UNIVERSIDAD NACIONAL DE COLOMBIA

Diplomado
BIG DATA ANALITYCS
K-Means

Por
Ing. Jorge E. Camargo, Ph.D.

Agenda

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

1. Agrupamiento
2. K-Means
3. Medidas de Validación

Ben Chams - Fotolia

Diplomado BIG DATA ANALITYCS
K-Means

Agrupamiento (Clustering)

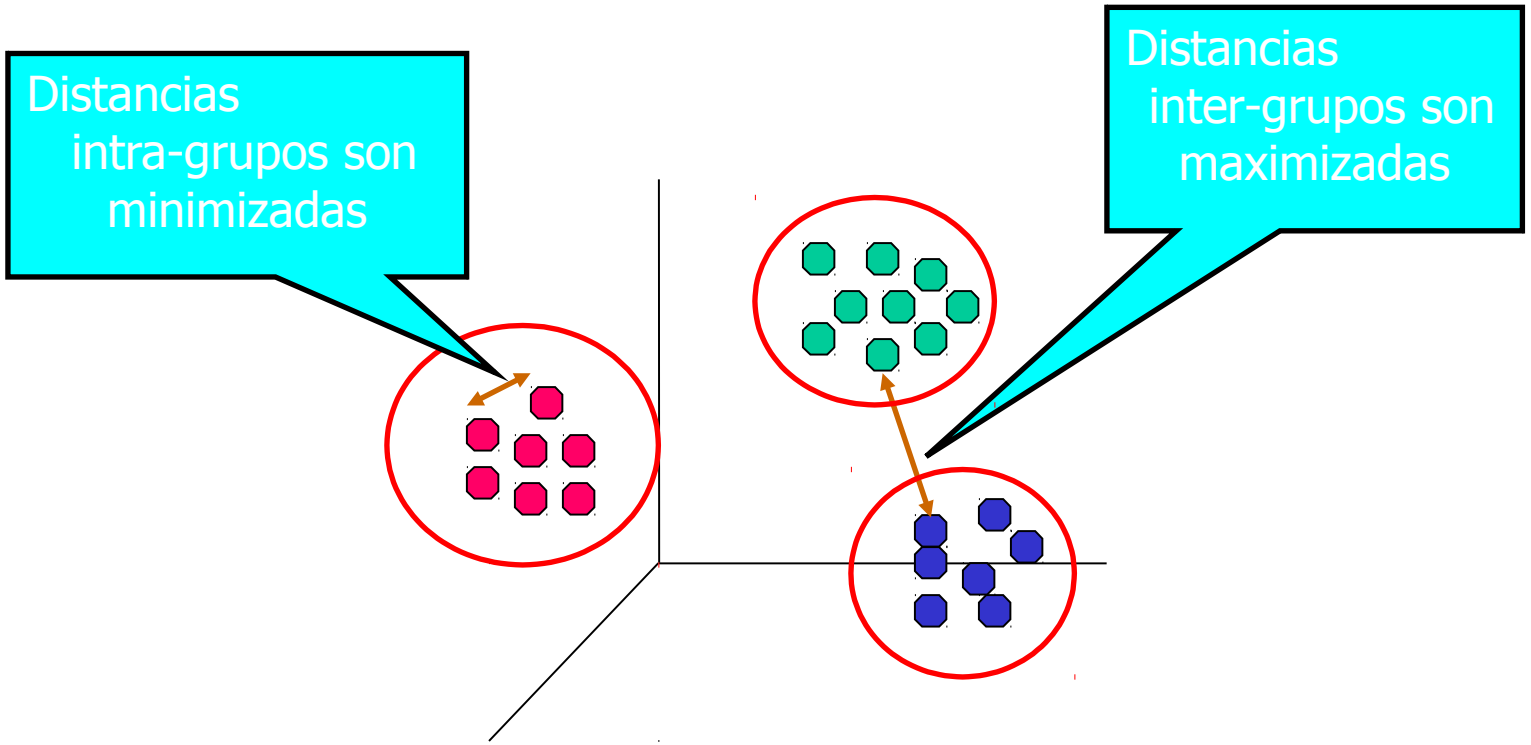
UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Dado un conjunto de puntos/datos, cada uno con un conjunto de atributos, y una medida de similitud entre ellos, encontrar grupos (clusters) tal que:
  - Los puntos dentro de un grupo son más similares entre ellos.
  - Los puntos que pertenecen a diferentes grupos son menos similares entre ellos.
- Medidas de Similitud:
  - Distancia Euclidiana si los atributos son continuos.
  - Otras medidas específicas del problema.

