DataSphere

Let's generate value

Workshop 2019: K-means

September-2019

info@datasphere.tech

www.datasphere.tech

Agenda

- Definición de Machine Learning
- Clustering/K-Means
- Cuantos Clusters
- Criterios de Validación
- Recomendaciones Pre y Pos procesamiento

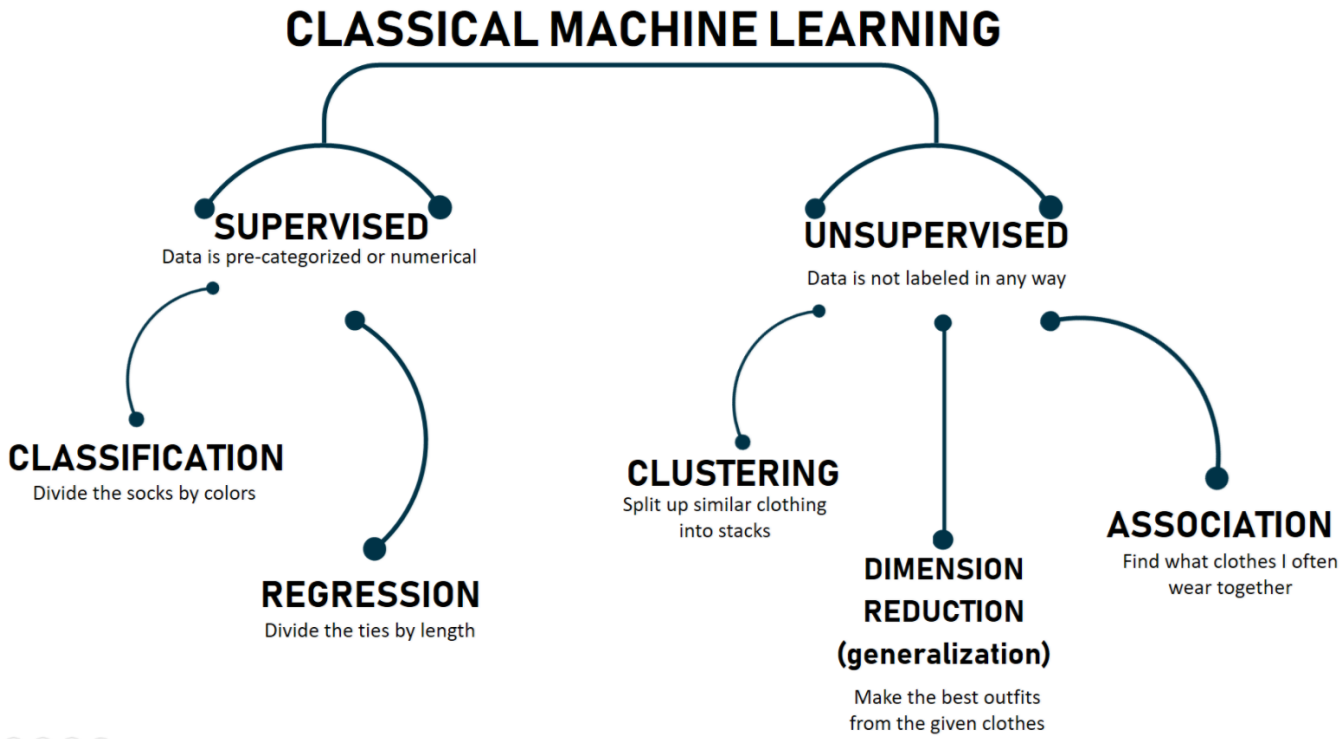
DataSphere
Let's generate value

Machine Learning

DataSphere
Let's generate value

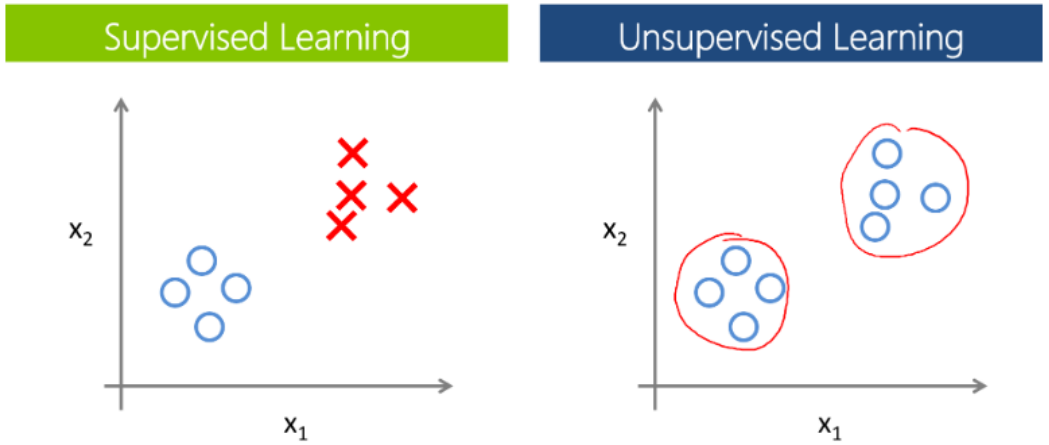
Machine Learning

- Machine Learning son un conjunto de métodos/algoritmos diseñados para encontrar patrones y tendencias en los datos. Se encuentra en la intersección entre las matemáticas y estadística con la ingeniería de software y ciencias de la computación.



