

MACHINE LEARNING

Msc Renzo Claure Aracena

1



Unsupervised learning

Reducción de dimensiones

Msc Renzo Claure Aracena

2

Introducción

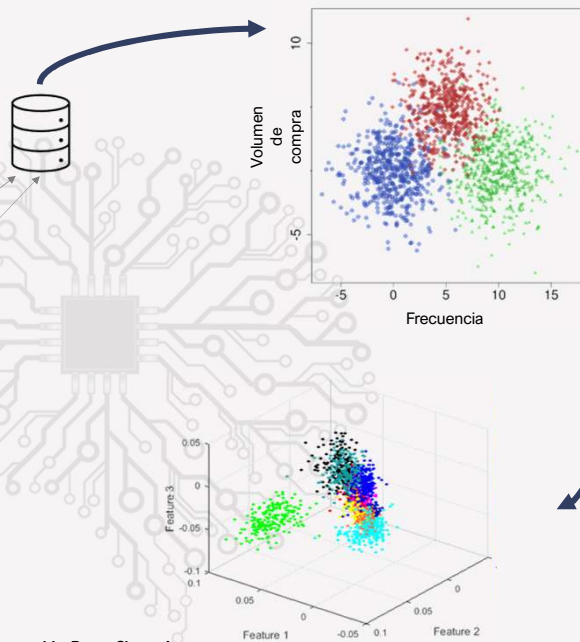
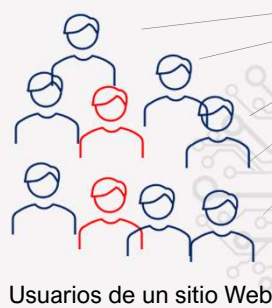
- El objetivo es identificar estructuras interesantes de información.
- No es supervisada por que no tiene una variable objetivo.
- Entre sus principales usos destacan:
 - Visualizar estructuras de datos complejos
 - Estimación de densidades para predecir probabilidades de eventos
 - Comprimir y resumir datos.
 - Extraer variables o características para el análisis supervisado
 - Encontrar outliers que podrían denotar comportamientos anómalos.



Msc Renzo Claire Aracena

3

Ejemplo

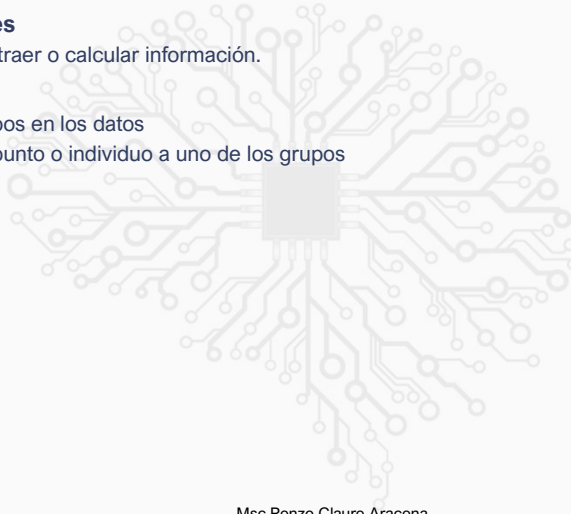


Msc Renzo Claire Aracena

4

Principales métodos del aprendizaje NS

- **Transformaciones**
 - Proceso de extraer o calcular información.
- **Clustering:**
 - Encontrar grupos en los datos
 - Asignar cada punto o individuo a uno de los grupos



