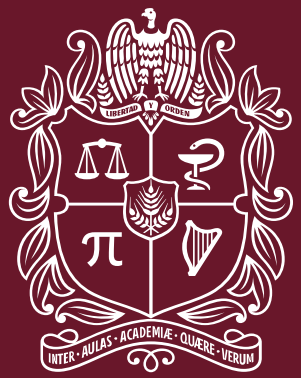


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Clustering – Aprendizaje No Supervisado

3007855 - Inteligencia Artificial
3010476 - Introducción a la Inteligencia Artificial

Semestre: 02/2021

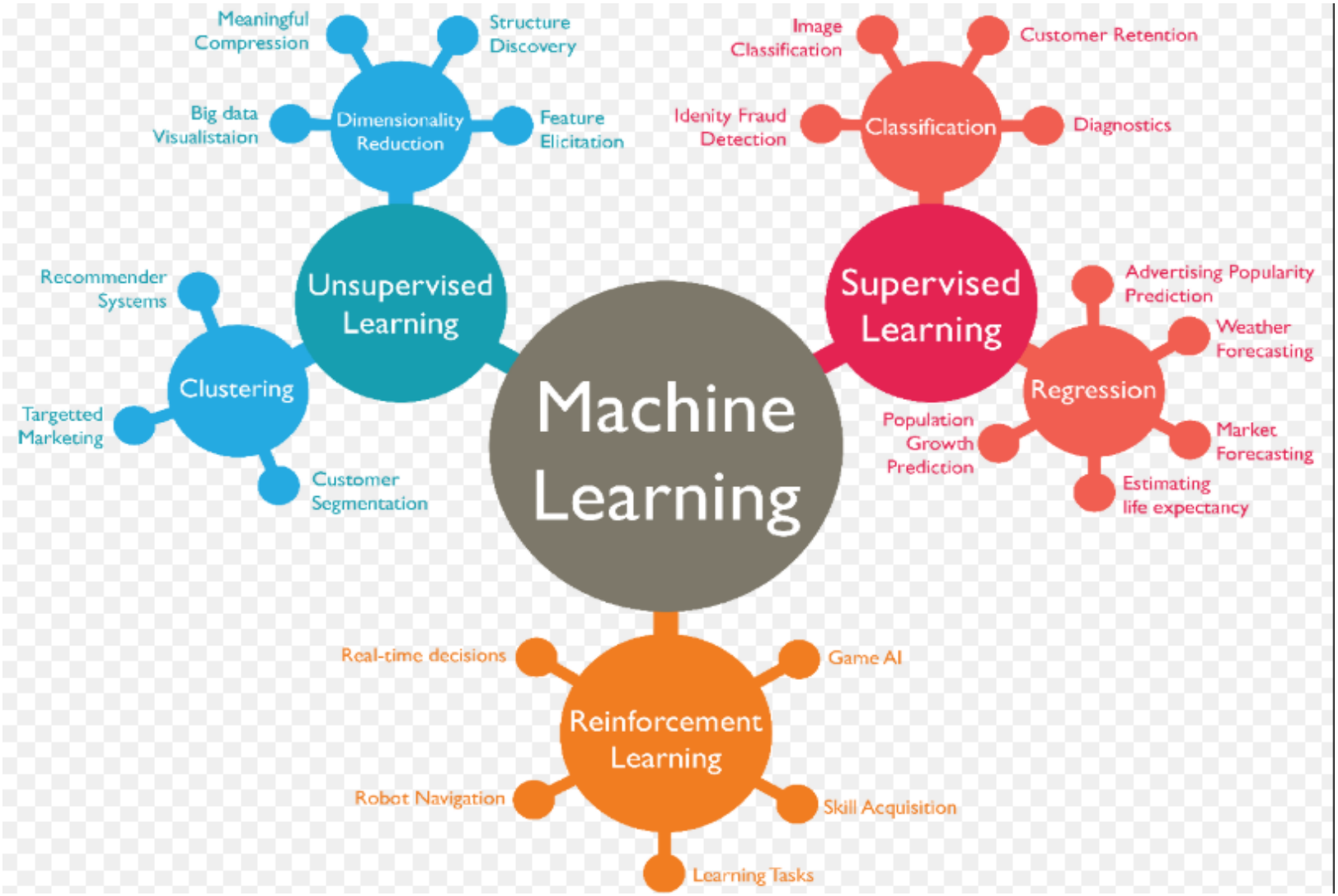
Prof. Demetrio Arturo Ovalle Carranza
Departamento de Ciencias de la Computación
y de la Decisión
Facultad de Minas

Noviembre 16 de 2021

LMS: https://minaslap.net/user/index.php?id=560

Link Clases: meet.google.com/quy-okvi-ugq

TECNICAS DE MACHINE LEARNING



APRENDIZAJE NO SUPERVISADO

El sistema intenta descubrir la estructura oculta de los datos o asociaciones entre variables (patrones).