Diplomado BIG DATA ANALITYCS
K-Means

Agrupamiento (Clustering)

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.



Diplomado BIG DATA ANALITYCS

K-Means

Agrupamiento - Aplicaciones

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Segmentación de Mercado:
  - Objetivo: Subdividir un mercado en diferentes subconjuntos de clientes donde cada uno pueda ser seleccionado como mercado objetivo que pueda ser alcanzado con distintas estrategias de marketing.
  - Enfoque:
    - Reunir diferentes atributos de los clientes con base en su información geográfica y estilo de vida.
    - Encontrar grupos de clientes similares.
    - Medir la calidad del agrupamiento al observar los patrones de compra de los clientes dentro de un mismo grupo vs. aquellos que pertenecen a otros grupos.

Agrupamiento - Aplicaciones

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Agrupamiento de Documentos:
  - Objetivo: Encontrar grupos de documentos que sean similares entre ellos con base en sus términos importantes.
  - Enfoque: Identificar los términos que aparecen frecuentemente en cada documento. Emplear una medida de similitud basada en las frecuencias de éstos. Usarla para agrupar.
  - Ganancia: En Recuperación de Información se pueden utilizar estos grupos para relacionar un nuevo documento o término de búsqueda con los documentos ya agrupados.

Diplomado BIG DATA ANALITYCS
K-Means

Agrupamiento de Documentos

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Datos a agrupar: 3204 Artículos de Los Angeles Times.
- Medida de Similitud: Qué tantas palabras son comunes en estos documentos (después de algún filtrado).

Categoría	Total Artículos	Bien Ubicados
Finanzas	555	364
Internacional	341	260
Nacional	273	36
Local	943	746
Deportes	738	573
Entretenimiento	354	278

Diplomado BIG DATA ANALITYCS
K-Means

K-Means

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

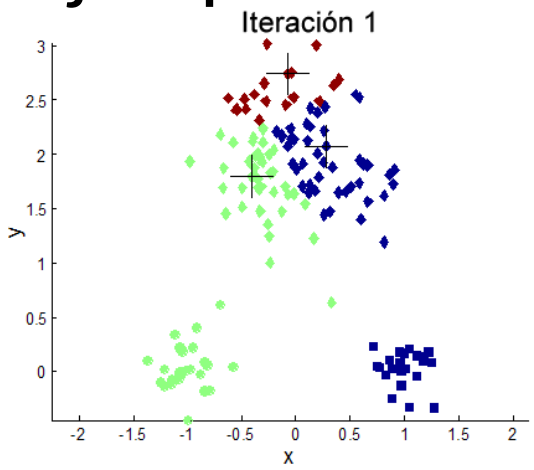
- Enfoque particional
- Cada grupo se asocia a un centroide
- Cada punto se asigna al grupo con el centroide más cercano
- El número de grupos *K* debe especificarse
- El algoritmo base es muy sencillo

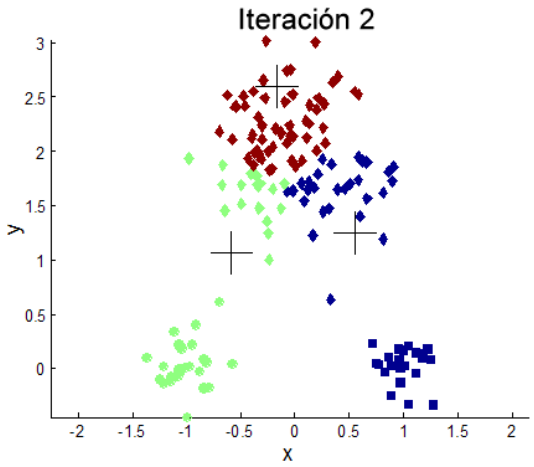
```
1.  Seleccionar K puntos como los centroides iniciales
2.  repetir
3.      Formar K grupos. Asignar cada punto al centroide más cercano
4.      Recalcular el centroide de cada grupo
5.  hasta: Los centroides no cambian
```