Msc Renzo Claire Aracena

5



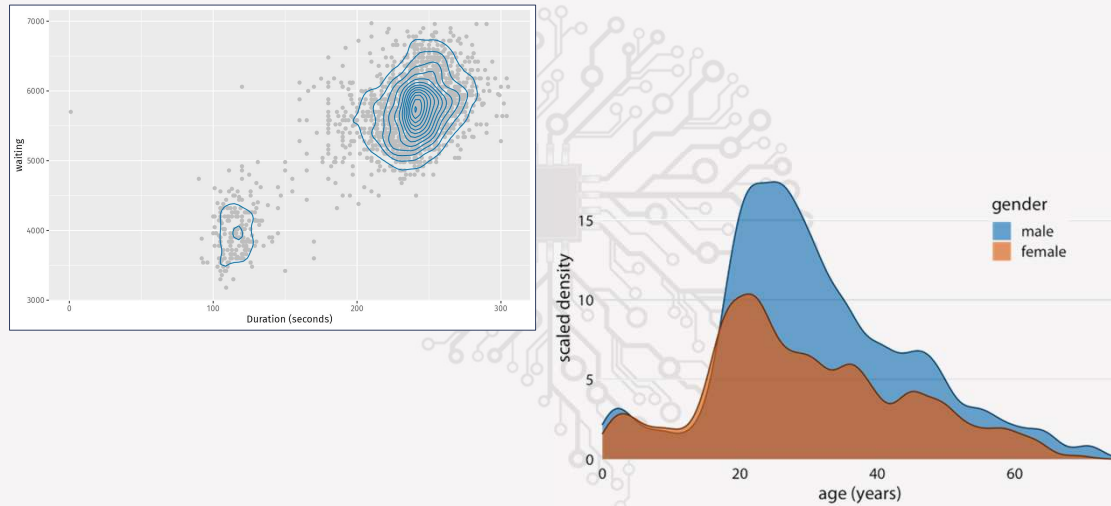
Transformaciones

Estimación de densidades y Reducción de dimensiones

Msc Renzo Claire Aracena

6

Transformación, Estimación de densidades

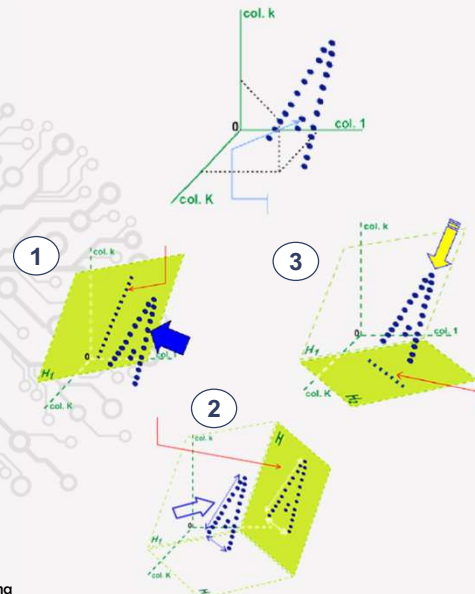


Msc Renzo Claire Aracena

7

Transformación: Dimensionality Reduction

- Busca la mejor representación de los datos usando menos dimensiones,
- Es utilizado especialmente para visualizar relaciones o agrupamientos en las variables de un set de datos.
- Utilizado también para comprimir o reducir la dimensionalidad en análisis supervisados.
- Es común utilizarlo para hacer visualizaciones en 2 dimensiones.

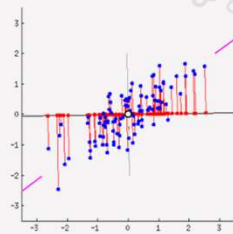


Msc Renzo Claire Aracena

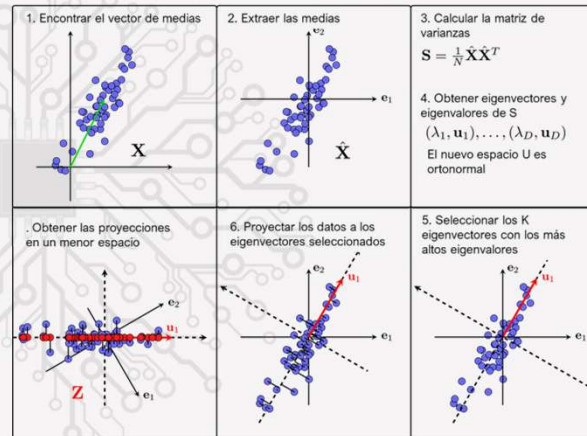
8

Análisis de componentes principales

- Basado en eigenvalores y eigenvectores.
- Busca minimizar la pérdida de información.
- Ordena de forma jerárquica los nuevos ejes, según su eigen valor que representa el nivel de aporte a la variabilidad total explicada.



