PROYECTO APLICADO

"You can't manage what you don't measure",
Tom De Marco, 1982

"We´re drowning in data but starving for knowledge",
John Naisbitt, 1982

UNIVERSIDAD ICESI

AGENDA

- Aprendizaje no supervisado
- Clustering por distancia
- Evaluación de clustering
- Análisis de componentes principales
- Proyecto de aprendizaje no supervisado

APRENDIZAJE NO SUPERVISADO

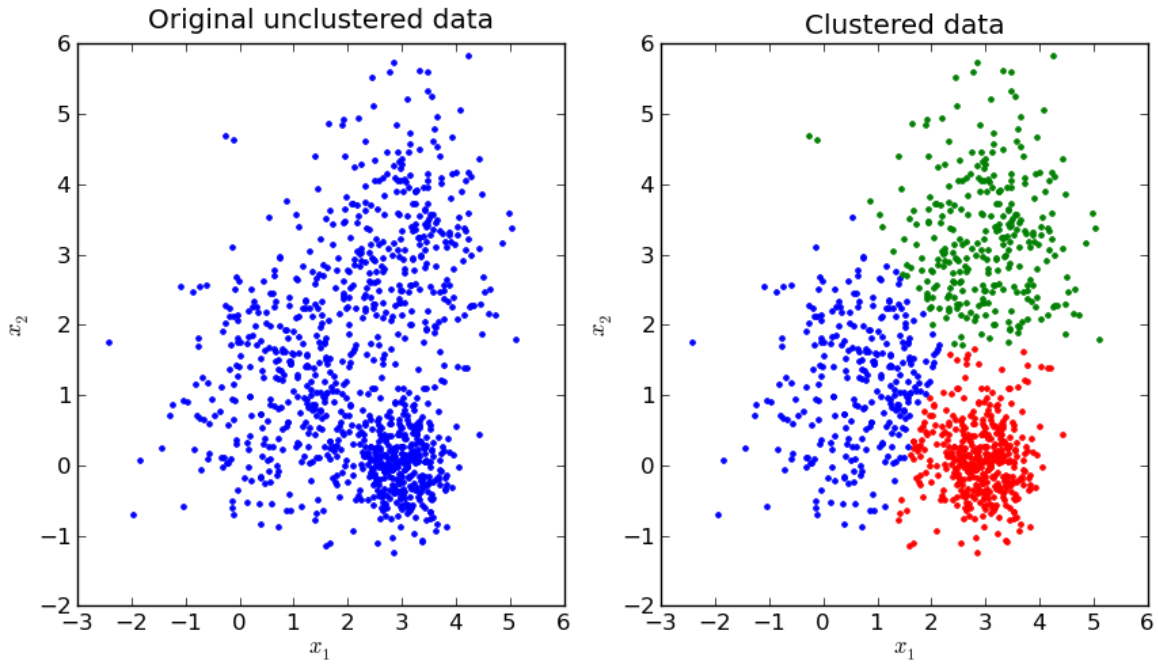
- No se interesa por la predicción sino por encontrar una estructura, un nuevo punto de vista, una simplificación o un resumen de los datos
- Usualmente se incluye en la fase exploratoria de datos
- Tipos de tareas:
  - Segmentación (clustering)
  - Cambio de representación (e.g. reducción de dimensiones, selección de factores)
  - Reglas de asociación
  - Detección de anomalías (i.e. excepciones)
- Difícil de validar los resultados, ya que no se cuenta con un "gold standard"

CLUSTERING POR DISTANCIA

- Objetivo: descubrir **k** **grupos** o **segmentos** desconocidos que
  - Minimicen la distancia dentro de los grupos
  - Maximicen la distancia por fuera de los grupos
- Se basan en una noción de **distancia**
  - Definición de la medida a utilizar
  - Unidades de los atributos tienen gran influencia