¿Qué es Aprendizaje No Supervisado?

- Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori.
- Su objetivo no se enfoca en una predicción o una respuesta.



Usos más Frecuentes

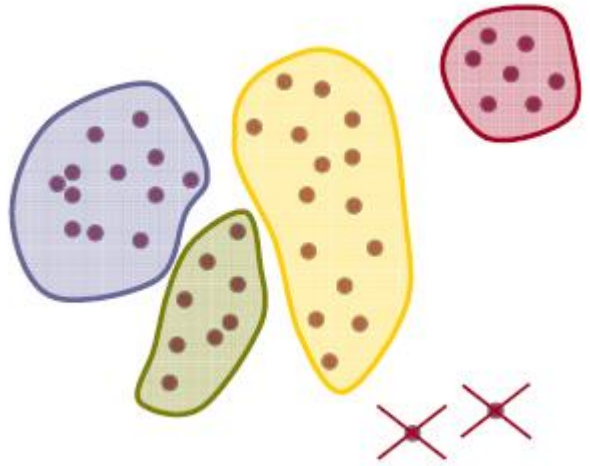
A continuación se enumeran algunas aplicaciones de aprendizaje no supervisado:

1. Clustering: Permite dividir los datos en diferentes grupos en función de su similaridad.

2. Detección de Anomalías: Permite descubrir observaciones o datos inusuales, es muy útil en la búsqueda de actividad fraudulenta.

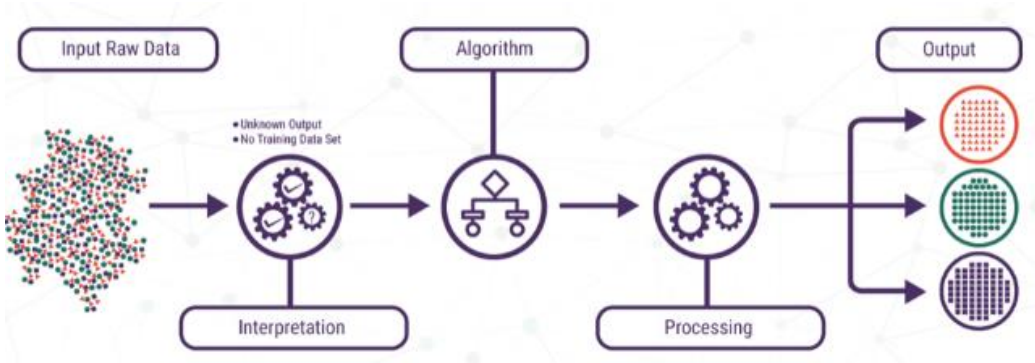
3. Asociación: Identifica objetos u eventos que frecuentemente ocurren juntos, el análisis de canasta es el ejemplo principal.