Procedimiento PCA

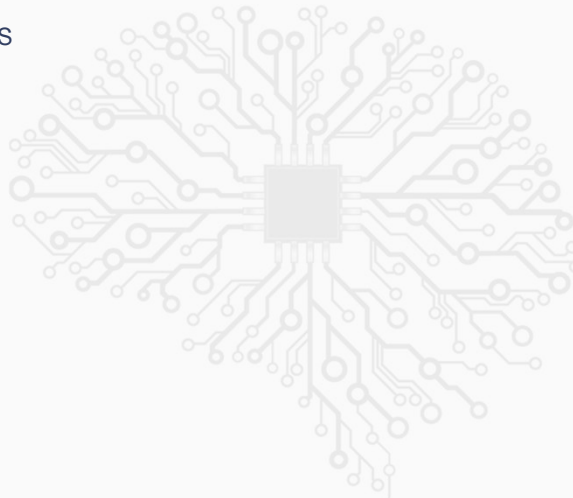


Msc Renzo Claire Aracena

9

Componentes principales PCA

- NB_16_PCA_MDS



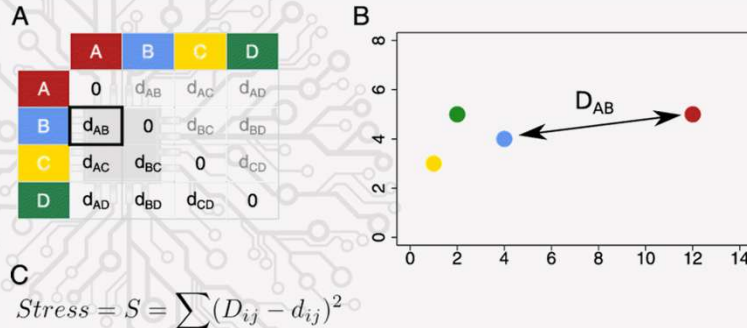
Msc Renzo Claire Aracena

10

Escalamiento multidimensional (MDS)

- Representación esquemática de la estrategia para el escalamiento multidimensional.

- a) Matriz positiva y simétrica de valores de distancia entre cuatro objetos.
- b) Representación de MDS que reduce la dimensión de las distancias en la matriz.
- c) Ecuación de estrés para calcular la diferencia general entre las distancias en el espacio de características (panel A, d_{ij}) y las distancias en el plano 2D (panel B, D_{ij}).



Msc Renzo Claire Aracena

11

Escalamiento multidimensional (MDS)

- Stress** mide la discrepancia entre las disimilitudes originales y las distancias en el espacio reducido.
- Fórmula del Stress**
- La fórmula más común para calcular el Stress es la siguiente:

- Interpretación del Stress**

- Valores bajos de Stress: Indican que las distancias en el espacio reducido se ajustan bien a las disimilitudes originales. Un Stress cercano a 0 significa un ajuste perfecto.
- Valores altos de Stress: Indican que hay una gran discrepancia entre las disimilitudes originales y las distancias en el espacio reducido, lo que sugiere un mal ajuste.

$$Stress = \sqrt{\frac{\sum_{i < j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i < j} d_{ij}^2}}$$

- Donde:**

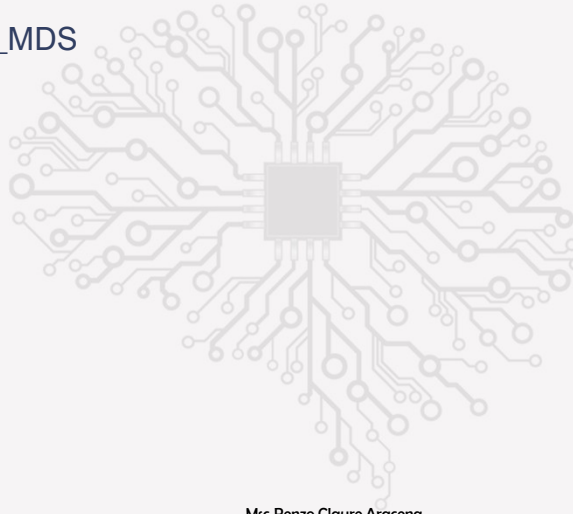
- d_{ij} : Es la disimilitud original entre los objetos i y j .
- \hat{d}_{ij} : Es la distancia entre los objetos i y j en el espacio de baja dimensión.
- La suma se realiza sobre todos los pares de objetos (i, j) .