  →**Normalizar**

  →**Estandarizar**



http://pypr.sourceforge.net/kmeans.html

UNIVERSIDAD ICESI

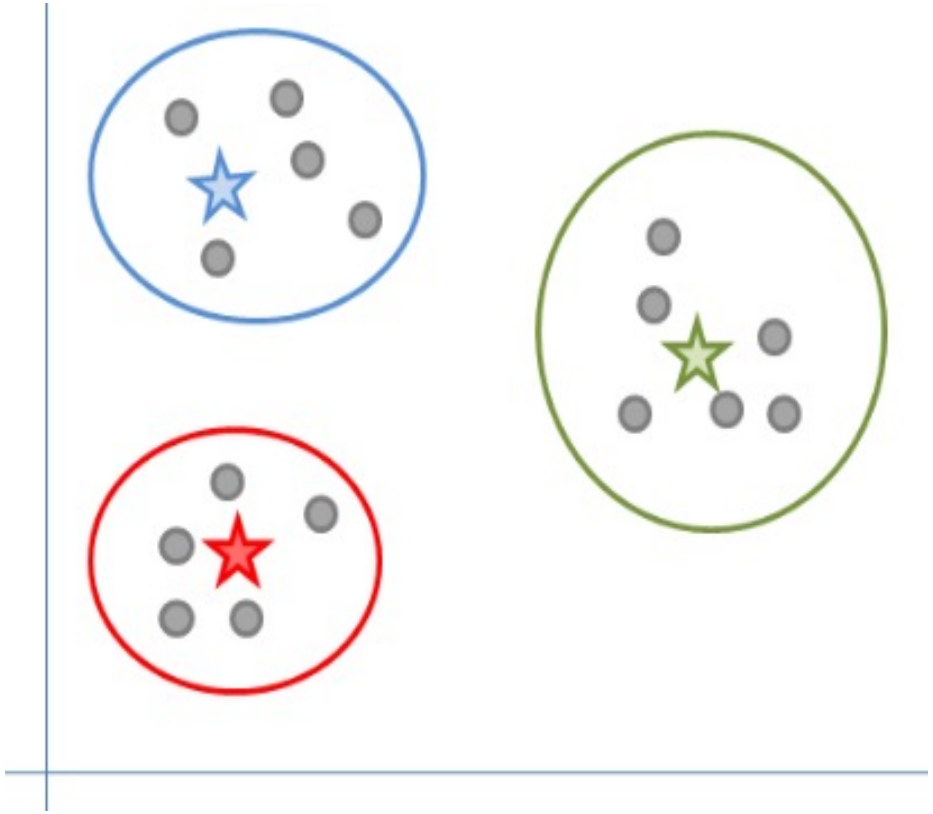
MODELOS DE CLUSTERING POR DISTANCIA

- **K-means**
- **Clustering jerárquico**

K-MEANS

- Algoritmo:
  1. Inicializar los K centroides
  2. Asignar cada instancia al cluster del centroide más cercano
  3. Re calcular los centroides de cada cluster (el baricentro/promedio)
  4. Repetir pasos 2 y 3 hasta convergencia (hasta que los centroides permanezcan estáticos)
- Cada observación se asigna a un solo cluster, de manera absoluta
- Los clusters no se sobrelapan
- Objetivo: minimizar la variación dentro de los clusters (Within Sum of Squares - WSS):

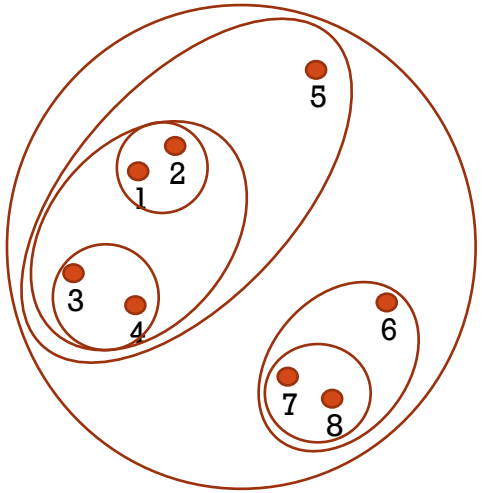
$$WSS = \sum_{i=1}^{\#instancias} distancia(\boldsymbol{x_i} - centroide(\boldsymbol{x_i}))^2$$