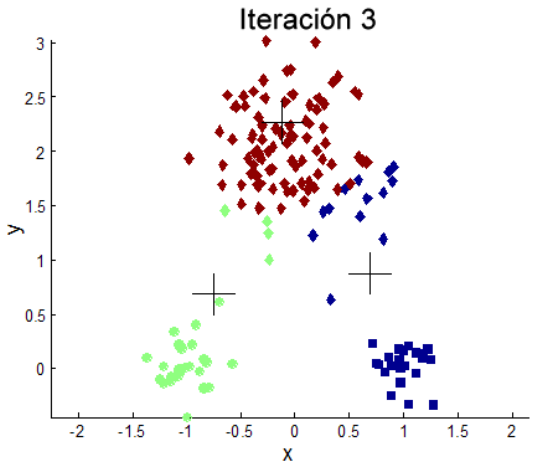
K-Means

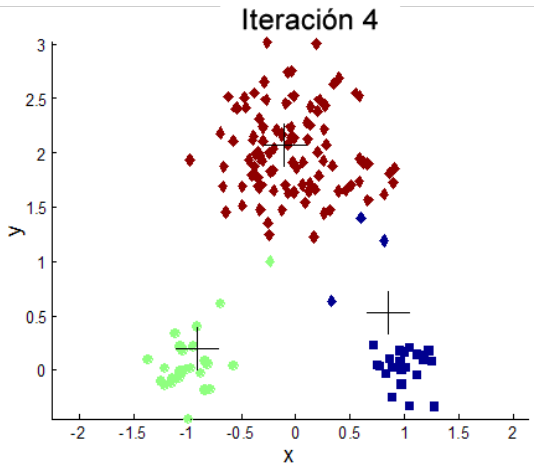
UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

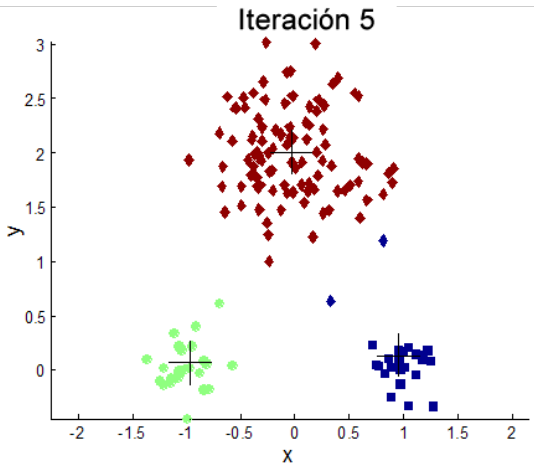
- Ejemplo

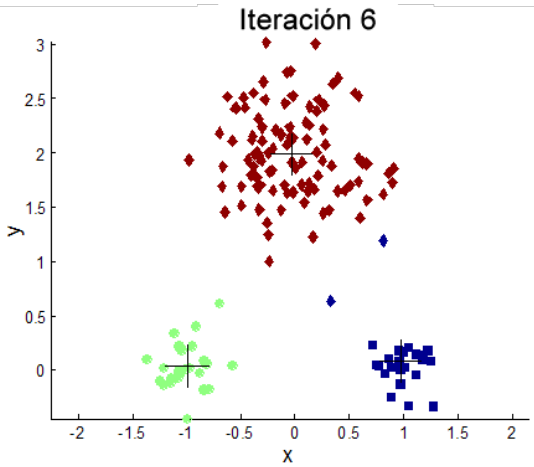












D

K-Means

K-Means

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Los centroides iniciales a menudo son elegidos al azar.
  - Los grupos producidos varían de una ejecución a otra.
- El centroide es (normalmente) la media de los puntos en el grupo.
- La 'cercanía' se mide con la distancia euclidiana, similitud de coseno, correlación, etc.

K-Means

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- K-Means converge con las medidas de similitud mencionadas anteriormente.
- La mayor parte de la convergencia ocurre en las primeras iteraciones.
- A menudo, la condición de parada se cambia por «Hasta que relativamente pocos puntos cambien de grupo“
- La complejidad es O(n * K * I * d)
  - n = número de puntos
  - K = número de grupos
  - I = número de iteraciones
  - d = número de atributos

K-Means. Evaluación de Grupos

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- La medida más común es la Suma de errores cuadrados(SSE)
  - Para cada punto, el error es la distancia al grupo más cercano.
  - Para calcular el SSE, se elevan estos errores al cuadrado y luego se suman.