Msc Renzo Claire Aracena

12

Escalamiento multidimensional (MDS)

- NB_16_PCA_MDS



Msc Renzo Claire Aracena

13

t-SNE t-Distributed Stochastic Neighbor Embedding

Geoffrey Hinton Facts



© Nobel Prize Outreach
Photo: Clement Morin

Geoffrey Hinton
Nobel Prize in Physics 2024

Born: 6 December 1947; London, United Kingdom

Affiliation at the time of the award: University of Toronto,
Toronto, Canada

Prize motivation: "for foundational discoveries and
inventions that enable machine learning with artificial
neural networks"

Prize share: 1/2



Laurens van der Maaten

Distinguished Research Scientist, Llama Team, Meta AI

Verified email at meta.com - Homepage

Artificial Intelligence Machine Learning Computer Vision

TITLE	CITED BY
Visualizing data using t-SNE L van der Maaten, G Hinton The Journal of Machine Learning Research 9 (2579-2605), 85	51929
Densely Connected Convolutional Networks G Huang, Z Liu, L van der Maaten, KQ Weinberger IEEE Conference on Computer Vision and Pattern Recognition	51900
Dimensionality reduction: A comparative review LJP Van der Maaten, EO Postma, HJ Van den Herik Technical Report TUG-TR-2009-006	4112
Accelerating t-SNE using Tree-Based Algorithms L Van Der Maaten The Journal of Machine Learning Research 15 (1), 3221-3245	3208
The Llama 3 Herd of Models A Dubey, A Jauhri, A Pandey, A Kadian, A Al-Dahle, A Letman, A Mathur, ... arXiv preprint arXiv:2407.21783	2962
CLaVE: A diagnostic dataset for compositional language and elementary visual reasoning J Johnson, B Harizan, L Van Der Maaten, L Fei-Fei, C Lawrence Zitnick, Proceedings of the IEEE conference on computer vision and pattern ...	2952
3d semantic segmentation with submanifold sparse convolutional networks B Graham, M Engelcke, L Van Der Maaten Proceedings of the IEEE conference on computer vision and pattern ...	1821

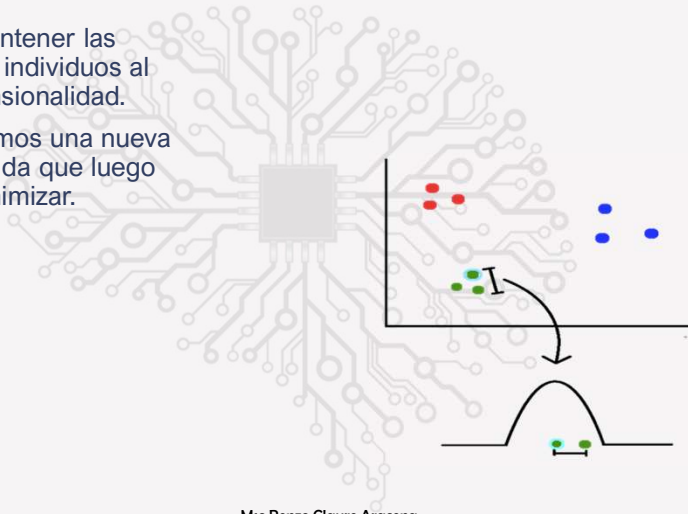
Msc Renzo Claire Aracena

14

t-SNE

t-Distributed Stochastic Neighbor Embedding

- Tratamos de mantener las distancias entre individuos al reducir la dimensionalidad.
- Para esto definimos una nueva función de pérdida que luego tratamos de minimizar.



