KIBERNUM
IT Academy

Módulo 6 – Aprendizaje de Máquina No Supervisado

Introducción al Aprendizaje No Supervisado

Especialización en Ciencia de Datos

Objetivos

- Utilizar los conceptos básicos de aprendizaje de máquinas no supervisado.
- Conocer los distintos tipos de algoritmos.
- Diferenciar entre supervisado y no supervisado.
- Casos de uso.

KIBERNUM

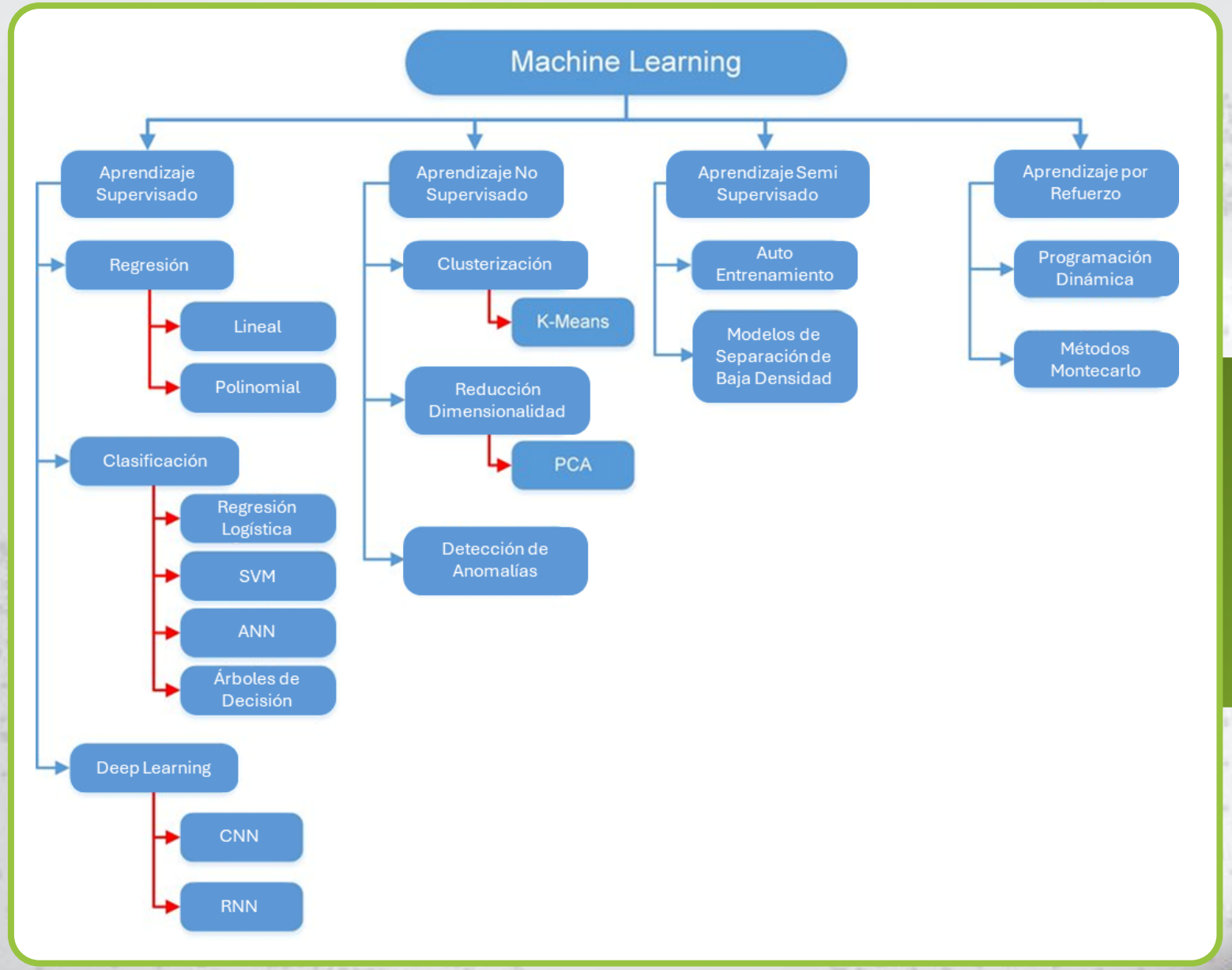
Contenido

- Aprendizaje de máquina no supervisado.
- Análisis de Clustering.
- Medidas de proximidad y similitud.
- Métodos jerárquicos.

