APRENDIZAJE AUTOMÁTICO CON PYTHON

CUARTA SESIÓN

TOULOUSE LAUTREC

TOULOUSE LAUTREC

EDUCACIÓN CONTINUA

ALEXANDER VALDEZ PORTOCARRERO

CONTENIDO

Introducción al Machine Learning

SESIÓN - 1

Introducción al Machine Learning

- Fundamentos de Machine Learning
- Tipos de Aprendizaje
- Librerías de Python para Machine Learning
- Algoritmos de Regresión
- Evaluación de Modelos de Regresión

Introducción al Machine Learning

SESIÓN - 2

Clasificación

- Algoritmos de Clasificación
- Balanceo de Datos
- Regresión Logística
- Evaluación de Modelos de Clasificación
- Árboles de Decisión

CONTENIDO

Introducción al Machine Learning

SESIÓN - 3

Clustering

- Algoritmos de Agrupación
- Clustering No Jerárquico: Kmeans, PAM, CLARA
- Clustering Jerárquico: AGNES, DIANA
- Clustering Basado en Densidad: DBSCAN

Introducción al Machine Learning

SESIÓN - 4

Reducción de Dimensionalidad

- Análisis de Componentes Principales
- Análisis Factorial
- Selección de Variables

Normas de clase online:

- Habrá un **break de 10 min** después de la parte teórica y antes de la parte práctica en Google Collaboratory.

- La evaluación se realizará **durante la sesión 4 como trabajo final** y obtendrás puntos adicionales en función de las tareas resueltas y enviadas al correo del profesor:

TAREAS PRESENTADAS	PUNTOS ADICIONALES
Ninguna	0
1	+2
2	+4
3	+6

Aprendizaje No Supervisado

1. Clustering K-Means (ejercicio paso a paso en español)
2. Clustering Jerárquico
3. DBSCAN

Reducción de Dimensionalidad

1. Principal Component Analysis
2. Análisis Factorial
3. Selección de Variables

Algoritmo no supervisado

K-Means es un algoritmo no supervisado de Clustering- NO JERARQUICO

El algoritmo de Clustering K-means es uno de los más usados para encontrar grupos ocultos, o sospechados en teoría sobre un conjunto de datos no etiquetado. Esto puede servir para confirmar -o descartar- alguna teoría que teníamos asumida de nuestros datos. Y también puede ayudarnos a descubrir relaciones asombrosas entre conjuntos de datos, que de manera manual, no hubiéramos reconocido. Una vez que el algoritmo ha ejecutado y obtenido las etiquetas, será fácil clasificar nuevos valores o muestras entre los grupos obtenidos.

$$\underset{c_i \in C}{\operatorname{argmin}}\ dist(c_i, x)^2$$

Paso de Asignación de datos

$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

Paso de Asignación de Centroid

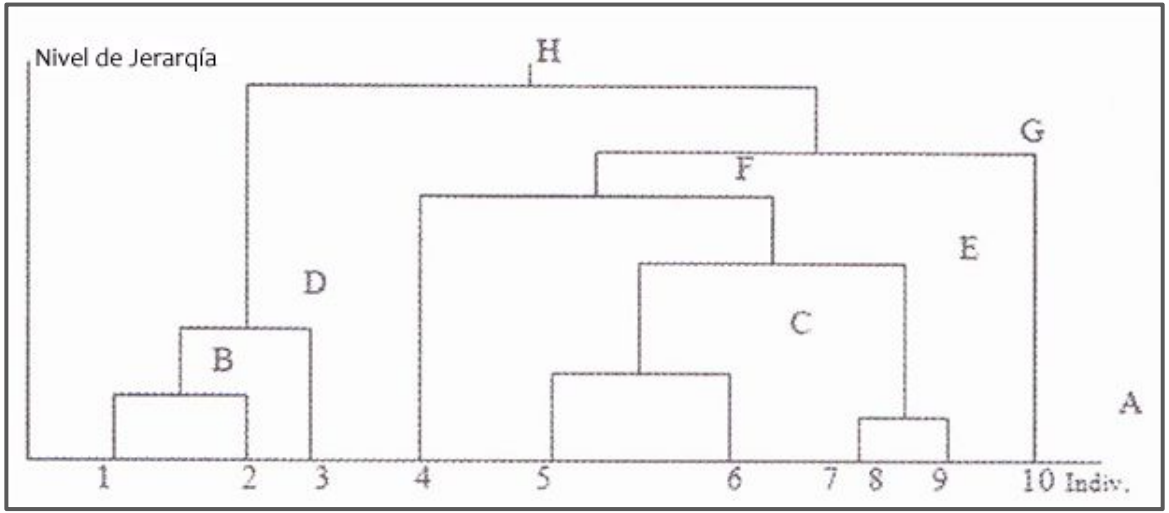
1. Identificar categorías dentro del dataset.
2. Visualización de datos
3. Definición de entrada
4. Obtención del valor de K
5. Ejecución del Kmeans
6. Clasificación de nuevas muestras.