4. Variables Latentes: Técnicas que se aplican durante la fase de pre-procesamiento tales como la reducción de variables en un dataset



Usos del Aprendizaje No supervisado

- la Recuperación de Información y la Minería de Textos
- El seguimiento y detección de sucesos en un flujo continuo de noticias
- La segmentación de imágenes
- La segmentación o perfilamiento
- La compresión de datos
- El procesamiento de bases de datos espaciales
- La clasificación de zonas geográficas
- La comprensión de imágenes de satélites
- La prospección geológica
- y en muchas otras aplicaciones como la estructuración de grandes volúmenes de datos.



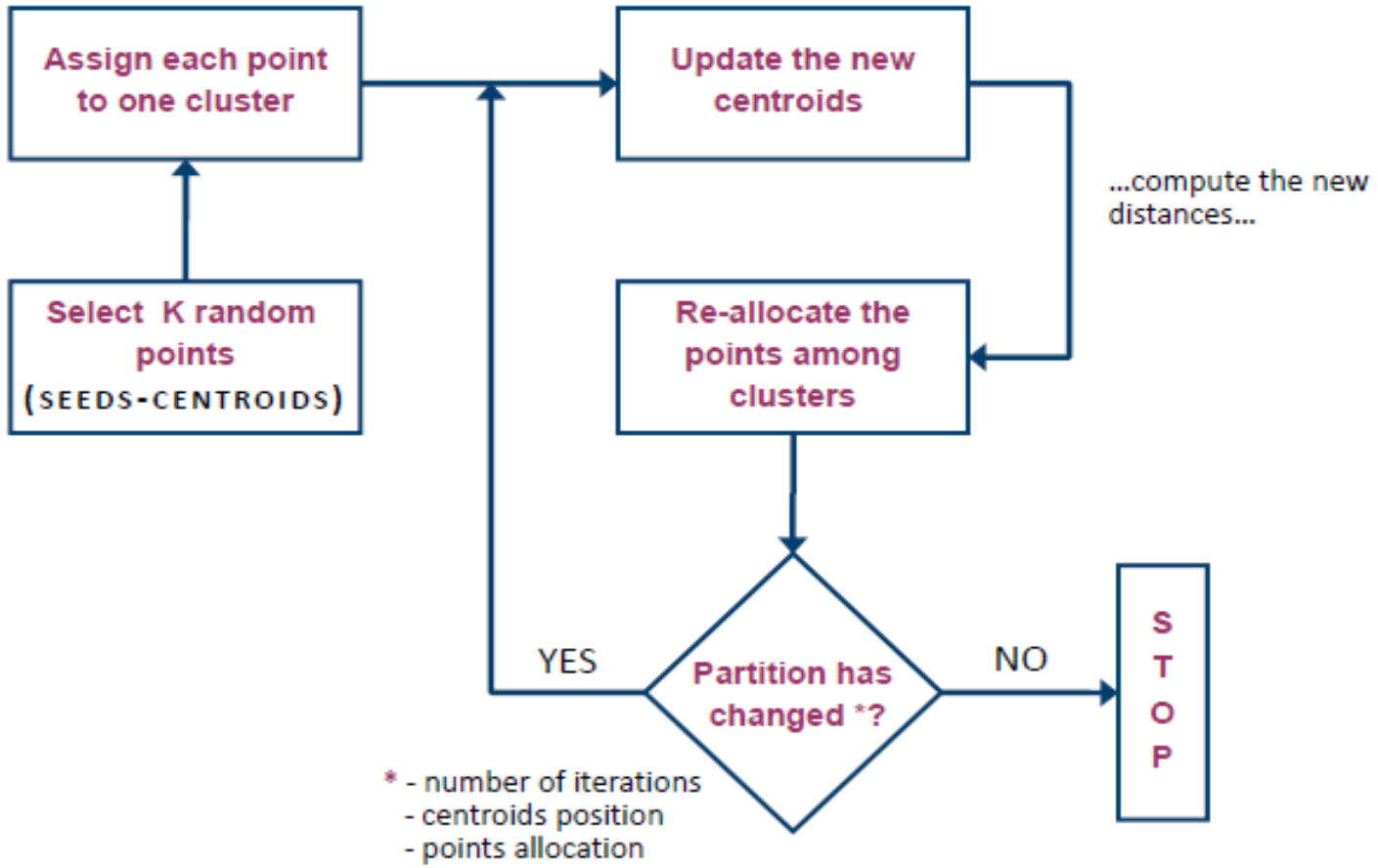
K-Means

K-Means

El algoritmo de las K-means (presentado por MacQueen en **1967**) es uno de los algoritmos de aprendizaje no supervisado más simples para resolver el problema de la clusterización.