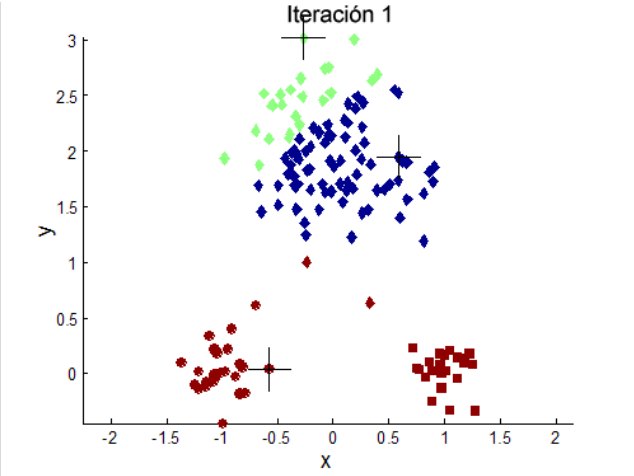
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

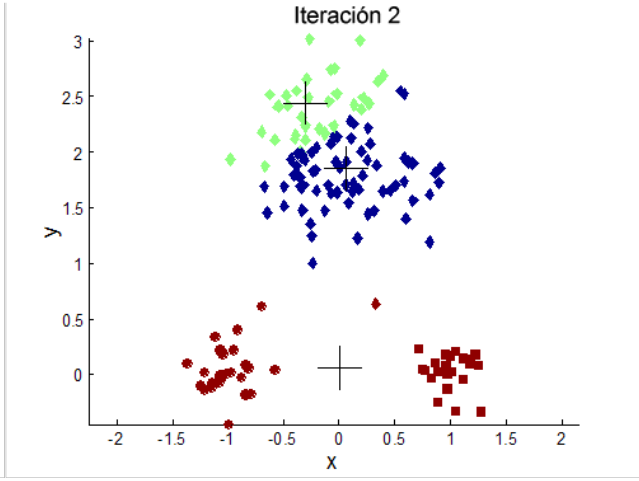
  - x es un punto en el grupo C_i y m_i es el punto representativo del grupo C_i
    - Se puede demostrar que m_i corresponde al centro (media) del grupo
  - Dados dos grupos, se puede elegir aquél con el error más pequeño
  - Una forma fácil de reducir el SSE es aumentar K, el número de grupos
    - Un buen agrupamiento con un menor K puede tener un menor SSE que un pobre agrupamiento con un mayor K

