Laboratorio de Datos

Aprendizaje No Supervisado

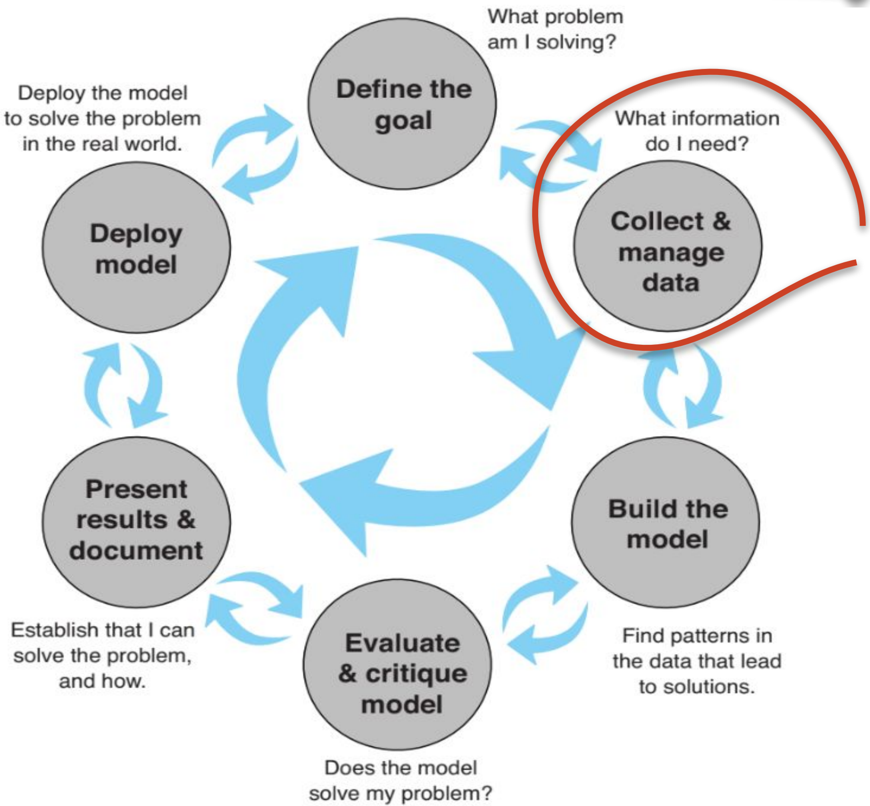
Verano 2024

UNIVERSIDAD · DE · BUENOS · AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES

DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

exactas UBA

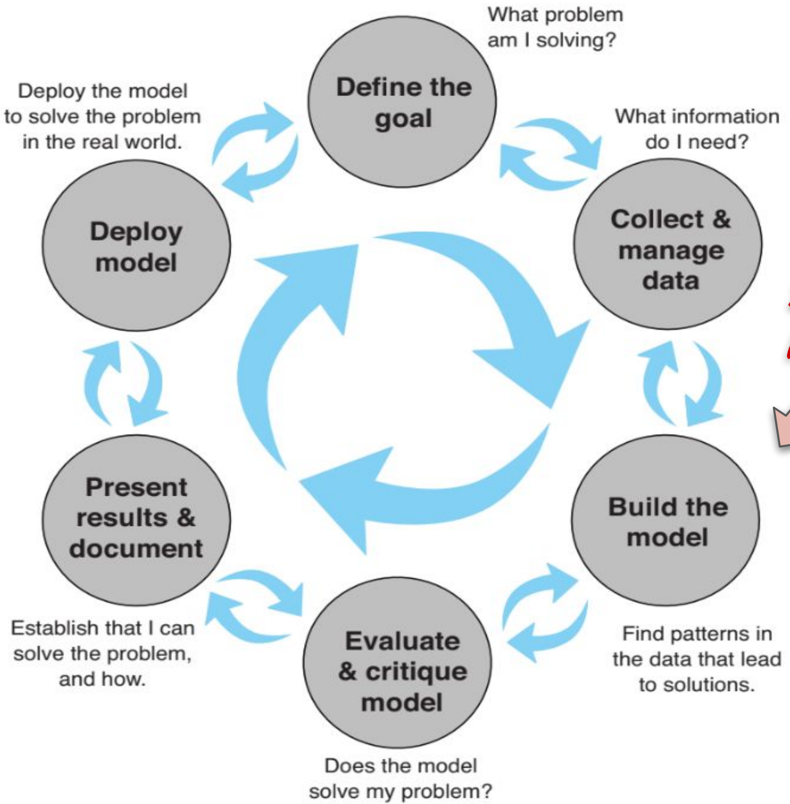
Recorrido de la materia (hasta ahora)