En ese caso, los datos consisten en instancias sin ninguna etiqueta correspondiente.

Enfoques: Clustering, Reglas de Asociación (e.g. Análisis de secuencias biológicas), etc.

CLUSTERING

- Se realiza una agrupación, es decir, una a partir de un conjunto de datos completo se realiza una separación en grupos de datos.
- Se busca que las instancias de datos que pertenecen al mismo grupo sean lo más similar posible, mientras que aquellas que pertenecen a diferentes grupos difieran tanto como sea posible.

APLICACIONES

- Segmentación de perfiles de usuario:
  - ✓ Segmentación por historial de compras
  - ✓ Segmentación por actividades en la organización, sitio web o plataforma
  - ✓ Clasificación de personas según intereses.
  - ✓ Creación de perfiles basados en el monitoreo de una actividad.

- Categorización de inventarios:
  - ✓ Agrupación de inventarios basada en actividad de ventas
  - ✓ Agrupación de inventarios basada en métricas de fabricación

APLICACIONES

- Clasificación de datos de un sensor:
  - ✓ Detección de tipos de actividad en sensores de movimiento
  - ✓ Imágenes grupales
  - ✓ Audio separado
  - ✓ Identificación de grupos en monitoreo de salud.
- Detección de anomalías y/o fraudes:
  - ✓ Creación de grupos de actividades que causan anomalías
  - ✓ Encontrar transacciones bancarias que pertenecen a diferentes grupos e identificarlas como fraudulentas.
  - ✓ Agrupación de actividades para limpieza de valores atípicos al detectar anomalías.
- Análisis de Imágenes y documentos:
  - ✓ Agrupar datos, documentos, etc. que tienen características similares.

Conceptos

- Datasets
- Histogramas de distribución de los datos
- Curva de codo (Elbow Curve) para determinar el número de clusters
- Diagramas de Voronoi
- Método K-means y centroides

Clustering K-means

- ✓ El clustering k-means es un método de cuantización vectorial, originalmente usado para procesamiento de señales y es popular para el análisis de clusters en minería de datos.
- ✓ Su objetivo es particionar n observaciones en k grupos o clústeres en los cuales cada observación pertenezca a un clúster con la media más cercana.
- ✓ La media sirve como modelo o prototipo para el clúster. Esto resulta en una partición del espacio de los datos en celdas Voronoi.
- ✓ Resolver un algoritmo k-means implica el análisis profundo de los datos y permite encontrar soluciones óptimas al utilizar algunas heurísticas.

Diagrama de Voronoi

- ✓ En matemáticas, un diagrama de Voronoi es una partición del plano en regiones.
- ✓ Se basa en la distancia de los puntos con respecto a un subconjunto especifico del plano.
- ✓ La cantidad de conjuntos de puntos se define de antemano y por cada punto existe una región de puntos más cercanos que corresponde a ese punto con respecto a otros puntos.
- ✓ Estas regiones se conocen como celdas de Voronoi.

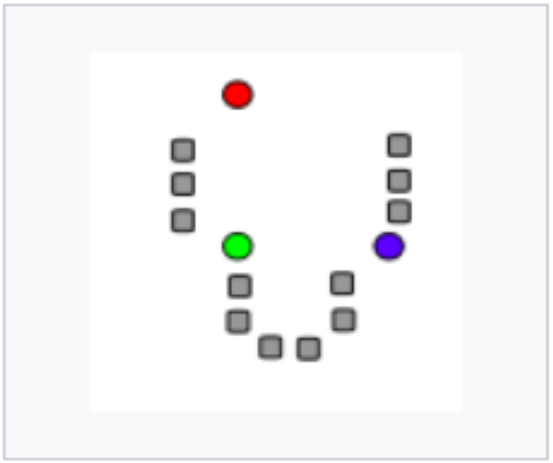
20 puntos con sus celdas Voronoi

Clustering K-means

Se parte de un vector X con un conjunto de n datos. Luego se ejecuta el algoritmo de la siguiente forma:

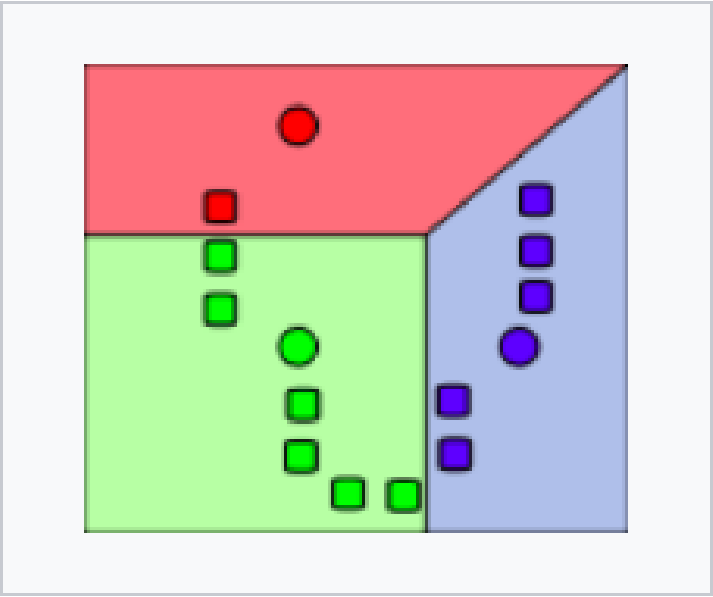
1. Se seleccionan k datos como centroides del vector X.
2. Asignar los otros datos al centroide más cercano. Cada nuevo grupo de datos se conoce como clúster.
3. Por cada clúster, encontrar un nuevo centroide más preciso (medida de similitud más acorde con el grupo de datos), calculando un nuevo centro entre los puntos.
4. Repetir los pasos 2 al 3 hasta que los centroides dejen de cambiar.

Demostración del algoritmo



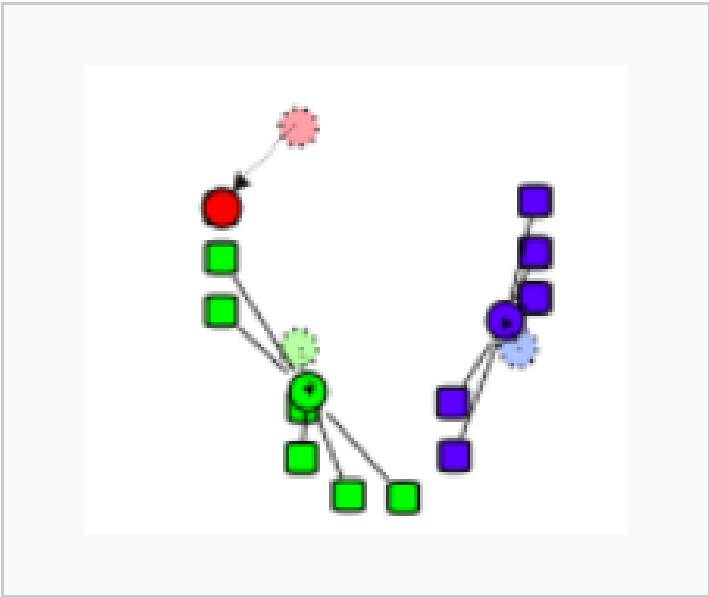
- ✓ Se define un k inicial, en este caso 3.
- ✓ Luego se genera aleatoriamente los tres clusters con sus medias usando los datos.