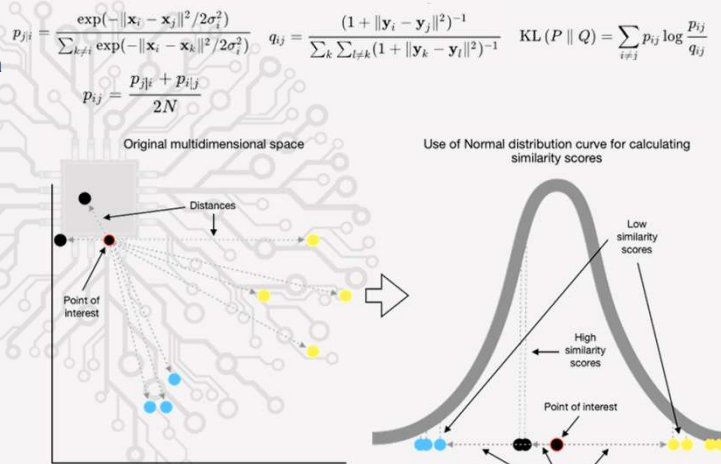
Msc Renzo Claire Aracena

15

t-SNE

t-Distributed Stochastic Neighbor Embedding

- Esta función de pérdida tiene la forma de una función de probabilidad, con una curva gaussiana que depende de la varianza.
- La varianza depende a su vez del principal parámetro: "perplexcity"
- Perplexity, controla la amplitud de las colas de la distribución. Se calcula a través de las entropías de las distribuciones. Representa el número de puntos que n punto tomará como vecinos.
- "t" de t-SNE, es una mejora, basada en la distribución t.
- Elevados valores de preplexicity sirven para modelos más complejos, pero demoran más, se deben probar varios valores, lo que lo hace más costoso.
- Más lento, pero mejor que PCA



Msc Renzo Claire Aracena

16

t-SNE t-Distributed Stochastic Neighbor Embedding

- Ventajas de t-SNE
 - Captura estructuras no lineales: Es muy efectivo para visualizar agrupamientos y patrones no lineales en los datos.
 - Enfoque en vecindades locales: Preserva mejor las distancias locales que las globales.
 - Visualización intuitiva: Los resultados suelen ser fáciles de interpretar visualmente.
- Limitaciones de t-SNE
 - No preserva distancias globales: Las distancias entre clusters en el espacio reducido no tienen significado directo.
 - Dependencia de hiperparámetros: La elección de la perplexity y la tasa de aprendizaje puede afectar significativamente los resultados.
 - No es determinista: Diferentes ejecuciones pueden dar resultados ligeramente diferentes.
 - No es adecuado para reducción de dimensionalidad general: t-SNE está diseñado principalmente para visualización, no para reducir dimensiones para otros algoritmos.

Msc Renzo Claire Aracena

17

t-SNE t-Distributed Stochastic Neighbor Embedding

- NB_17_tsne_umap

Msc Renzo Claire Aracena

18

UMAP

- Tiene el mismo principio que tSNE
- Sirve para espacios con mayor dimensión.
- El objetivo de UMAP es preservar tanto la estructura local como la global de los datos al mapearlos a un espacio de baja dimensión. Esto significa que no solo se enfoca en mantener las relaciones de vecindad cercana (como t-SNE), sino que también intenta preservar las relaciones entre clusters y la estructura general de los datos.
- En lugar de usar curvas gaussianas usa grafos
- Se construyen grafos, con los "k" vecinos próximos, estos grafos son ponderados con las distancias entre los vecinos.
- Distancia mínima entre puntos en el espacio reducido. Controla qué tan "compactos" son los grupos (0.1).
- Se repite este proceso para cada punto.
- De este modo se seleccionan grafos con los menores pesos.
- Se calcula una función de pérdida con la matriz de adyacencia.
- Es más rápido.

