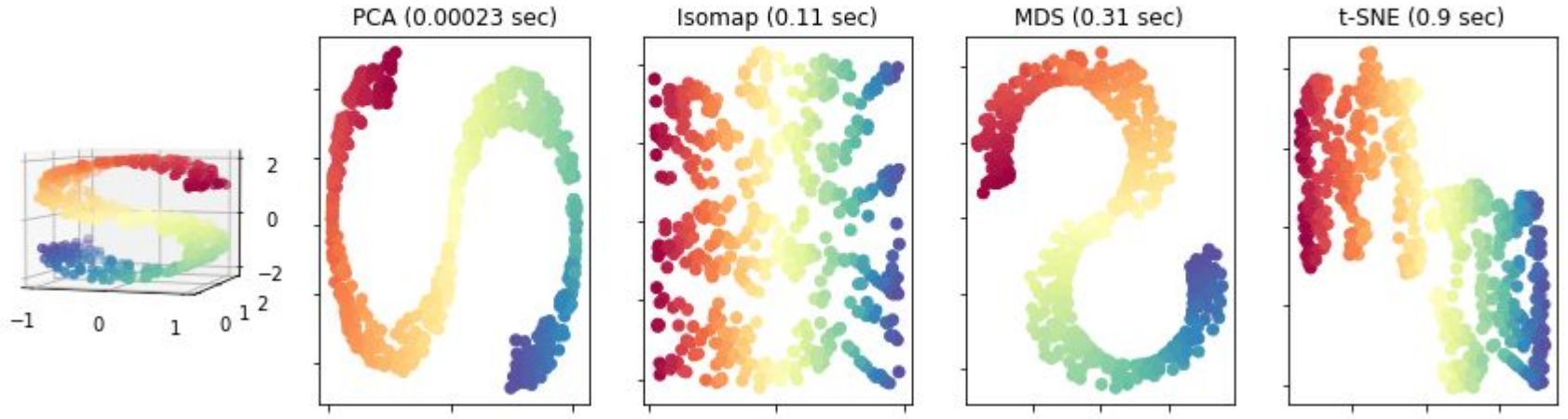
$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

PVE de PCm

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} = \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

También se vincula con la distancia al subespacio generado por las componentes principales.

$$\underbrace{\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2}_{\text{Var. of data}} = \underbrace{\sum_{m=1}^{M} \frac{1}{n} \sum_{i=1}^{n} z_{im}^2}_{\text{Var. of first } M \text{ PCs}} + \underbrace{\frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_{ij} - \sum_{m=1}^{M} z_{im} \phi_{jm} \right)^2}_{\text{MSE of } M\text{-dimensional approximation}}$$


Comparación de métodos
PCA (0.00023 sec)
Isomap (0.11 sec)
MDS (0.31 sec)
t-SNE (0.9 sec)
2
0
−2
−1
0
1
0
1
2

PCA con scikit-learn

Cierre de Aprendizaje Automático

Modelos

Modelo: Cómo planeo modelar el proceso.
Por ejemplo, ¿árbol de decisión o knn? ¿Regresión lineal o polinomial?

Hiperparámetros: Valores que se especifican previo a realizar el aprendizaje, es decir determinan el proceso de aprendizaje automático.

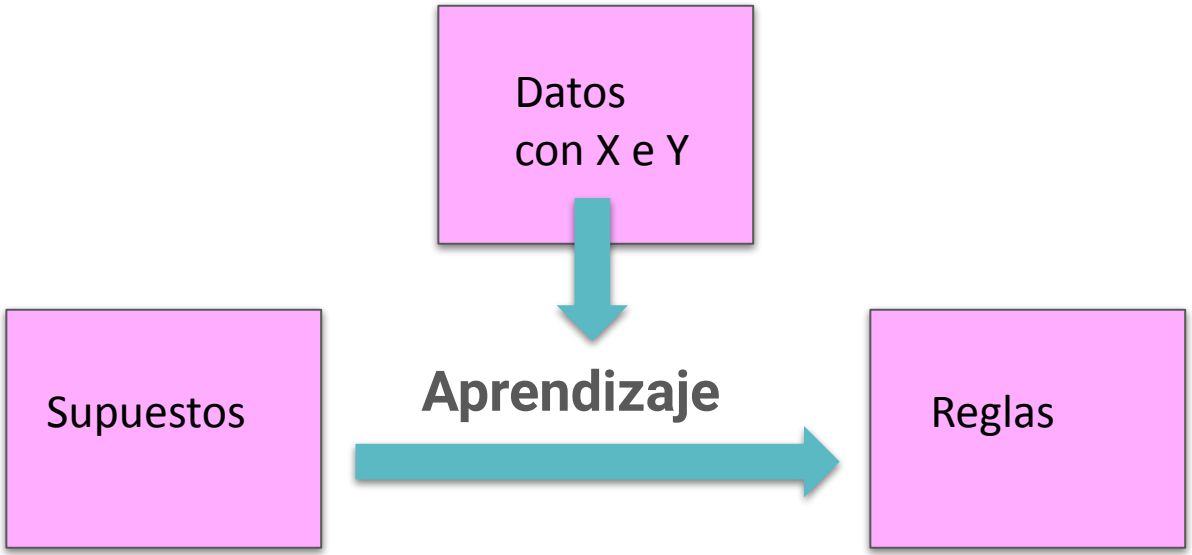
Por ejemplo, en árboles: máxima profundidad permitida, criterio de elección de atributos en cada nodo (Gini Gain, Information Gain), etc. O el valor de k en knn.