Clustering Jerárquico

El clustering Jerárquico construye una jerarquía de clusters para realizar el análisis y existen dos categorías para este tipo de clustering:

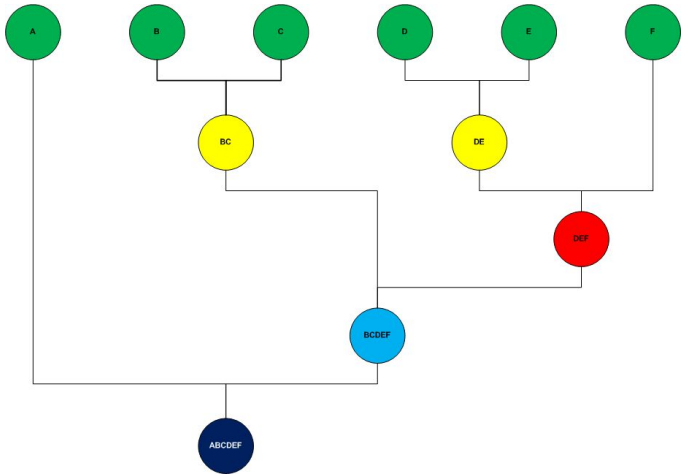
- Aglomerante
- Divisivo

Para representar los resultados de la jerarquía de grupos se usa el dendograma que muestra las jerarquías de acuerdo a las distancias que existen entre los elementos del conjunto de datos, las cuales se pueden representar en una matriz de distancias.



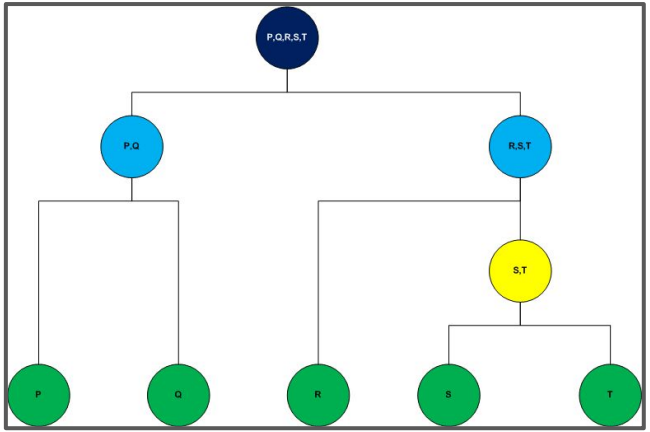
Clustering Jerárquico Aglomerante

Es una aproximación de abajo hacia arriba (bottom-up) donde se dividen los clusters en subclusters y así sucesivamente. Iniciando asignando cada muestra simple a un cluster y en cada iteración sucesiva va aglomerando (mezclando) el par de clusters más cercanos satisfaciendo algún criterio de similaridad, hasta que todos los elementos pertenecen a un solo cluster. Los clusters generados en los primeros pasos son anidados con los clusters generados en los siguientes pasos.



Clustering Jerárquico Divisivo

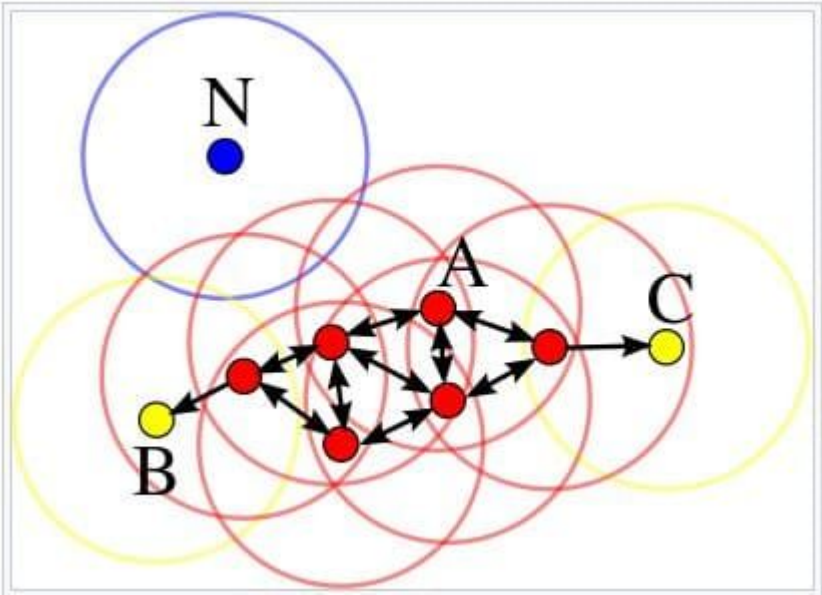
Este tipo de clustering se lleva a cabo con un enfoque de arriba hacia abajo (top-down), Se inicia con todos los elementos asignado a un solo cluster y sigue el algoritmo hasta que cada elemento es un cluster individual.
A diferencia del enfoque de abajo hacia arriba donde las decisiones para generar los clusters se basan en lo patrones locales sin tomar en cuenta la distribución global, el enfoque de arriba hacia abajo se beneficia de la información completa sobre la distribución global al ir haciendo las particiones.
El siguiente diagrama muestra el proceso divisivo



