Economía y ciencia de los datos

Introducción a Machine Learning : Modelos No Supervisados

Carlos Alvarado & Pablo González

¿Qué es el aprendizaje no supervisado?

Entendemos como aprendizaje no supervisado al uso de algoritmos de machine learning para analizar y agrupar datos sin la presencia de una variable dependiente (no hay “y”)

Al no requerir de una variable objetivo o dependiente para realizar su labor, decimos que el aprendizaje es no supervisado, ya que estos logran descubrir patrones en lo datos sin la necesidad de intervención humana.

Tipos de algoritmos (Algunos ejemplos)

Clustering

- K-Means
- DBSCAN
- Espectral

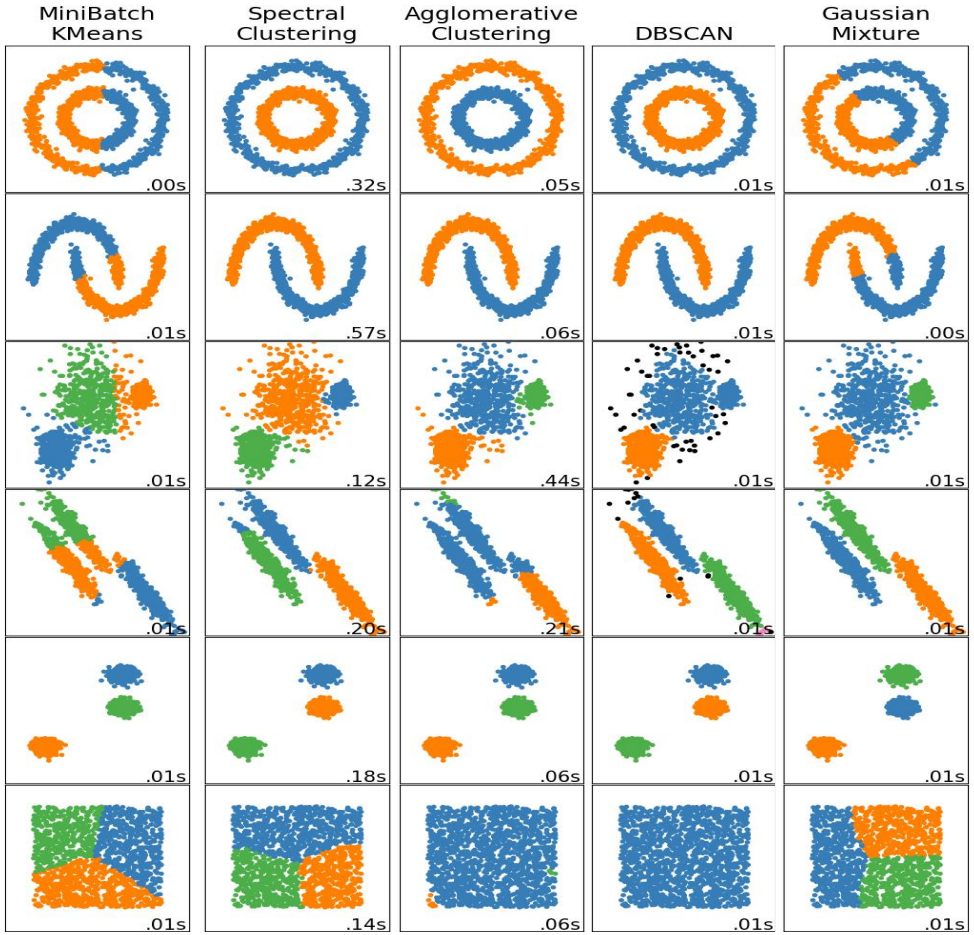
Reducción de dimensionalidad

- PCA
- ISOMAP

¿En qué sentido se podráin parecer los primeros?