Modelo entrenado o Parámetros: Valores internos del modelo resultante **luego del aprendizaje**. Representan las **reglas aprendidas**. Representan el algoritmo final, que es un producto obtenido en el aprendizaje automático.

Por ejemplo, en árboles: las preguntas a realizar en cada nodo del árbol y la clase asignada a cada hoja del árbol.

Aprendizaje supervisado

A partir de un modelo basado en algunos supuestos, se entrena un algoritmo de manera de construir las reglas a partir de la observación de datos etiquetados.



Clasificación vs. Regresión

En clasificación buscamos explicar una variables que es categórica.

- True - False
- Setosa - Versicolor - Virginica
- Sobrevivió - No sobrevivió
- Ceibo - Pindó - Eucaliptus - Jacarandá

En regresión buscamos explicar una variable que es contínua (puede tomar valores en R o en Z)

- Altura de una persona
- Precio de una propiedad
- Temperatura

Clasificación vs. Regresión

Algunos modelos para clasificación

- Umbral
- Árboles de decisión
- knn

Métricas para clasificación

- Exactitud/accuracy
- Precisión/precision
- Exhaustividad/Recall
- F1

Algunos modelos para regresión

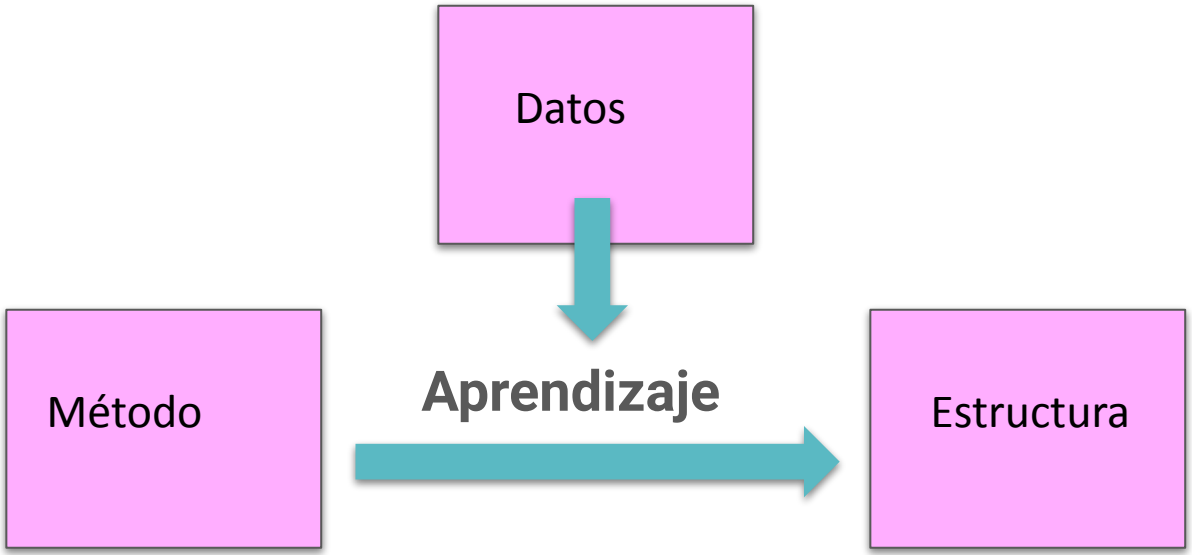
- Modelos lineales
- knn

Métricas para regresión

- R^2
- MSE

Aprendizaje no supervisado

A partir de un modelo o método, se buscan patrones o estructura en los datos. No se cuenta con datos etiquetados.