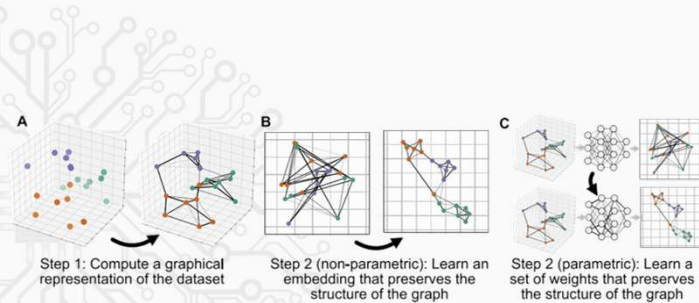


Figure 1: Overview of UMAP (A → B) and Parametric UMAP (A → C).

Msc Renzo Claire Aracena

19

UMAP

- Hiperparámetros:
 - `n_neighbors`: Controla el número de vecinos considerados para construir el grafo en el espacio de alta dimensión. Valores más altos preservan más la estructura global, mientras que valores más bajos se enfocan en la estructura local.
 - `min_dist`: Controla la distancia mínima entre puntos en el espacio de baja dimensión. Valores más bajos permiten que los puntos estén más juntos, mientras que valores más altos los separan.
 - `n_components`: Número de dimensiones en el espacio de baja dimensión (generalmente 2 o 3 para visualización).
 - `metric`: Métrica de distancia utilizada para calcular las distancias en el espacio de alta dimensión (por ejemplo, 'euclidean', 'cosine', 'manhattan').
- Ventajas:
 - Preservación de la estructura global: A diferencia de t-SNE, UMAP intenta preservar tanto la estructura local como la global.
 - Mayor velocidad.
 - Flexibilidad.
 - Resultados reproducibles.
- Desventajas:
 - Dependencia de hiperparámetros
 - Interpretación de distancias.
 - No es determinista en todos los casos.

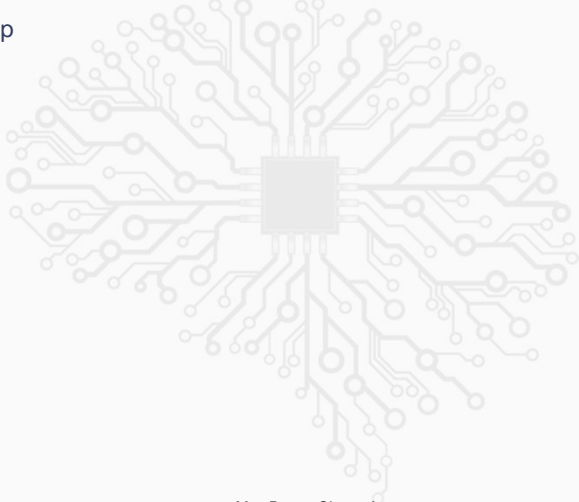
Msc Renzo Claire Aracena

20

MACHINE LEARNING

UMAP

- NB_17_tsne_umap



Msc Renzo Claire Aracena

21



Unsupervised learning

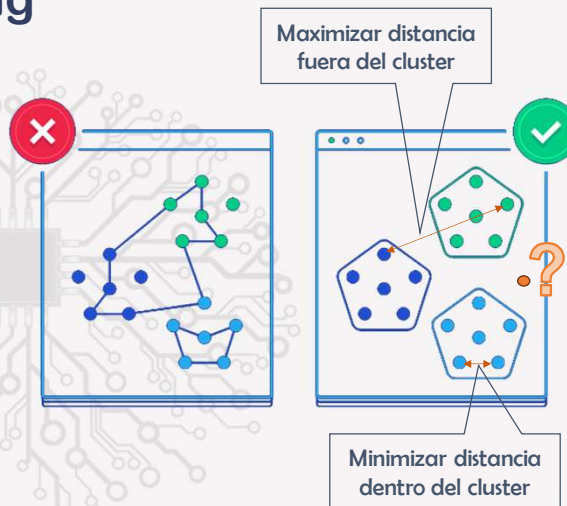
Clustering

Msc Renzo Claire Aracena

22

Métodos de Clustering

- Busca encontrar similitudes entre individuos, generalmente a través de su "distancia".
- Puntos dentro del mismo cluster son más semejantes y a su vez más distintos a puntos de otros clusters.
- El algoritmo devuelve los agrupamientos, es decir, la pertenencia a un grupo determinado.
 - Hard: todos los puntos son asignados.
 - Difuso: se asigna a los puntos una probabilidad de pertenencia a un grupo.