Demostración del algoritmo



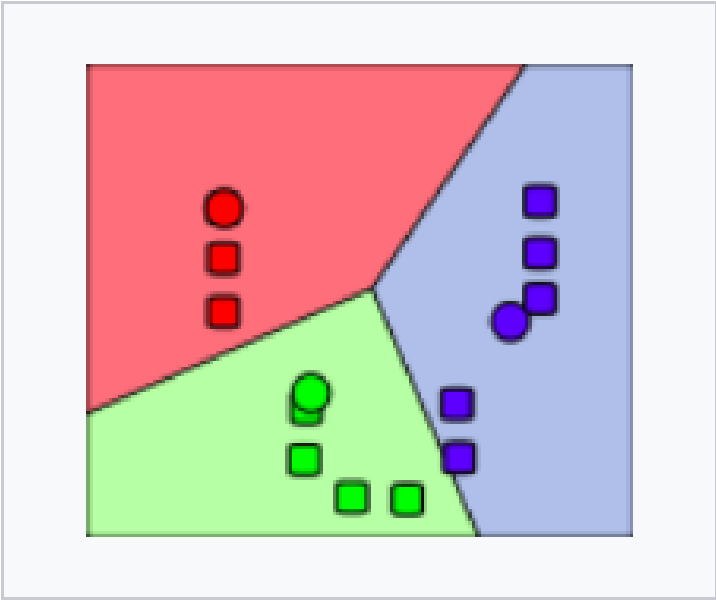
- ✓ Se crean k clusters, ejm. 3, al asociar cada observación con la media más cercana.
- ✓ Las particiones mostradas usan el diagrama de Voronoi, estas particiones son generadas gracias a la media.

Demostración del algoritmo



- ✓ El centroide de cada uno de los k clusters, se convierte en la nueva media.

Demostración del algoritmo



Se repiten los pasos 2 y 3 hasta que que se alcance la convergencia.

- ✓ **Paso 2:** Asignar los otros datos al centroide más cercano. Cada nuevo grupo de datos se conoce como cluster.
- ✓ **Paso 3.** Por cada cluster, encontrar un nuevo centroide más preciso (medida de similitud más acorde con el grupo de datos), calculando un nuevo centro entre los puntos.

Clustering k-means Personalidad

Objetivo: Agrupar usuarios de Twitter según su personalidad usando Clusterización (clustering) k-means

analisis.csv → 140 registros con características de personalidades
usuario,op,co,ex,ag,ne,wordcount,categoria

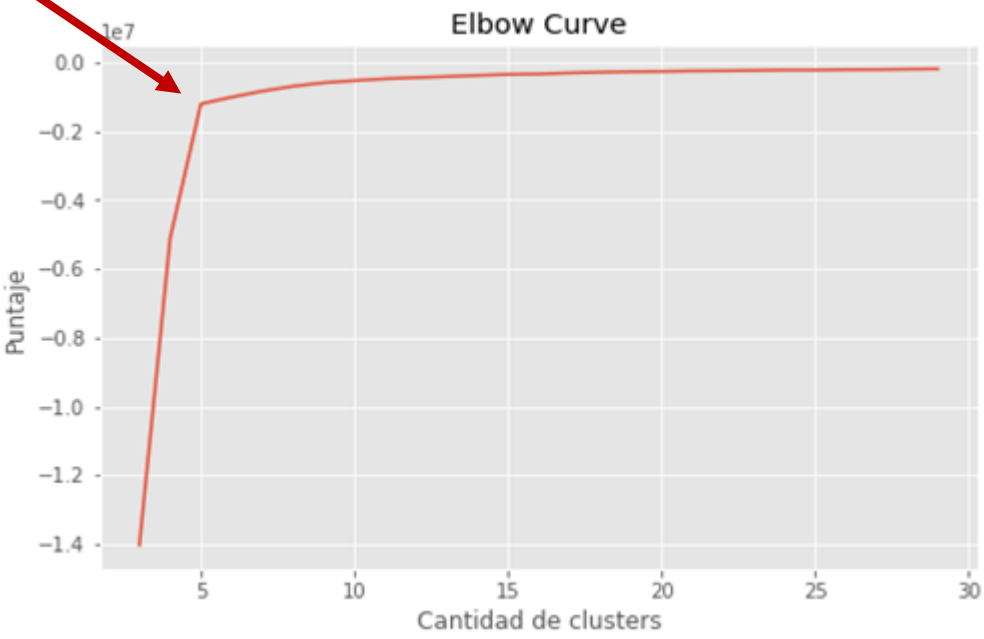
categoria son 9:
1. Actor/actriz (muestra:27)
2. Cantante (muestra:34)
3. Modelo (muestra:9)
4. Tv, series (muestra:19)
5. Radio (muestra:4)
6. Tecnología (muestra:8)
7. Deportes (muestra:17)
8. Política (muestra:16)
9. Escritor (muestra:6)