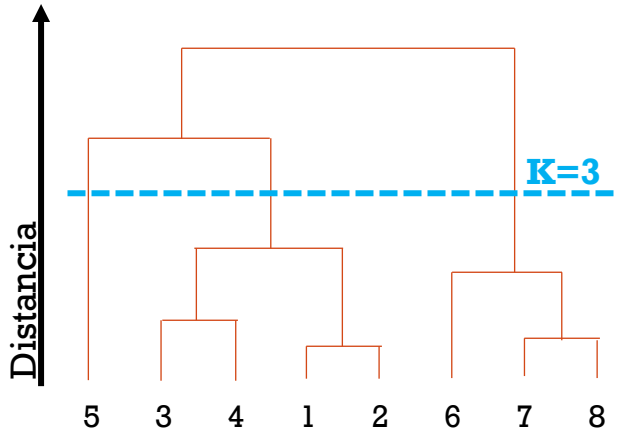


UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

- Algoritmo (iterativo):
  1. Al inicio cada instancia es un cluster (n clusters)
  2. Se identifica el par de clusters más cercanos y se fusionan (n-1 clusters)
  3. Se repite el paso anterior hasta que queda un solo cluster con todas las instancias
  4. Se escoge un punto de corte
- Los clusters se pueden organizar en forma de **dendrograma**
- Es necesario definir como **fusionar** clusters y la **distancia** a utilizar

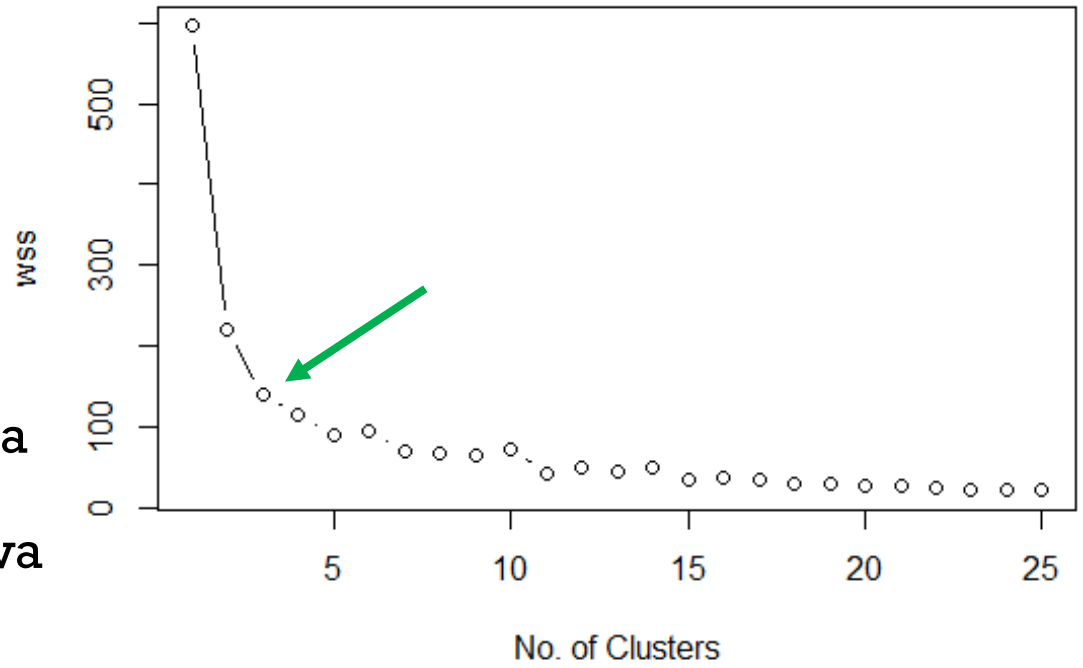




UNIVERSIDAD ICESI

ESCOGENCIA DEL K – CODO

- Heurísticos:
  - No hay un método absoluto
  - Dependen del juicio del analista, se requiere conocimiento del negocio
- Método "del codo":
  - Plotear WSS para cada valor de K
  - Escoger el último valor de K que implica una reducción "considerable" del WSS del clustering resultante, cuando la curva se vuelve aproximadamente lineal



$$WSS = \sum_{i=1}^{\#instancias} distancia(\boldsymbol{x_i} - centroide(\boldsymbol{x_i}))^2$$