KIBERNUM

Aprendizaje de Máquina No Supervisado

KIBERNUM



ML Aprendizaje No Supervisado

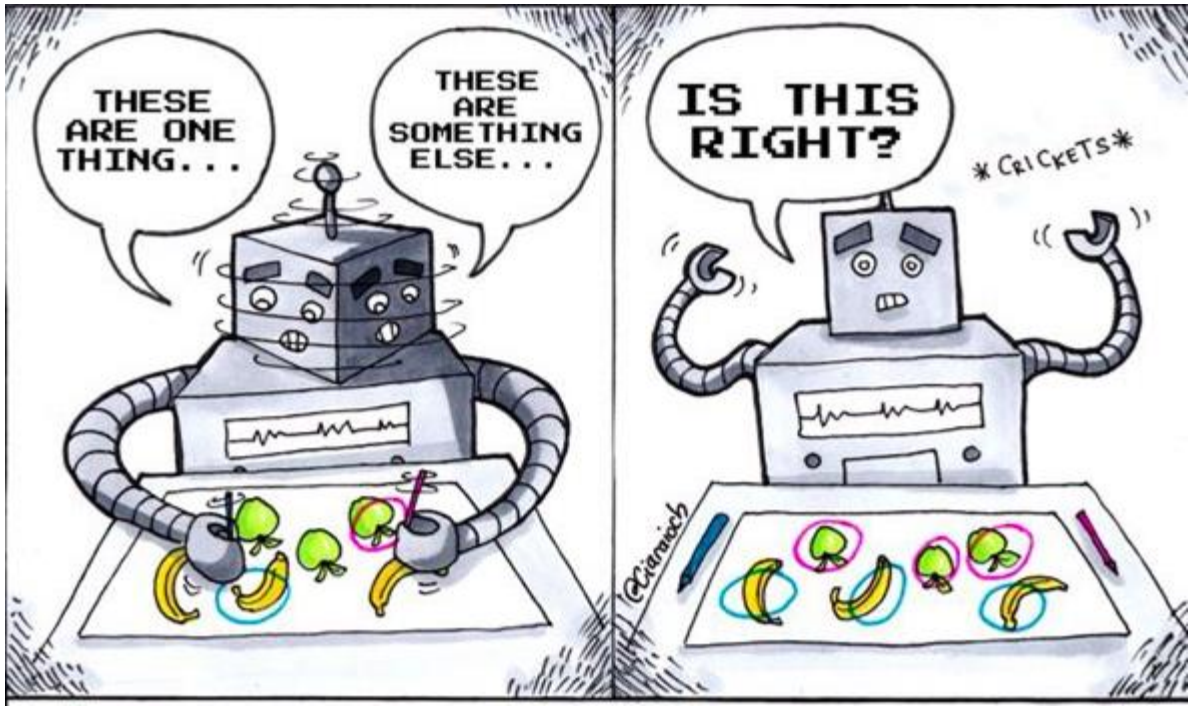
“Puedo aprender por mí mismo”

El modelo identifica relaciones y patrones automáticamente en el conjunto de datos sin etiquetar.