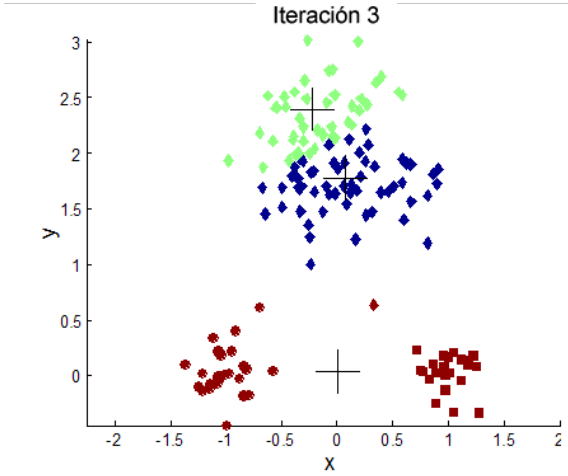
Diplomado BIG DATA ANALITYCS
K-Means

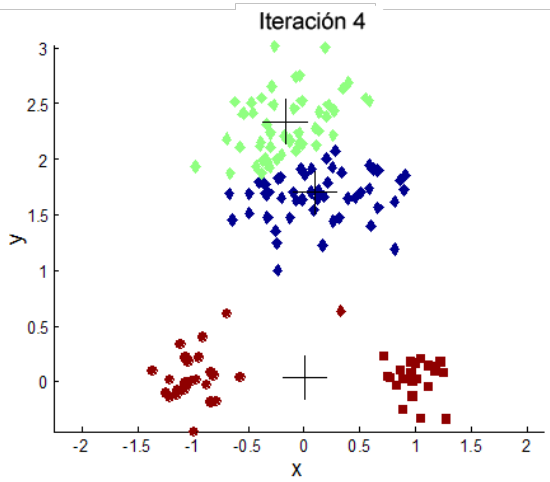
K-Means

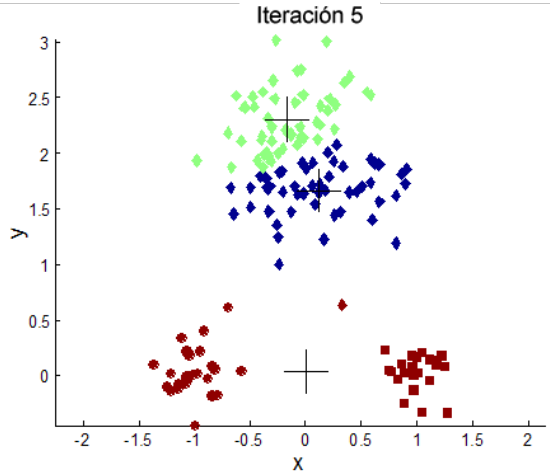
- Importancia de escoger bien los centroides