UNIVERSIDAD ICESI

ESCOGENCIA DEL K – SILHOUETTE

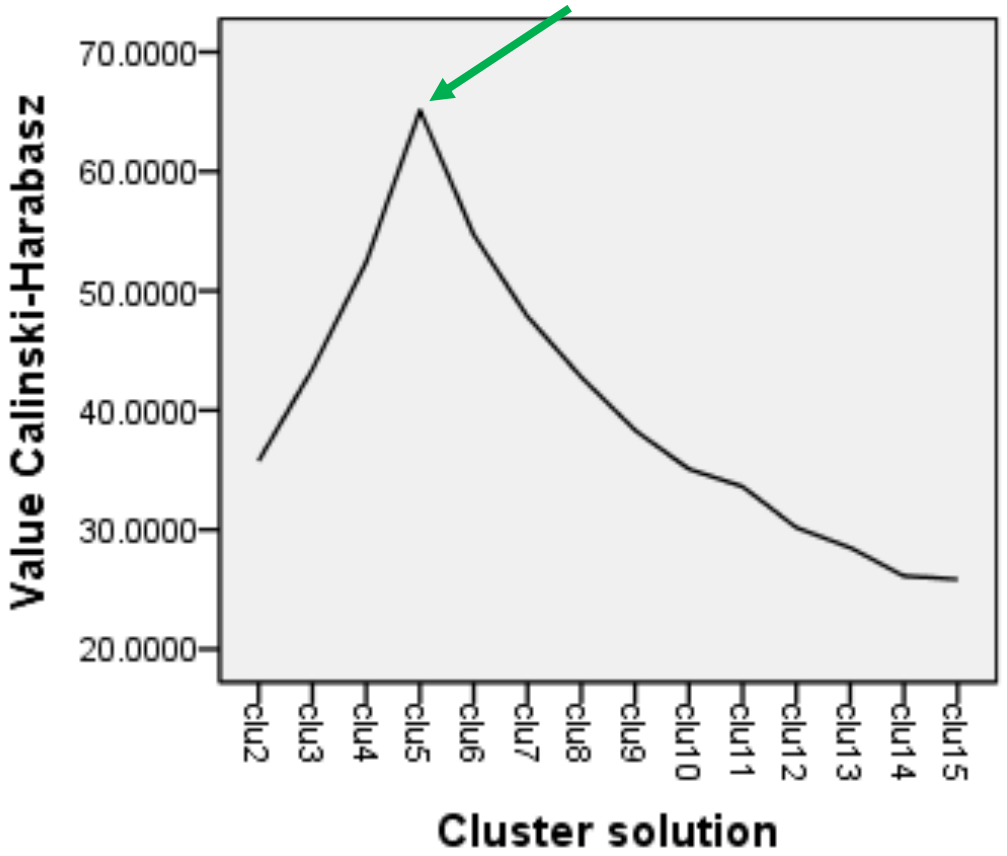
- Método Silhouette
  - Se busca el K que maximice la **separación** entre clusters, con clusters lo más **compactos** posibles
  - Analizar el ajuste de cada instancia al cluster al que fue asignado
  - Qué tan cerca está cada observación de las demás de su propio cluster
    - 0,7-1,0: el cluster es fuertemente robusto
    - 0,5-0,7: el cluster es razonablemente robusto
    - 0,25-0,5: el cluster puede ser artificial y puede no denotar una noción de estructura necesariamente
    - Inferior a 0,25: el cluster debería descartarse, no indica estructura
  - Se busca la maximización del valor Silhouette promedio de los clusters