KIBERNUM

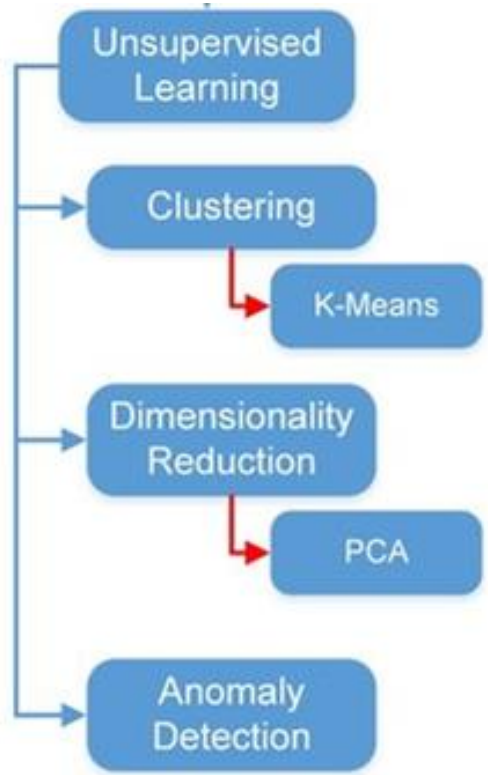
¿Qué es Aprendizaje No Supervisado?

El aprendizaje de máquina no supervisado es una técnica de aprendizaje automático en la que un algoritmo se utiliza para **descubrir patrones ocultos y relaciones en los datos sin la necesidad de etiquetas o categorías previas**. En otras palabras, no se le dice al algoritmo qué buscar o cómo clasificar los datos, sino que se le permite aprender por sí solo a partir de la estructura subyacente de los datos.



KIBERNUM

Tareas de Aprendizaje No Supervisado



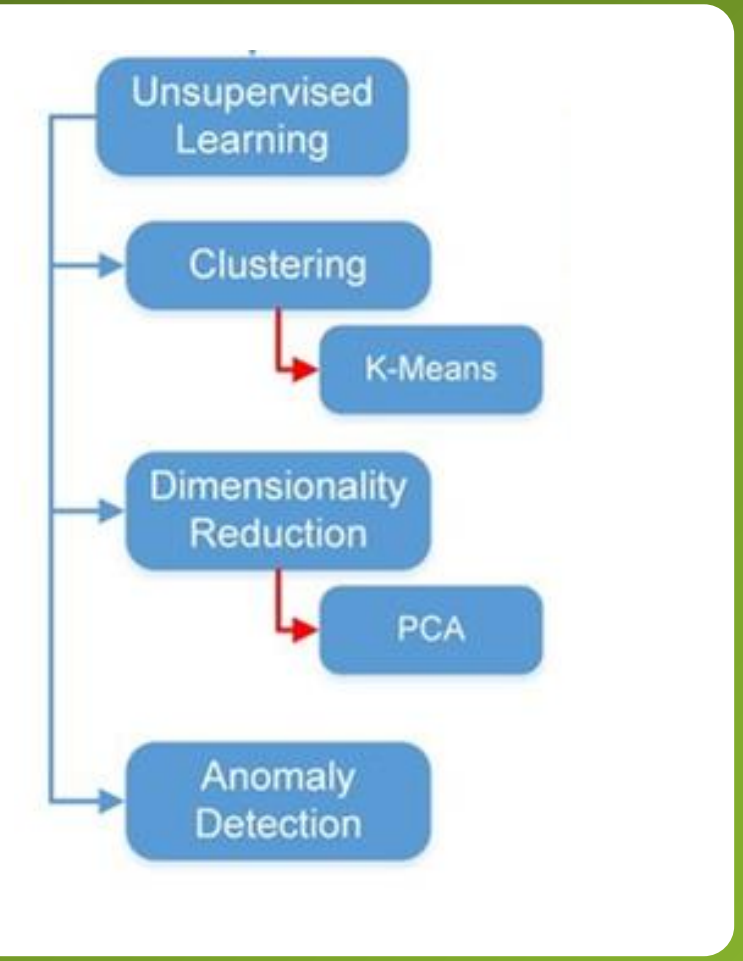
Las tareas más comunes en el aprendizaje no supervisado son:

1. **Clustering o agrupamiento:** identificar grupos o clústeres de objetos o datos similares.