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
- ✓ Aprendizaje Supervisado
  - ❏ Clasificación: Árboles de decisión, KNN
  - ❏ Regresión: Regresión Lineal, KNN
  - ❏ Evaluación

Reducción de la dimensión

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

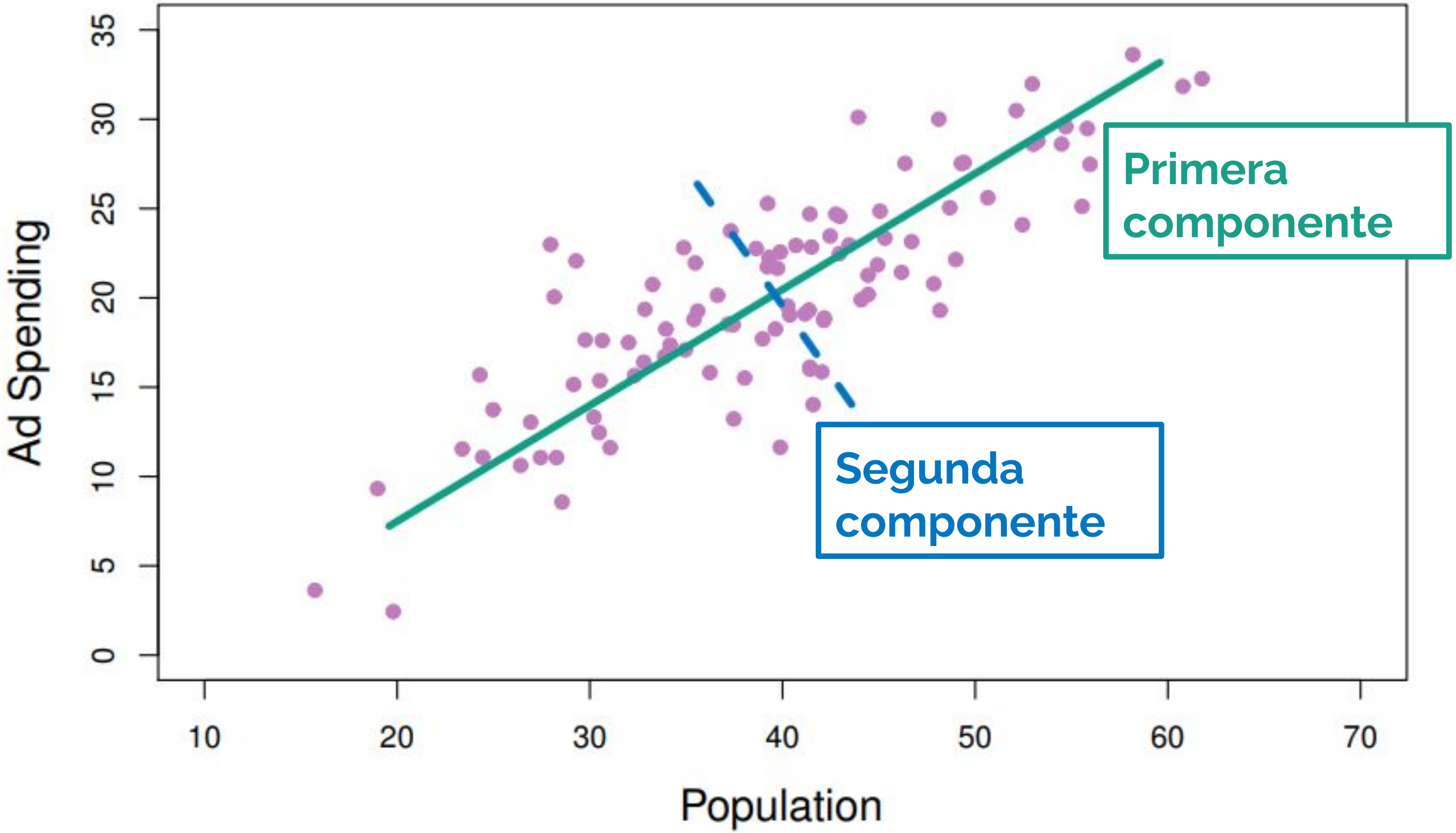
Técnicas (hay más)

- 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.