Clustering

- Consiste en la tarea de agrupar un conjunto de puntos/casos a través de un determinado criterio y alguna medida de distancia basada en el set de atributos (variables).
- Algunos ejemplos serían los siguientes:



Fuente: [Scikitlearn]

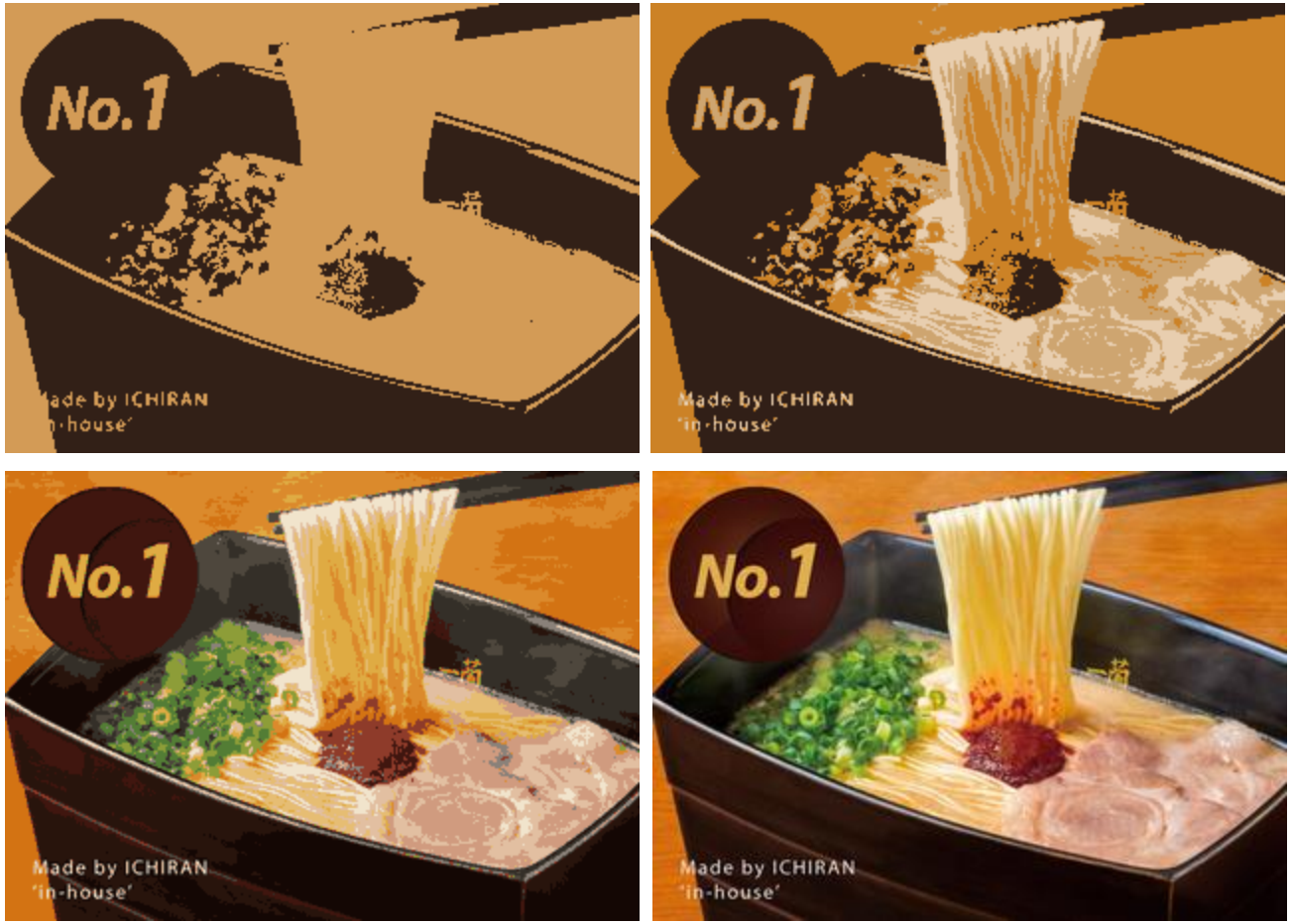
K-Means

Inputs: (1) $\mathrm{D} = \{d_1, d_2, \ldots, d_n\}$; (2) k (número de clusters)

Output: Conjunto de clusters.