1. **Reducción de la dimensionalidad:** reducir la complejidad de los datos al transformarlos en un espacio de menor dimensión.

1. **Detección de anomalías:** identificar objetos o datos que difieren significativamente de la mayoría del conjunto de datos.

Tareas de Aprendizaje No Supervisado



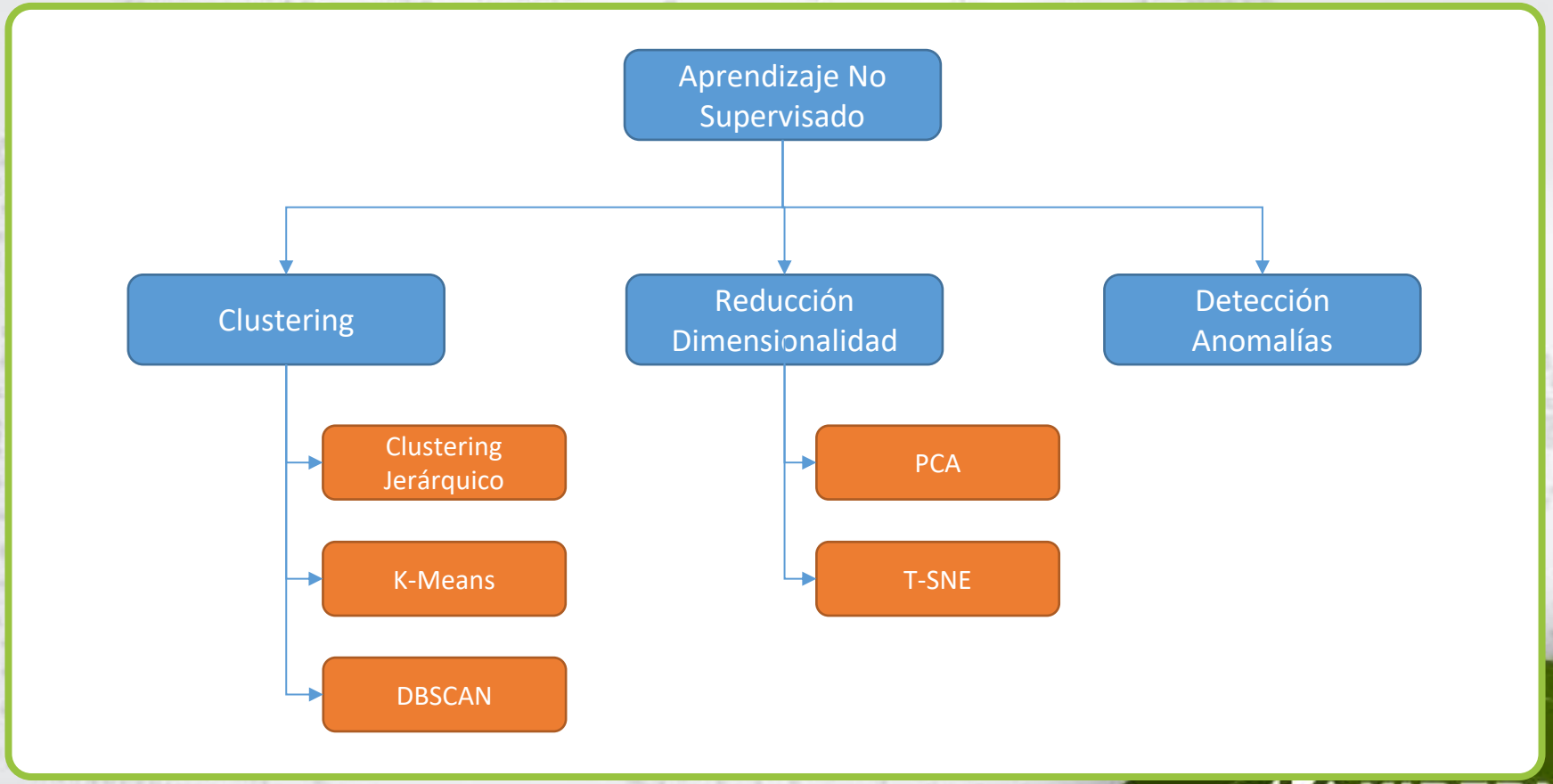
También, se pueden considerar como parte de las tareas de aprendizaje de máquina no supervisado a las siguientes:

4. **Asociación**: Esta tarea busca identificar patrones de asociación entre las diferentes variables en un conjunto de datos. Por ejemplo, en un conjunto de datos de compras de un supermercado, se puede utilizar el aprendizaje de máquina no supervisado para identificar que los clientes que compran pan también suelen comprar mantequilla.

5. **Generación de datos**: En esta tarea, se busca generar nuevos datos que sean similares a los datos originales. Esto se logra mediante la identificación de patrones en los datos originales y la creación de nuevas instancias que sigan esos patrones.

Tareas de Aprendizaje No Supervisado

En este curso, nos enfocaremos en las siguientes tareas y algoritmos de aprendizaje no supervisado:



KIBERNUM

Medidas de Proximidad y Similitud

KIBERNUM

Medidas de Proximidad y Similitud

- Las medidas de **proximidad (proximity)** y **similitud (similarity)** son herramientas estadísticas y matemáticas que se utilizan para cuantificar la similitud entre dos objetos o conjuntos de datos. A menudo se utilizan en aplicaciones de aprendizaje automático y minería de datos para comparar y clasificar objetos en función de su similitud.

- La medida de similitud, por otro lado, se utiliza para cuantificar la similitud entre dos objetos o conjuntos de datos. A diferencia de la medida de proximidad, la medida de similitud no necesariamente mide la distancia entre dos objetos, sino que se enfoca en cuantificar la similitud entre ellos. Por ejemplo, en un conjunto de datos que representa la preferencia de los usuarios en películas, la medida de similitud puede ser la similitud del coseno, que mide el ángulo entre dos vectores que representan las preferencias de los usuarios.

- Ambas medidas se utilizan comúnmente en técnicas de aprendizaje automático, como el clustering y la clasificación, para agrupar objetos similares y clasificar nuevos objetos en función de su similitud con objetos previamente clasificados. Las medidas de proximidad y similitud son fundamentales para muchas aplicaciones en minería de datos, reconocimiento de patrones, procesamiento del lenguaje natural y otras áreas relacionadas con la inteligencia artificial.

Medidas de Proximidad y Similitud

Las **medidas de proximidad** se utilizan para cuantificar la distancia o cercanía entre dos objetos en un espacio n-dimensional. Por ejemplo, en un espacio bidimensional, la medida de proximidad se puede calcular como la distancia euclidiana entre dos puntos. En un espacio n-dimensional, la medida de proximidad puede calcularse utilizando diversas técnicas, como la distancia de Manhattan, la distancia de Chebyshev, la distancia de Mahalanobis, etc.

Sean los vectores X e Y:

$$X = (X_1, X_2, \dots, X_p)$$
$$Y = (Y_1, Y_2, \dots, Y_p)$$

¿Cómo medir la cercanía entre X e Y?

KIBERNUM

Medidas de Proximidad y Similitud

Disimilitud

Una relación $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ es una medida de disimilaridad, si satisface las siguientes condiciones:

$$d(x, y) \geq 0$$

$$d(x, y) = 0 \Leftrightarrow x = y$$

$$d(x, y) = d(y, x)$$

Medidas de Proximidad y Similitud

Existen varias medidas de proximidad o distancia que se utilizan en estadística, minería de datos y aprendizaje automático. Algunas de las medidas de proximidad más comunes son:

Distancia Euclidiana: es la distancia lineal entre dos puntos en un espacio n-dimensional. Esta medida se calcula utilizando el teorema de Pitágoras y se puede utilizar en espacios con cualquier número de dimensiones.