UNIVERSIDAD ICESI

ESCOGENCIA DEL K – CALINSKI-HARABASZ

- Método de Calinski-Harabasz:
  - Se busca el K que maximice la **separación** entre clusters, con clusters lo más **compactos** posibles
  - TSS = variación total (entre todos los datos y el centro global)
  - WSS = variación intra-cluster (entre los puntos de cada cluster y sus centroides
  - BSS = variación inter-cluster (entre los centroides de los clusters y el centro global). BSS = TSS - WSS
  - CH = ratio entre la variación entre clusters (BSS) y el promedio de la variación interna de los clusters (WSS). Se busca maximizar CH:

$$CH = \frac{BSS}{WSS} * \frac{N-k}{k-1}$$



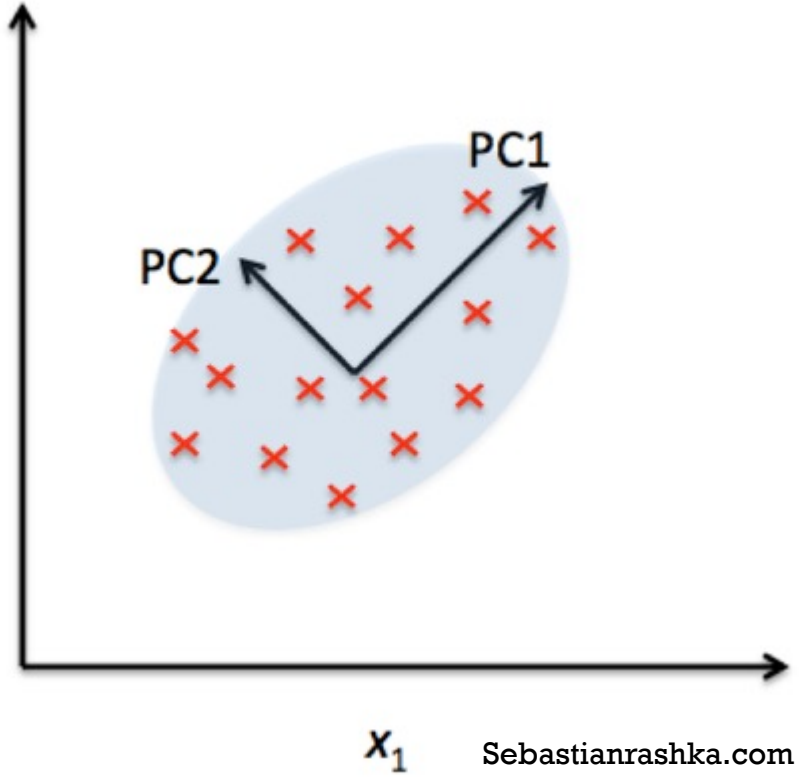
UNIVERSIDAD ICESI

COMPONENTES PRINCIPALES

PCA: Principal Component Analysis

Objetivo: Simplificar el dataset, encontrando una representación de **baja dimensionalidad** que conserva la mayor parte de la información

- **Combinación lineal** de las dimensiones (atributos) originales del dataset que maximiza la varianza
- **Rotación** de los ejes originales
- Permite una **visualización** los datos en problemas de aprendizaje supervisado y no supervisado
- Se limitan las dimensiones que estén altamente **correlacionadas** entre ellas
- PCA permite encontrar la superficie lineal de menos dimensiones más cercana a los puntos en el espacio original (en distancia Euclidiana)



UNIVERSIDAD ICESI

Proyecto de aprendizaje no supervisado

- Exploración de datos
  - Estadísticos y visualización
    - Variables numéricas
    - Variables categóricas
    - Pairplots
    - Distribuciones
    - Barplots
    - Correlación de variables
- Implementación de modelos
  - PCA
    - Análisis de los componentes
    - Creación de variables nuevas
    - Caracterización de variables nuevas
  - K-means
    - Evaluación del k
      - Método del codo
      - Método de la silueta
      - Método de Calinski-Harabasz
    - Implementación del modelo
    - Análisis y caracterización de clusters
      - Diagramas de densidad
      - Scatterplots
      - Polar scatterplots
      - Scatter geo
    - Clustering Jerárquico
      - Diversos enlazamientos
      - Dendograma

      Análisis y caracterización de clusters
      - Diagramas de densidad
      - Scatterplots
      - Polar scatterplots
      - Scatter geo

Ejercicio

Replicar los mismos pasos realizados con el conjunto de datos de países, pero en este caso con respecto al conjunto de datos de salarios de nba.

Con este dataset se caracterizar a los distintos jugadores de la NBA.