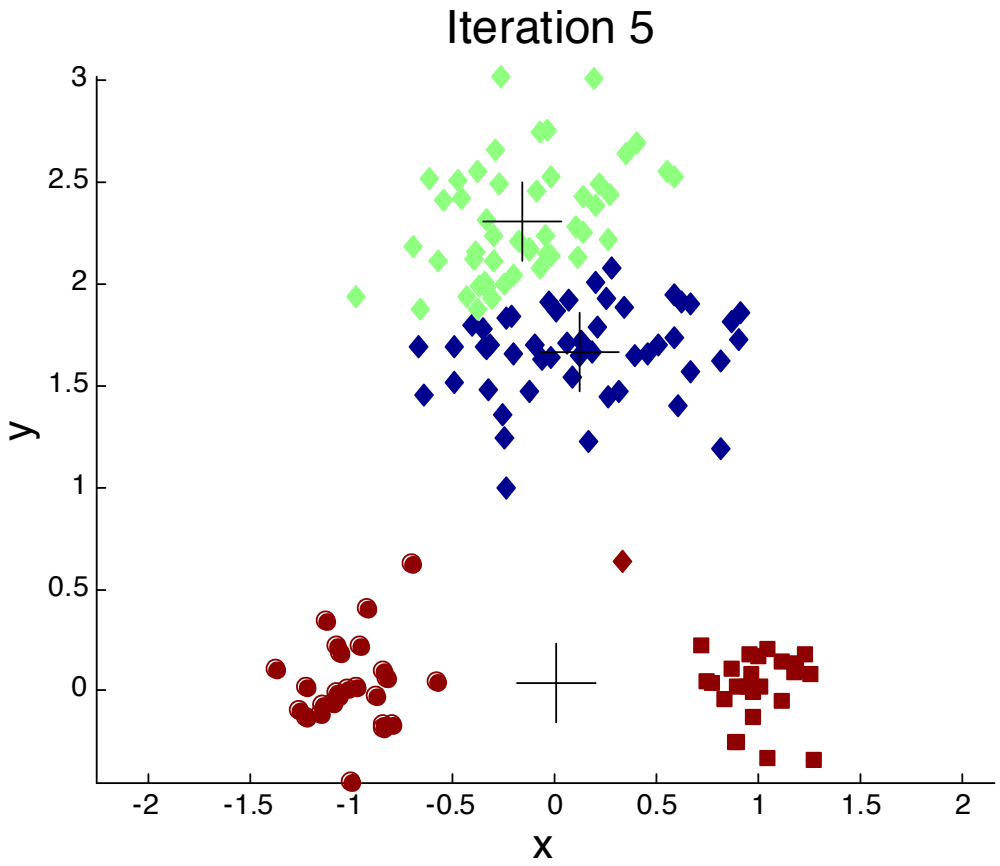


Diplomado BIG DATA ANALITYCS

K-Means

K-Means

- Importancia de escoger bien los centroides



K-Means. Problema con los Centroides Iniciales

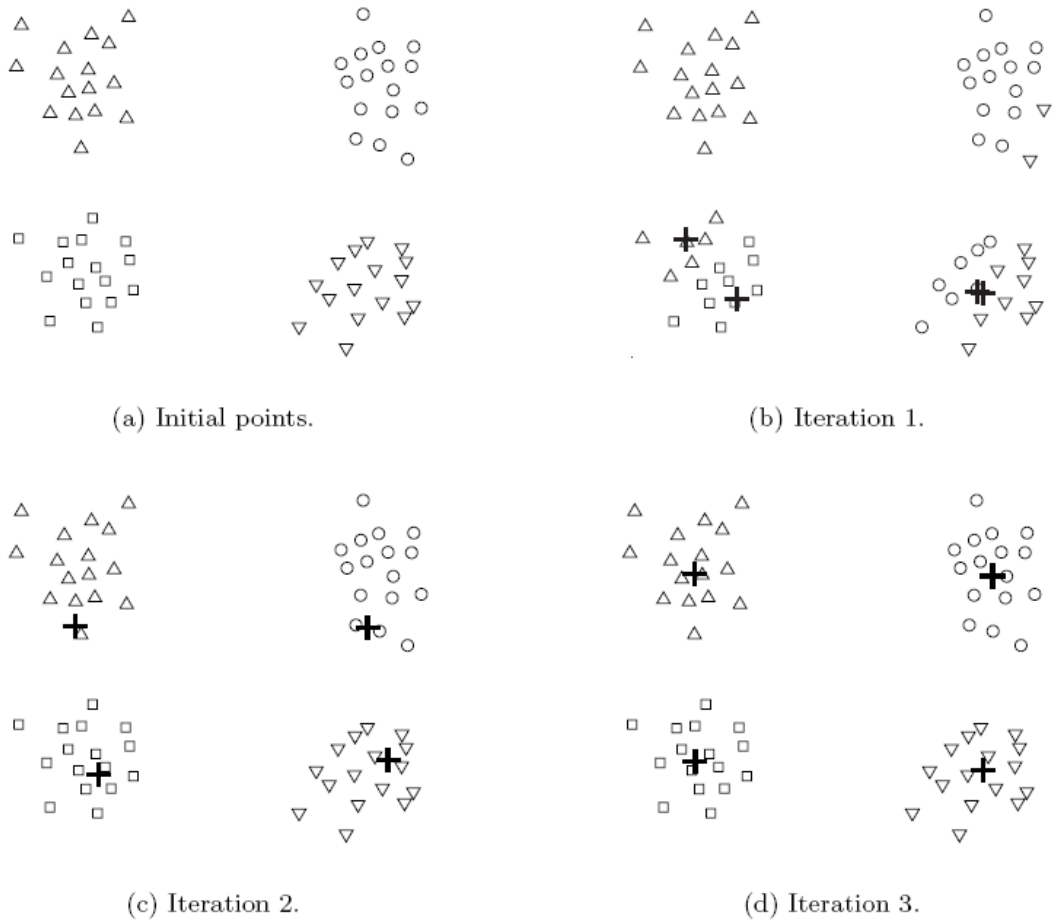
UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Si hay K grupos 'reales' entonces la oportunidad de seleccionar un centroide de cada grupo es pequeña.
  - La oportunidad es relativamente pequeña cuando K es grande
  - Si los grupos son del mismo tamaño, n, entonces.

$$p = \frac{\#\text{ de posibilidades de seleccionar un centroide de cada grupo}}{\#\text{ de posibilidades de seleccionar } K \text{ centroides}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

  - Por ejemplo, si K = 10, entonces la probabilidad $10!/10^{10} = 0.00036$.
  - A veces, los centroides iniciales se reajustan en la posición 'correcta', y a veces no.
  - Considere un ejemplo de cinco pares de grupos.

K-Means. Problema con los Centroides Iniciales



(a) Initial points.

(b) Iteration 1.