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y})}$$

KIBERNUM

Medidas de Proximidad y Similitud

- **Distancia Euclidiana Generalizada**: es una extensión de la distancia euclidiana clásica que se utiliza para medir la distancia entre dos puntos en un espacio n-dimensional con diferentes ponderaciones en cada dimensión. En la distancia euclidiana clásica, todas las dimensiones se ponderan por igual, lo que significa que la distancia entre dos puntos se calcula como la raíz cuadrada de la suma de las diferencias al cuadrado entre cada dimensión.

- Sin embargo, en algunos casos, puede ser necesario asignar diferentes pesos o importancias a cada dimensión en función de la relevancia de cada variable en un problema específico. Por ejemplo, en un conjunto de datos de análisis de sentimientos, la importancia de algunas palabras en un texto puede ser mayor que la importancia de otras palabras en la determinación del sentimiento general del texto.

Medidas de Proximidad y Similitud

En estos casos, se puede utilizar la distancia euclidiana generalizada para medir la distancia entre dos puntos. La distancia euclidiana generalizada se calcula como la raíz cuadrada de la suma de las diferencias al cuadrado entre cada dimensión, pero cada diferencia se multiplica por un peso que refleja la importancia relativa de esa dimensión.

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} (x_i - y_i)(x_j - y_j)} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})}$$

KIBERNUM

Medidas de Proximidad y Similitud

- **Distancia Mahalanobis :** es una medida de distancia utilizada en estadística multivariada para medir la distancia entre dos puntos en un espacio n-dimensional, teniendo en cuenta la correlación entre las diferentes variables. Es una medida de distancia que se utiliza comúnmente en el análisis de datos y el aprendizaje automático para evaluar la similitud entre dos conjuntos de datos.

- La distancia de Mahalanobis se calcula como la distancia euclidiana entre el punto de interés y la media de los datos, dividida por la matriz de covarianza de los datos. La matriz de covarianza tiene en cuenta la variabilidad y la correlación entre las diferentes variables. Esta distancia es útil en casos donde la varianza de los datos es diferente en diferentes direcciones o cuando las diferentes variables están correlacionadas entre sí.

Medidas de Proximidad y Similitud

La distancia de Mahalanobis se utiliza en una variedad de aplicaciones en estadística y aprendizaje automático, incluyendo clasificación, agrupamiento y detección de anomalías. Es especialmente útil en problemas en los que los datos tienen diferentes escalas o varianzas y cuando las variables están correlacionadas entre sí.

$$d\,(\text{Mahalanobis}) = [(x_B - x_A)^T * \boldsymbol{C}^{-1} * (x_B - x_A)]^{0.5}$$

En donde C es la matriz de covarianza de la muestra.

KIBERNUM

Medidas de Proximidad y Similitud

Distancia Manhattan: también conocida como distancia de ciudad, del taxista o distancia L1, es una medida de distancia utilizada en matemáticas y ciencias de la computación para medir la distancia entre dos puntos en un espacio n-dimensional.

A diferencia de la distancia euclidiana, la distancia Manhattan mide la distancia "en línea recta" entre dos puntos a lo largo de los ejes de las coordenadas, en lugar de medir la distancia en línea recta entre los dos puntos. En otras palabras, la distancia Manhattan se calcula como la suma de las diferencias absolutas entre las coordenadas de los dos puntos.

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} |x_i - y_i|$$

KIBERNUM

Medidas de Proximidad y Similitud

Distancia de Minkowsky: es una medida de distancia utilizada en matemáticas y ciencias de la computación para medir la distancia entre dos puntos en un espacio n-dimensional. Esta medida es una generalización de la distancia Manhattan y la distancia euclidiana, y se puede ajustar para variar el peso de cada dimensión en la distancia calculada.

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{1/m}$$

Nótese que si m=1, entonces corresponde a la distancia de Manhattan. Y, con m=2, corresponde a la distancia euclideana.