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
Ad Spending
Population
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.

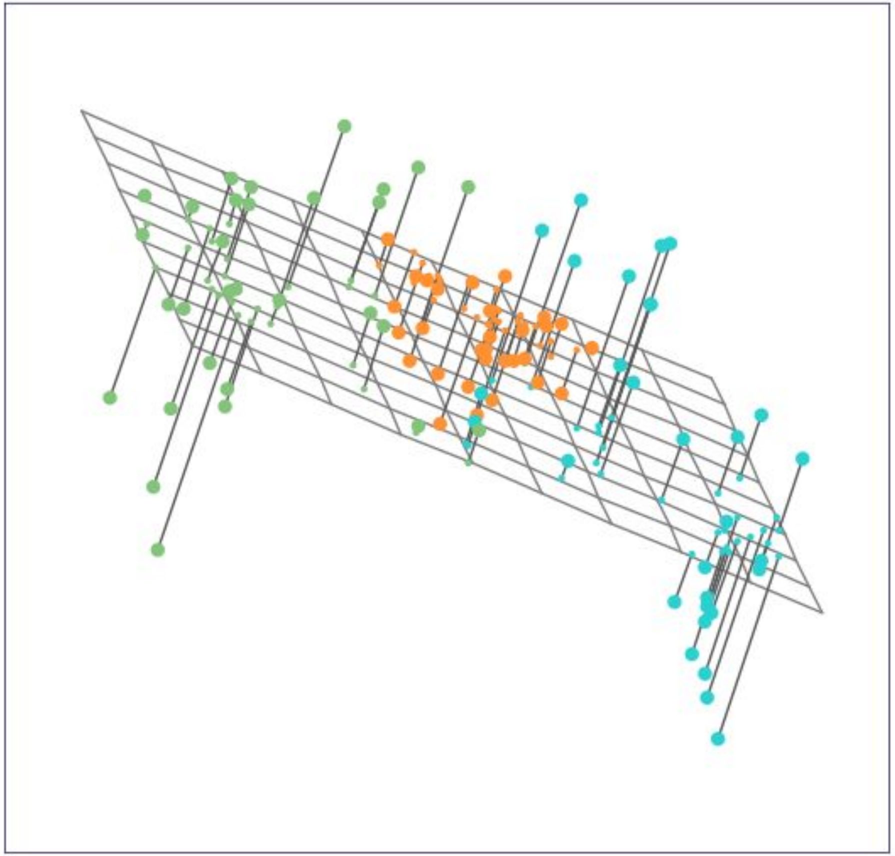
$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

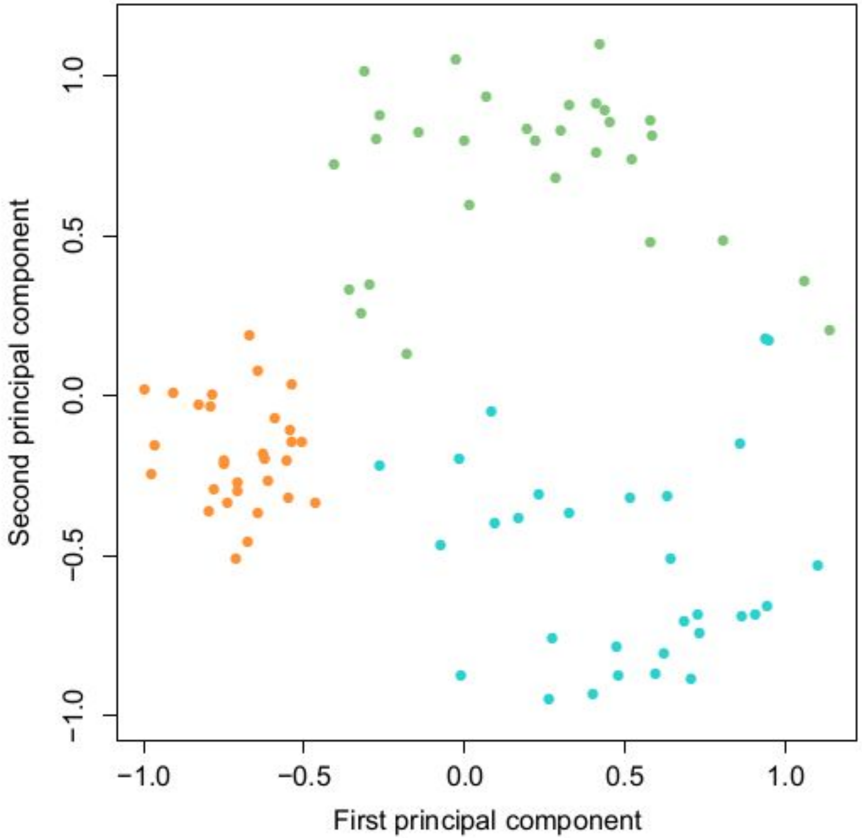
La segunda componente, **PC2**, es la dirección de **mayor varianza**, dentro de las direcciones **ortogonales** a PC1.