K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

El procedimiento aproxima por etapas sucesivas un cierto número (prefijado) de clusters haciendo uso de los centroides de los puntos que deben representar.



https://es.wikipedia.org/wiki/K-means

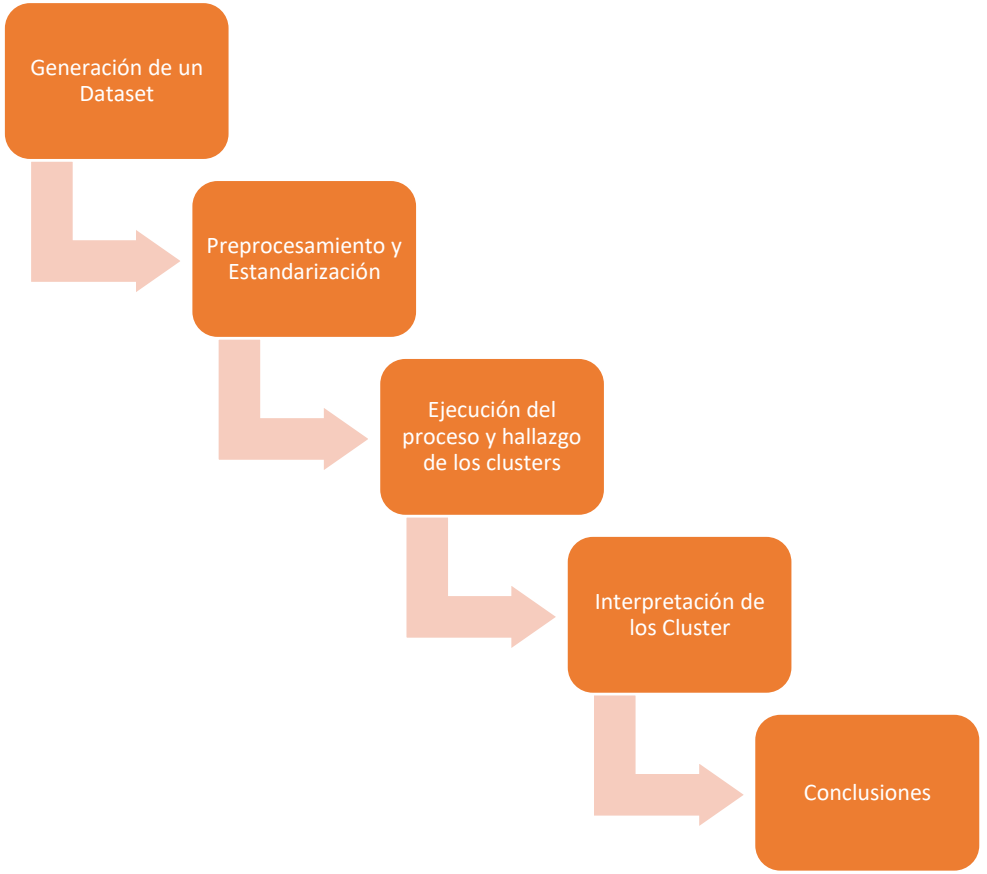
Análisis de Clusters

- Los algoritmos de clusterización buscan cumplir con 3 requerimientos primordiales:

1. **Flexibilidad**: Se debe poder incluir atributos numéricos y categóricos.

2. **Robustez**: Estabilidad en los clusters ante cualquier ruido.

3. **Eficiencia**: Tiempos adecuados de procesamiento.