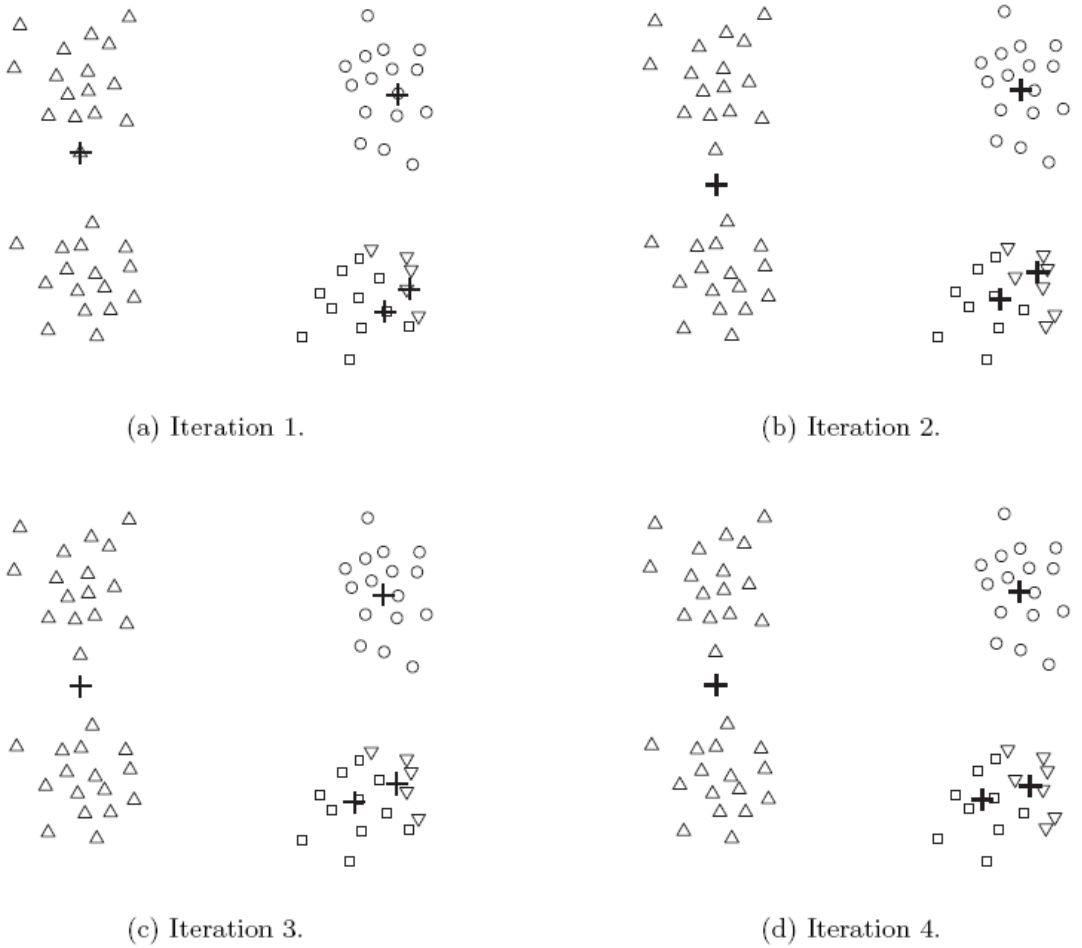
(c) Iteration 2.

(d) Iteration 3.

Figure 8.6. Two pairs of clusters with a pair of initial centroids within each pair of clusters.

Diplomado
K-Means

K-Means. Problema con los Centroides Iniciales



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Figure 8.7. Two pairs of clusters with more or fewer than two initial centroids within a pair of clusters.

Diplomado
K-Means

K-Means. Solución al Problema de los centroides

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Múltiples ejecuciones
  - Ayuda, pero la probabilidad no está de su lado
- Muestreo y uso de agrupamiento jerárquico para determinar centroides iniciales
- Seleccionar más de k centroides iniciales y después, seleccionar entre éstos
  - Seleccionar los más ampliamente separados
  - Postprocesamiento
- Bisecting K-means
  - No es tan susceptible a los problemas de inicialización

Medidas de Validación

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Medidas numéricas que se aplican para evaluar diversos aspectos de la validez de un grupo, se clasifican en los siguientes tres tipos.
  - Índice externo: Se utiliza para medir el grado en que las etiquetas de un grupo coinciden con las etiquetas externas.
    - Entropía
  - Índice interno: Se utiliza para medir qué tan buena es la estructura del agrupamiento sin información externa.
    - Suma de errores cuadrados (SSE)
  - Índice relativo: se utiliza para comparar dos grupos o agrupamientos diferentes.
    - A menudo, un índice externo o interno se utiliza para esta función, e.g., la entropía o SSE

- A veces, estas se conocen como criterios en lugar de índices
  - Sin embargo, a veces el criterio es la estrategia general y el índice es la medida numérica que se aplica en el criterio.

Diplomado BIG DATA ANALITYCS
K-Means

Referencias

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2005, Introduction to Data Mining, Addison-Wesley.

¿Preguntas?

jecamargom@unal.edu.co

http://www.mindlaboratory.org