DBSCAN

Algoritmo de aprendizaje no supervisado basado en densidad. Calcula gráficos de vecinos más cercanos para encontrar valores atípicos y grupos de formas arbitrarias. Mientras que la agrupación de K-medias genera agrupaciones de forma esférica.

DBSCAN no requiere grupos K inicialmente. En cambio, requiere dos parámetros: eps y minPts.
* **eps :** es el radio de barrios específicos. Si la distancia entre dos puntos es menor o igual a esp, se considerarán sus vecinos.
* **minPts :** número mínimo de puntos de datos en un vecindario determinado para formar los grupos.
DBSCAN utiliza estos dos parámetros para definir un punto central, un punto fronterizo o un valor atípico.



DBSCAN

¿Cómo funciona el algoritmo de agrupamiento DBSCAN?

1. Seleccionando aleatoriamente cualquier punto **p** . También se le llama **punto central** si hay más puntos de datos que **minPts** en un vecindario.
2. Utilizará **eps** y **minPts** para identificar todos los puntos de densidad alcanzables.
3. Creará un clúster usando **eps** y **minPts** si **p** es un punto central.
4. Pasará al siguiente punto de datos si **p** es un **punto fronterizo** . Un punto de datos se denomina punto fronterizo si tiene menos puntos que **minPts** en la vecindad.
5. El algoritmo continuará hasta que se visiten todos los puntos.

PCA

¿No sería mejor tomar menos variables, pero más valiosas?

Técnica de Extracción de Características donde combinamos las entradas de una manera específica y podemos eliminar algunas de las variables “menos importantes” manteniendo la parte más importante todas las variables. Como valor añadido, luego de aplicar PCA conseguiremos que todas las nuevas variables sean independientes una de otra.

PCA

¿No sería mejor tomar menos variables, pero más valiosas?

Técnica de Extracción de Características donde combinamos las entradas de una manera específica y podemos eliminar algunas de las variables “menos importantes” manteniendo la parte más importante todas las variables. Como valor añadido, luego de aplicar PCA conseguiremos que todas las nuevas variables sean independientes una de otra.

Reducción de Dimensiones

Análisis Factorial

La reducción de dimensionalidad es una técnica crucial en el análisis de datos para simplificar conjuntos complejos de datos y encontrar patrones ocultos. Uno de los métodos más efectivos para reducir la dimensionalidad de datos mixtos es el Análisis Factorial de Datos Mixtos (FAMD). En este artículo, explicaremos qué es el FAMD y cómo se puede utilizar para reducir la dimensionalidad de conjuntos de datos complejos.

¿Qué es FAMD?

El FAMD es un método de análisis factorial que se utiliza para conjuntos de datos con múltiples tipos de variables, incluyendo variables cuantitativas y cualitativas. El FAMD utiliza una matriz de covarianza mixta para estimar las varianzas y covarianzas de los datos, y luego aplica un análisis de valores propios para reducir la dimensionalidad del conjunto de datos. El resultado es un conjunto de factores que explican la mayor parte de la variabilidad en los datos originales.

Selección de Variables

- Selección de Variables atendiendo a la multicolinealidad
- **Entonces, ¿qué ocurre si las variables «independientes» no son del todo independientes?**

  En ocasiones puede ocurrir que, en efecto, existe una dependencia entre las propias variables predictoras. A este hecho se le conoce como multicolinealidad, y su presencia no es que se diga muy satisfactoria cuando se trata de resolver problemas de predicción de variables.

- Una solución a este problema es utilizar el **Factor de Inflación de la Varianza (VIF, de sus siglas en inglés, Variance Inflation Factor)**, que permite cuantificar la intensidad de la multicolinealidad.

Valor de VIF	Grado de Muticolinealidad
Hasta 5	Débil/Moderado
De 5 a 10	Elevado
Mayor a 10	Muy elevado

APRENDIZAJE AUTOMÁTICO CON PYTHON

Gracias por su atencion.