Medidas de Proximidad y Similitud

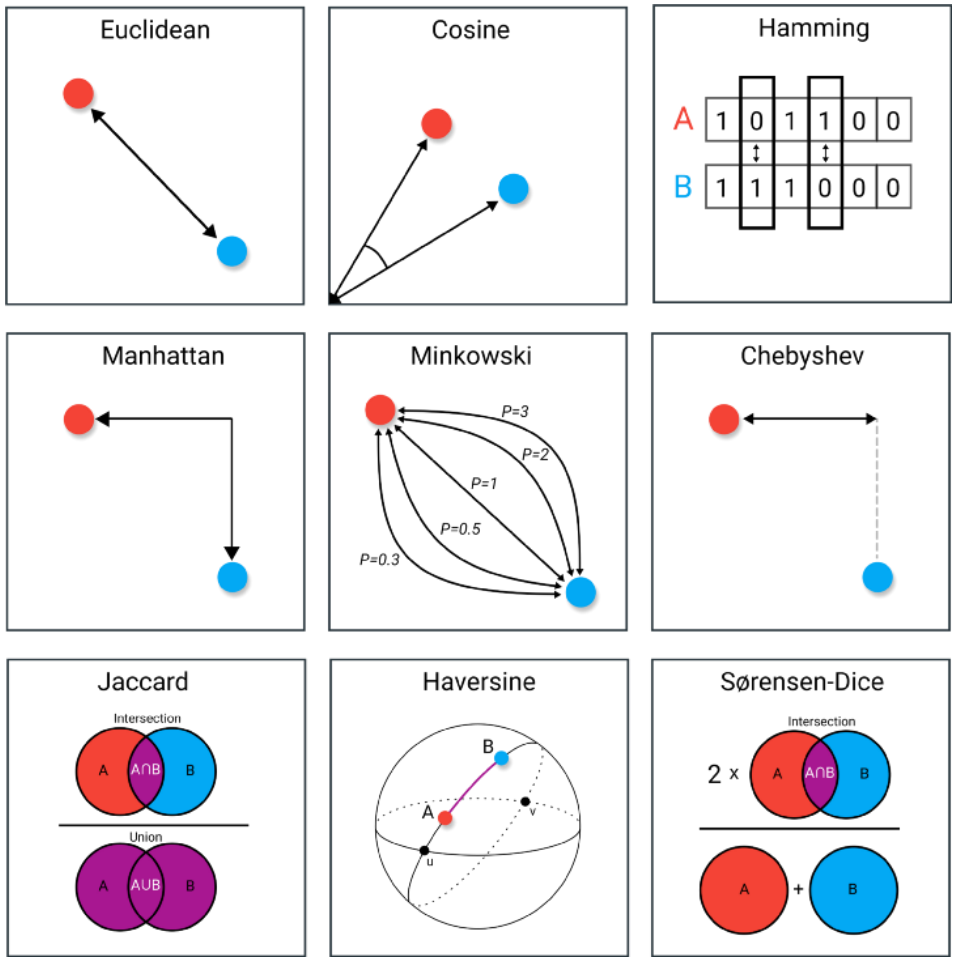
Existen otras medidas que sólo las mencionaremos.

Métrica de Canberra

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Coeficiente de Sorenzen-Dice

$$d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{2\sum_{i=1}^{p} \text{mín}\{|x_i|, |y_i|\}}{\sum_{i=1}^{p} |x_i| + |y_i|}$$



Medidas de Proximidad y Similitud

En el caso de variables categóricas ordinales, con M categorías, se puede realizar la siguiente transformación, para ser tratadas como variables continuas.

$$\frac{i-0{,}5}{M}, \quad i = 1, 2, \ldots, M.$$

En el caso de variables categóricas nominales, con M categorías, se suele asumir una disimilitud 0 ó 1, donde 1 se obtiene cuando coinciden en la categoría y 0 en otro caso.

Dudas y consultas

KIBERNUM

Fin Presentación

KIBERNUM