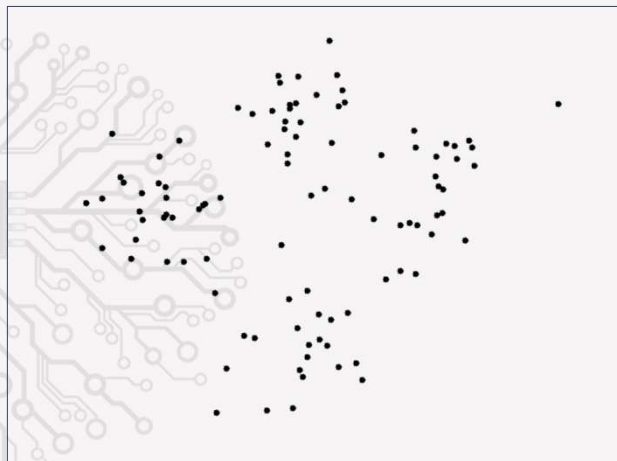


Msc Renzo Claire Aracena

23

K-mean clustering

- Método recursivo de agrupamiento
- Inicia con la definición de la cantidad "k" de clusters y la definición de los centroides aleatorios.
- Asignar cada punto al centroide más cercano.
- Recalcular los centroides con la media de la posición de los puntos asignados.
- Volver a medir distancias y reasignar los puntos a los centroides más cercanos hasta obtener una solución estable.
- Sus problemas se basan en:
 - La aleatoriedad de los centroides iniciales.
 - La dimensionalidad del espacio.



Msc Renzo Claire Aracena

24

K-means limitaciones

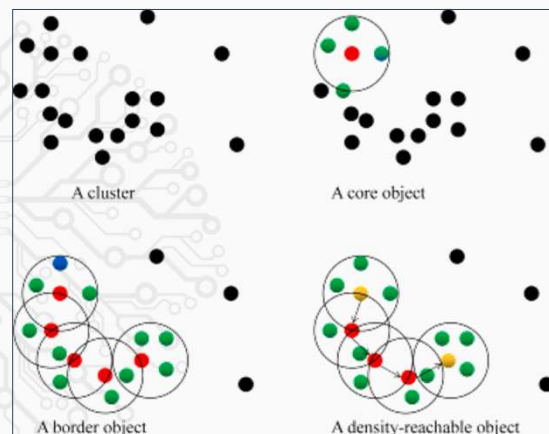
- Funciona bien y es fácil de comprender las agrupaciones con bases de datos pequeñas y pocas características
- No funciona con clusters complejos e irregulares
- Existen variantes para el uso de datos categóricos después de una adecuada adaptación.

Msc Renzo Claire Aracena

25

DBSCAN Clustering

- El funcionamiento del algoritmo DBSCAN se basa en clasificar las observaciones en tres tipos:
 - Puntos núcleo (Core points): son aquellos puntos que cumplen con las condiciones de densidad que hemos establecido.
 - Puntos alcanzables (Achievable points): son aquellos puntos que, aunque no cumplen con las condiciones de densidad, están cerca de otros puntos núcleo.
- Ruido (Noise): son los puntos que no cumplen con las condiciones de densidad y, además, en su radio no tienen otros puntos.
- Por lo tanto, DBSCAN se basa en lo siguiente:
 - Calcula la matriz de distancias entre los diferentes puntos. Generalmente se utiliza la distancia euclídea, aunque se pueden usar otras.
 - Teniendo en cuenta los parámetros del modelo, clasifica cada punto entre punto núcleo, punto de borde y ruido. En este sentido, pueden surgir diferentes puntos núcleo, ya que puede haber varias zonas de densidad. Cada uno de esos puntos núcleo pertenecerá a un cluster.
 - Asigna los puntos alcanzables de cada cluster al cluster correspondiente.