Transformación de los Datos

Transformación con la Mediana y MAD

Este método es más robusto que la transformación lineal (Z), se debe extraer la mediana de cada valor y luego hay que dividirlo entre la desviación absoluta media

Median / MAD

$$Z = \frac{X - Me}{MAD}$$

$$MAD = median(|x_i - \tilde{x}|)$$

Transformación Lineal

El método más utilizado se conoce como el proceso de transformación lineal o recentralización (Recenter en Ingles), dicho proceso debe extraer primero la media y luego la desviación de cada variable.

Recenter – Z score

$$Z = \frac{X - \mu}{\sigma}$$

Transformación Logarítmica

Este método es muy utilizado cuando estamos ante escenarios cuya distribución de datos presentan un sesgo elevado ya sea a la izquierda o a la derecha.

Esta transformación puede requerir más trabajo ya que los valores que produce pueden tender al infinito.

Transformación con mínimos y máximos

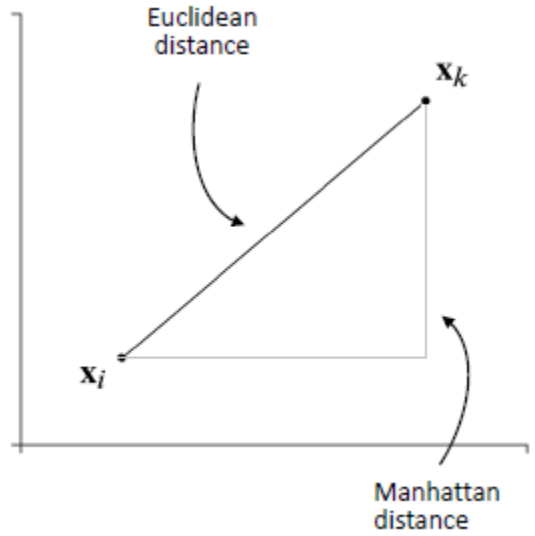
Este es otro método de transformación y su resultado nos devuelve un dataset cuyas variables numéricas están en el rango de 0 a 1.

Scale [0-1]

$$Z_i = \frac{X_i - min\,(x)}{max(x) - min\,(x)}$$

Tipos de Distancias

Distancia	Definición
Distancia Euclideana	Distancia proveniente de la raíz cuadrada entre 2 vectores.
Distancia Máxima	Distancia máxima entre 2 componentes de X y Y
Distancia Manhattan	Distancia absoluta entre 2 vectores
Distancia Canberra	Distancia Manhattan ponderada
Distancia Binaria	Los vectores son tratados como bits, si un elemento tiene valor se representa con un 1, de lo contrario son 0
Distancia Pearson	Distancia de tipo Euclidea, conocida como Pearson No Centrada sum(x_i y_i) / sqrt [sum(x_i^2) sum(y_i^2)]
Distancia por Correlación	Tambien conocida como Pearson Centrada. 1 - corr(x,y)
Distancia Spearman	Calcula la distancia basada en un ranking



Determinando K

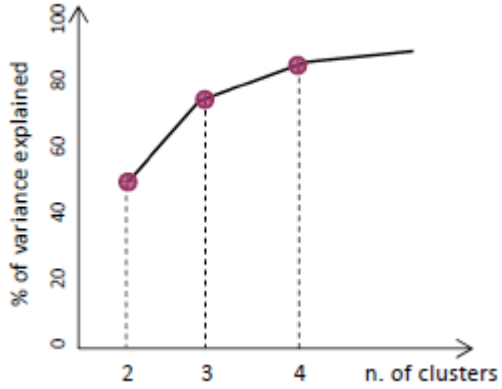
DataSphere
Let's generate value

Número de Clusters

- Regla del Pulgar: Corresponde a la raíz cuadrada del total de observaciones dividido por 2.

$$K \approx \left(\frac{m}{2}\right)^{1/2}$$

- Método del Codo (Elbow Method): El porcentaje de varianza que se puede explicar esta en función del número de clusters.