Variables:
op (openness - apertura mental),
ex (extraversion – grado de extraversión)
ag (agreeableness - amabilidad)
co (conscientiousness – escrupulosidad)
ne (neurocitismo – instabilidad emocional)
wordcount (# promedio de palabras del mensaje)
categoria (9 categorías de personajes)

Determinación del Número de Clusters: Método 1. Curva de Codo (Elbow Curve)

Punto de Inflexión

Método estadístico que utiliza la **distancia media de las observaciones a su centroide**. Cuanto menor es la distancia intra-cluster mejor, ya que significa que los clústers son más compactos. El método del codo busca el valor k que satisfaga que un incremento de k, no mejore sustancialmente la distancia media intra-cluster. No es un método exacto, sin embargo, es potencialmente útil.



NOTA: El término distancia se emplea en la técnica de ***clustering*** como cuantificación de la similitud o diferencia entre observaciones.

Método 2. Estadístico de Gap

El **estadístico *gap*** fue publicado por *R.Tibshirani, G.Walther y T. Hastie*, autores del libro ***Introduction to Statistical Learning***.

Este estadístico compara, para diferentes valores de *k*, la varianza total *intra-cluster* observada frente al valor esperado acorde a una distribución uniforme de referencia.

La estimación del número óptimo de *clusters* es **el valor *k* con el que se consigue maximizar el estadístico *gap***, es decir, encuentra el valor de *k* con el que se consigue una estructura de *clusters* lo más alejada posible de una distribución uniforme aleatoria. Este método puede aplicarse a cualquier tipo de *clustering*.

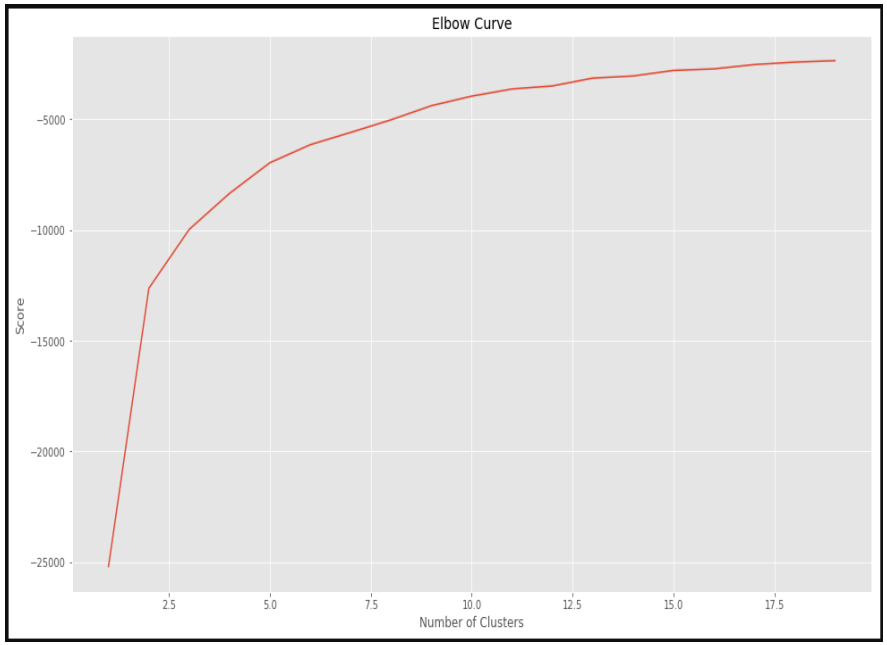
Clustering k-means Personalidad

op (apertura mental),
ex (extraversion)
ag (amabilidad)

140 registros con personalidades
9 categorías son:

1. Actor/actriz (muestra:27)
2. Cantante (muestra:34)
3. Modelo (muestra:9)
4. Tv, series (muestra:19)
5. Radio (muestra:4)
6. Tecnología (muestra:8)
7. Deportes (muestra:17)
8. Politica (muestra:16)
9. Escritor (muestra:6)

Curva de codo para clusters



Clusters = 5

Clustering k-means Personalidad

op (apertura mental),
ex (extraversion)
ag (amabilidad)

Centroides

[[39.86685966 45.20847056 25.30614166]
[49.80086386 40.8972579 17.48224326]
[34.82702519 47.11690063 34.66889141]
[42.302263 33.65449587 20.812626]
[58.58657531 31.02839375 15.6120435]]

140 datos , Clusters = 5