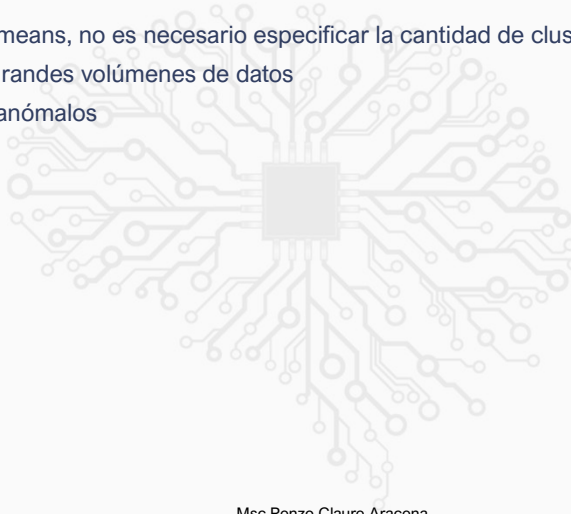


Msc Renzo Claire Aracena

26

DBSCAN Clustering

- A diferencia de Kmeans, no es necesario especificar la cantidad de clusters.
- Es eficiente con grandes volúmenes de datos
- Identifica puntos anómalos



Msc Renzo Claire Aracena

27

Métricas de evaluación de clusters

- El coeficiente de silueta es una métrica que mide la calidad de los clusters. Para cada punto, se calcula de la siguiente manera:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

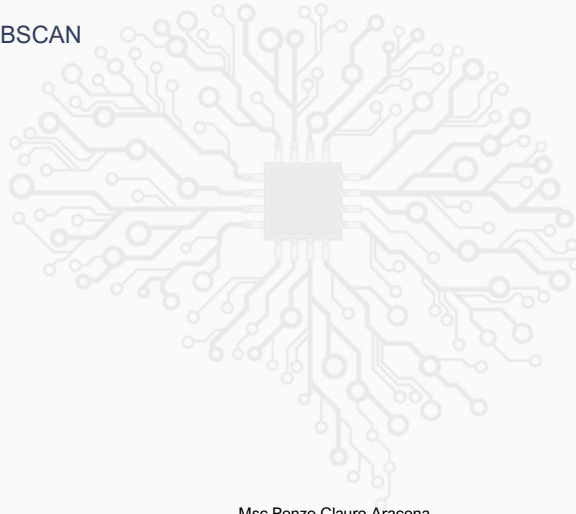
- Donde:
 - $a(i)$: Distancia promedio del punto i a todos los demás puntos en el mismo cluster (cohesión).
 - $b(i)$: Distancia promedio del punto i a todos los puntos en el cluster más cercano (separación).
- El coeficiente de silueta $s(i)$ varía entre **-1 y 1**:
 - **Cercano a 1**: El punto está bien asignado a su cluster (buena cohesión y separación).
 - **Cercano a 0**: El punto está cerca del límite entre dos clusters.
 - **Cercano a -1**: El punto está probablemente asignado al cluster incorrecto.
- El **coeficiente de silueta promedio** es la media de $s(i)$ para todos los puntos y se utiliza para evaluar la calidad general de los clusters.
- El coeficiente de silueta es una métrica útil, pero tiene limitaciones importantes. Para obtener una evaluación más completa de los clusters, es recomendable combinar el coeficiente de silueta con otras métricas (como el índice de Calinski-Harabasz o el índice de Davies-Bouldin) y técnicas visuales (como gráficos de dispersión o t-SNE/UMAP).
- Entropía, es un concepto fundamental en teoría de la información y estadística que mide la incertidumbre o el desorden en un sistema. En el contexto de clustering (agrupamiento), la entropía se utiliza para evaluar la pureza de los clusters, es decir, qué tan bien separadas están las clases dentro de cada cluster.

Msc Renzo Claire Aracena

28

Kmeans - DBSCAN

- NB_18_kmean_DBSCAN



Msc Renzo Claire Aracena