Proceso:

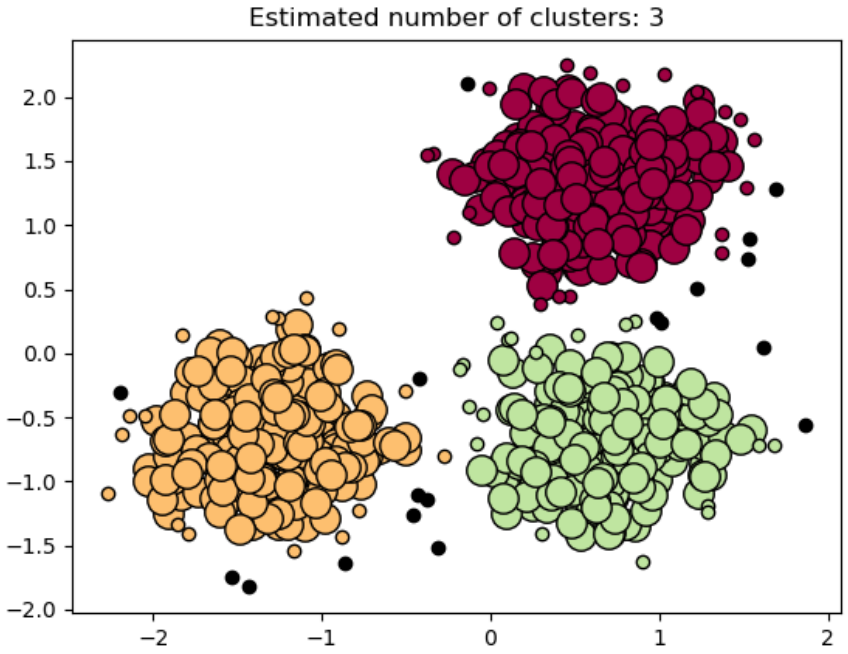
1. Seleccionar aleatoriamente k observaciones de D como centroides iniciales.
2. Repetir hasta convergencia:
    1. Asignar carda punto al cluster que se encuentra más cerca a un centroide
    2. Recomputar media para cada cluster (centroide)



K-Means con K = 2, 4 y 16 (+ original)

DBSCAN

- Este algoritmo se caracteriza por ser capaz de generar clusters (agrupar) sin la necesidad de que previamente se le indique el número de grupos/clusters.
- Es un algoritmo basado en densidad, basado en la idea de que un punto pertenece a un cluster si se encuentre cerca de múltiples otros puntos de éste.
- Frecuentemente es utilizado en casos con ruido o para detectar clusters con outliers



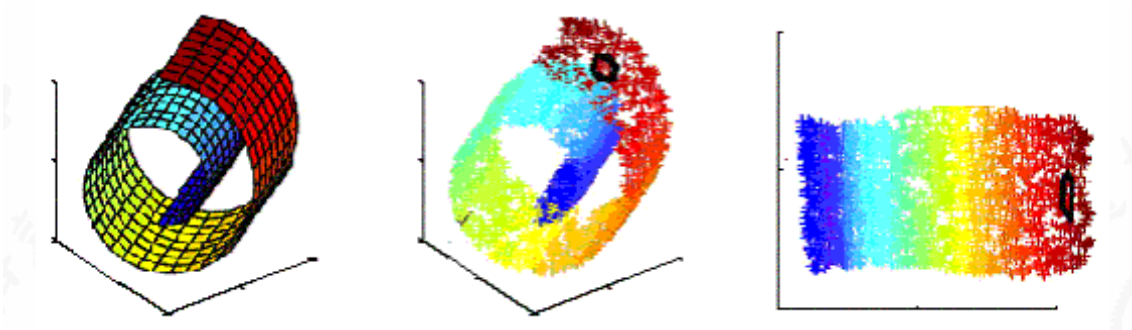
Ejemplo scikit-learn

Reducción de dimensionalidad

Generalmente, al momento de recolectar información, se suele optar por incluir la mayor cantidad de variables relacionadas posible. Ocasionalmente, algunas de estas variables están correlacionadas o, directamente, son transformaciones lineales de otra.