Aprendizaje no supervisado

En clustering buscamos agrupar los datos en fragmentos homogéneos.

Métodos:

- K-means
- DBSCAN
- Jerárquico

Medidas:

- WCSS

En reducción de la dimensión buscamos resumir información de los datos.

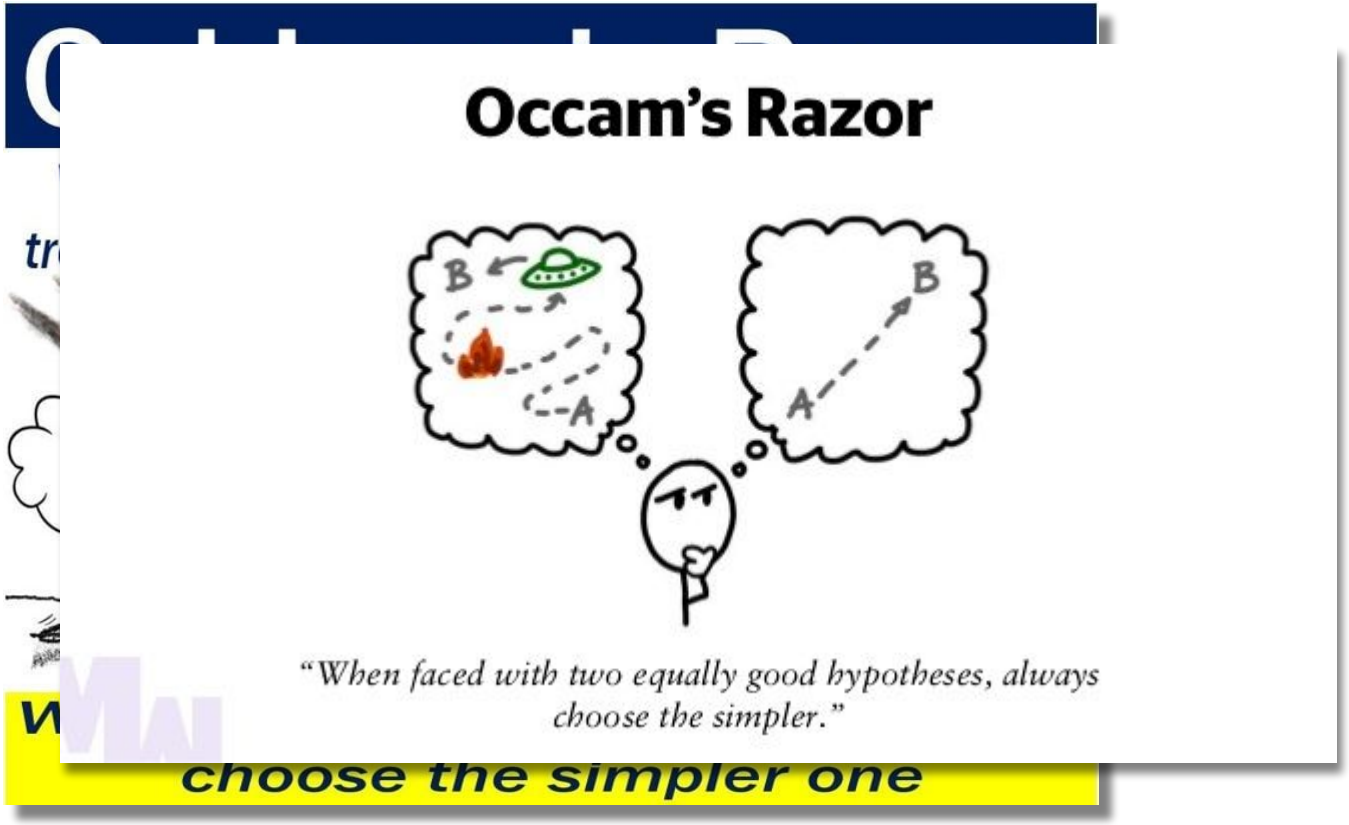
Métodos:

- PCA

Medidas:

- Variabilidad explicada

Principio de la Navaja de Ockham



Principio de la Navaja de Ockham

Ante dos explicaciones posibles, elegiremos la más sencilla.

Cuando tenemos dos modelos que compiten, y que tienen una predicción similar, debemos elegir el más simple. Es decir, con menos parámetros o menor complejidad.

- Evita el sobreajuste y por lo tanto mejora el rendimiento en datos no vistos.
- Mejora la interpretación: Los modelos más simples son más fáciles de interpretar y entender.
- Reduce los costos computacionales.