Así, hasta la p-ésima componente (eran p-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).





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

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:

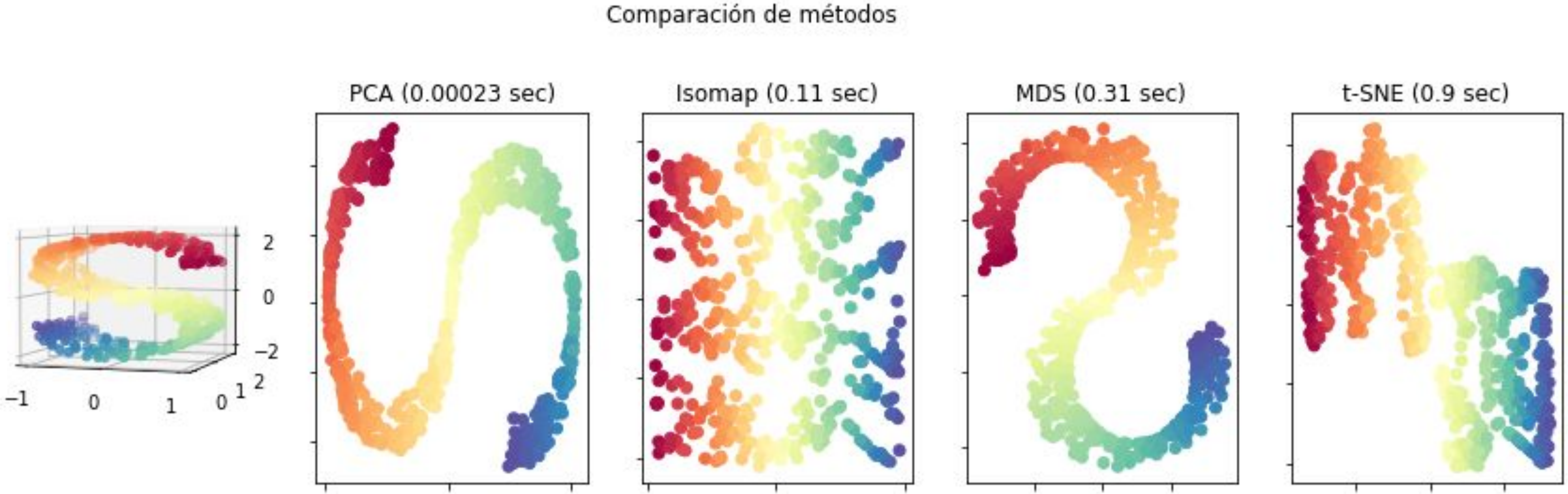
$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

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

Aprendizaje Automático

- Aprendizaje supervisado
  Hay una variable de interés a explicar o predecir, y datos etiquetados para un entrenamiento.
  - Problemas de regresión
    - Modelos lineales, knn
  - Problemas de clasificación
    - knn, árboles de decisión

- Aprendizaje no supervisado
  No se cuenta con datos para entrenamiento ni con una variable de interés en particular.
  - Reducción de la dimensión
    - PCA
  - Clustering
    - K-medias, DBSCAN, jerárquico