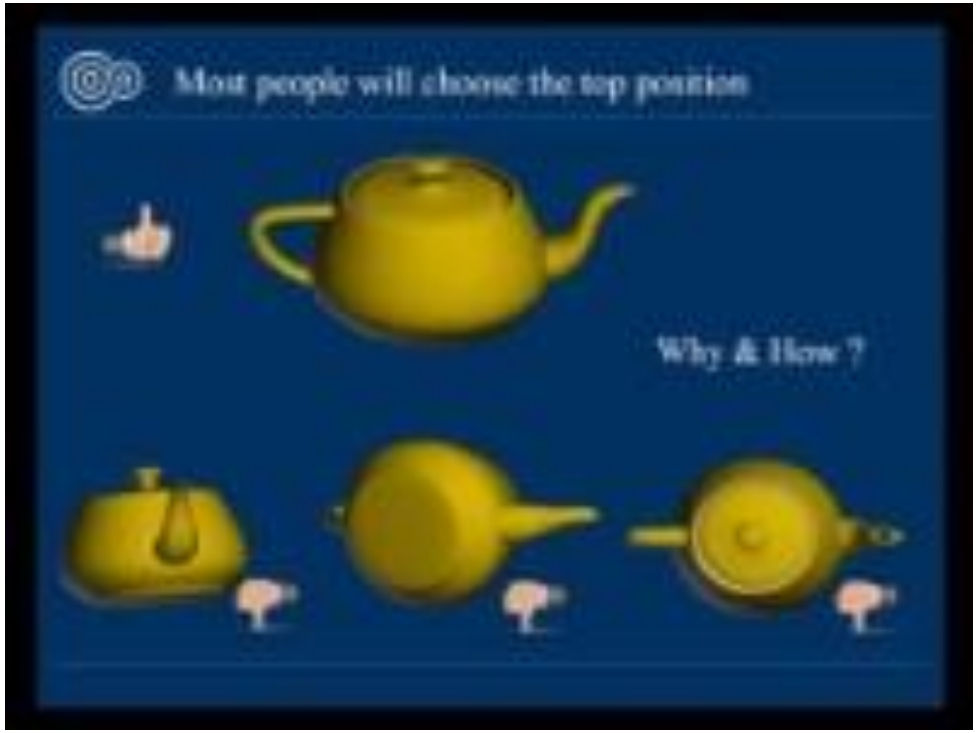
Lo anterior provoca diversos problemas, uno de los principales siendo la cantidad de valores no informativos presentes en los datos y el costo computacional que puede tener el uso de ciertos algoritmos sobre grandes bases de datos de alta dimensionalidad.

Los algoritmos de reducción de dimensionalidad reducen el número de variables, minimizando la ´pérdida de información. Así, nos permiten resolver o reducir este problema mediante tareas símiles a la construcción de índices sobre dichas variables.



PCA

- Consiste en transformar el conjunto original de variables en otro conjunto de nuevas variables no correlacionadas entre si.
- Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.
- Se buscan m<p variables que sean combinaciones lineales de las p originales, que no estén correlacionadas y que recojan la mayor parte de la información o variabilidad de los datos.