- Método de Silhouette: Silhouette indica que tan similar es una observación con respecto a las demás observaciones del mismo clusters con respecto a otros clusters.

$$\text{silh}\,(\mathbf{x}_i) = \frac{v_i - u_i}{\max\,(u_i, v_i)}$$

- within [-1,1]
- the closer to 1 the better
- average silhouette

DataSphere
Let's generate value

Validación y Recomendaciones

Criterios de Validación

- **Cohesión:** El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster, la afinidad entre las observaciones se puede medir mediante el coeficiente de Jaccard o bien el coeficiente de afinidad.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

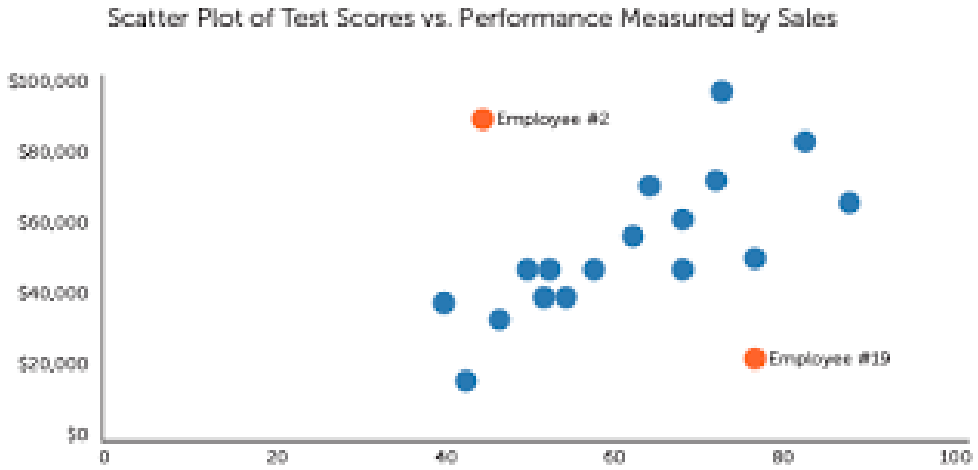
- **Separación:** Los clústeres deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster:
  - Distancia entre el miembro más cercano,
  - Distancia entre los miembros más distantes
  - Distancia entre los centroides.

$$SSB = \sum_{j=1}^{k} n_j \, dist^2 \, (c_j - \bar{x})$$

Recomendaciones

Pre-procesamiento:

- Estandarizar los datos
- Descartar los outliers



Post-procesamiento:

- Descartar los clusters pequeños que pueden representar outliers
- Re-clusterizar los clusters que son de gran tamaño
- Unificar los clusters que son cercanos

Q & A

DataSphere
Let's generate value

Bibliografía

- Unsupervised Learning with R
- By Erik Rodríguez Pacheco, 2015

- Data Clustering
- By Charu C. Aggarwal; Chandan K. Reddy, 2016

- A Framework for Analysis of Data Quality Research
- by Richard Y. Wang,1995

- An Introduction to Data Cleaning with R
- by Edwin de Jonge & Mark van der Loo,2013

DataSphere
Let's generate value

Learning Path: Machine Learning

Unsupervised Learning	• 16 Hrs (Clustering, Hierarchical, Association, PCA)
Supervised Learning	• 20 Hrs (Regression, Naive Bayes, KNN, Logistic)
Supervised Learning: Trees	• 16 Hrs (Decision Trees, Random Forest)
Time Series	• 8 Hrs (ARMA, ARIMA, Prophet)

DataSphere
Let's generate value

Learning Path: Data Visualization

Fundamentals	• 16 Hrs (Power BI)
Essentials Design Principles	• 12 Hrs (PowerBI)
Visual Analytics	• 12 Hrs (Power BI)
Dashboard – Story Telling	• 8 Hrs (Power BI)

DataSphere
Let's generate value

Learning Path: Big Data

Big Data Ingestion	• 16 Hrs
Hive/Impala	• 16 Hrs
Spark	• 16 Hrs
Oozie/Nifi	• 16 Hrs

Contáctanos

- Ventas y Servicios : ventas@datasphere.tech
- Mkt & Clients Director : edenilson.ruiz@datasphere.tech
- Chief Analytics Officer : nelson.zepeda@datasphere.tech

- https://www.datasphere.tech

- https://www.linkedin.com/company/datasphere-consulting

- https://twitter.com/datasphere_cns

- https://www.facebook.com/DataSphereSV/

DataSphere

Let's generate value