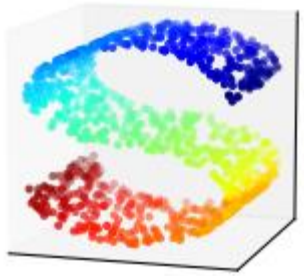


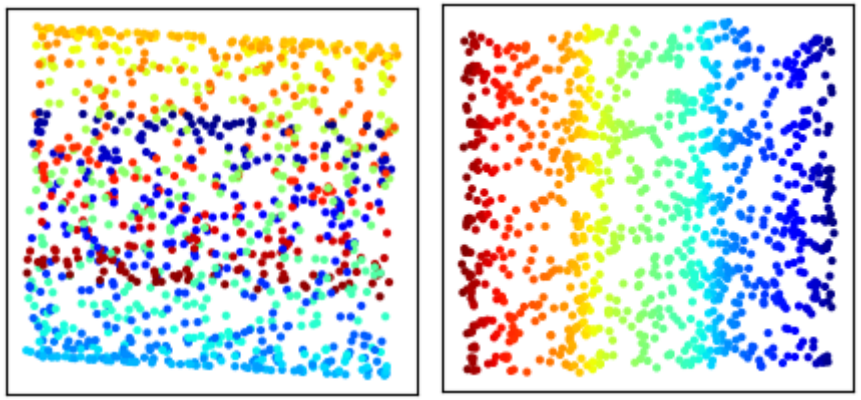
https://www.youtube.com/watch?v=BfTMmoDFXyE

ISOMAP

ISOMAP (Isometric Mapping.) corresponde a un método de reducción de dimensionalidad **no lineal.**

La meta consiste en mantener las distancias geodésicas (camino mas corto en la superficie) y utilizarlas para aproximar la geometría de la data antes de proyectarla a una menor dimensionalidad.





PCA - ISOMAP

Fuente de ejemplo: AstroML

Otros algoritmos interesantes

Mini Batch K-Means

Hierarchical Clustering

Spectral Clustering

Gaussian Mixture Models

BIRCH

Y más…

Sobre scikit-learn

scikit-learn

Machine Learning in Python

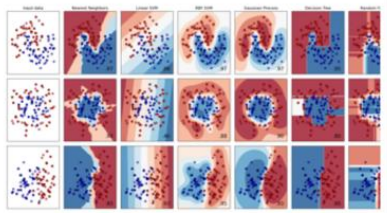
Getting Started | Release Highlights for 1.1 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

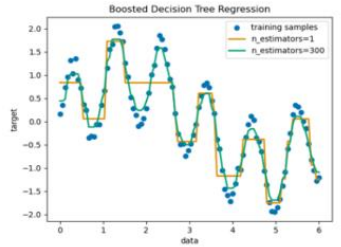


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

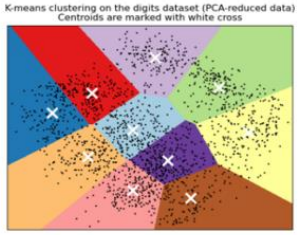


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

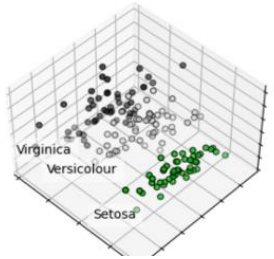


Examples

Dimensionality reduction

Reducing the number of random variables to consider.

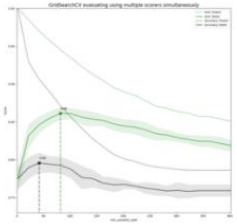
Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization, and more...



Model selection

Comparing, validating and choosing parameters and models.

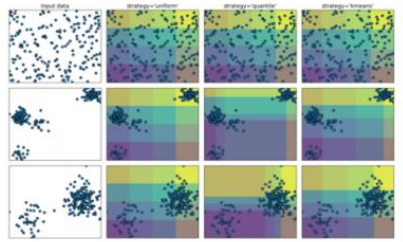
Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...



Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...



Sobre

Documentación

Casos de ejemplo:

- Clustering
  - K-Means
  - DbScan
- Reducción de dimensionalidad
  - PCA
  - Isomap