Clustering basado en prototipos

Christopher Expósito Izquierdo
Airam Expósito Márquez
Israel López Plata
Belén Melián Batista
J. Marcos Moreno Vega

{**cexposit, aexposim, ilopezpl, mbmelian, jmmoreno**}**@ull.edu.es**

Departamento de Ingeniería Informática y de Sistemas
Universidad de La Laguna

Índice

CLUSTERING BASADO EN PROTOTIPOS

Algoritmo K-means

- K-means [1] es un algoritmo simple con el que obtener clusters basados en prototipos (centroides).
- Consta de los siguientes pasos:
  - Sean dados K centroides iniciales.
  - Asignar cada punto al centroide más cercano. Los puntos asignados a un centroide forman un cluster.
  - Recalcular el centroide de cada cluster.
  - Repetir los anteriores dos pasos hasta que los puntos no cambien de cluster.

Algoritmo K-means. Ejemplo

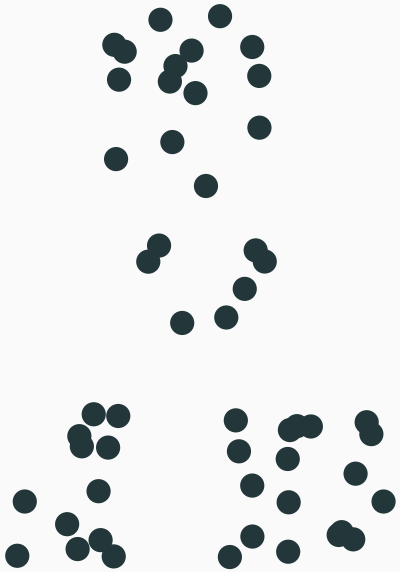

Figura 1: Iteración 1.

Algoritmo K-means. Ejemplo

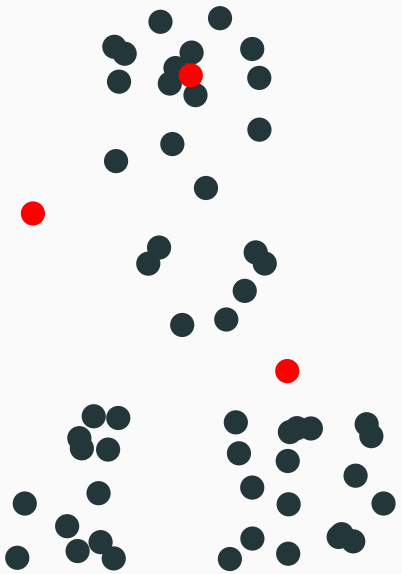

Figura 1: Iteración 1.

Algoritmo K-means. Ejemplo

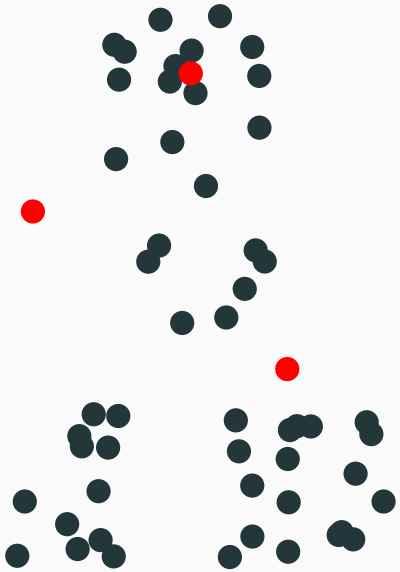

Figura 1: Iteración 1.

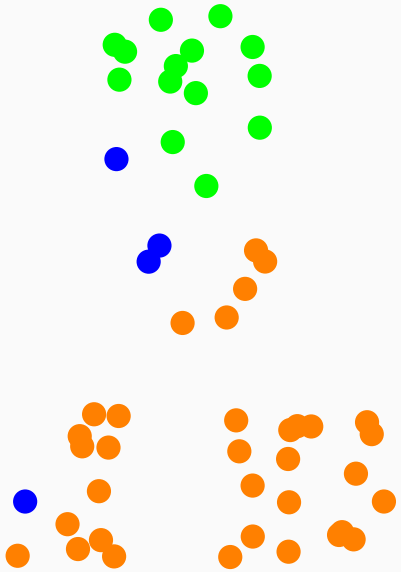

Figura 2: Iteración 2.

Algoritmo K-means. Ejemplo

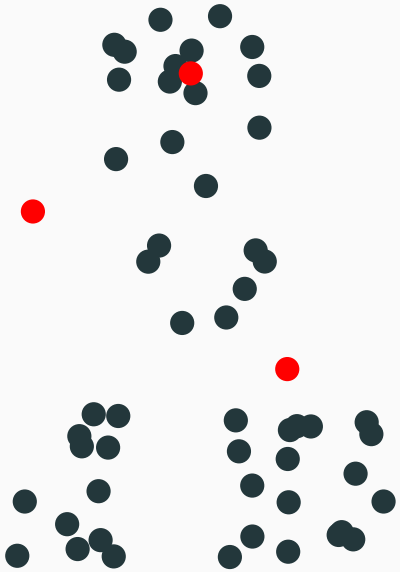

Figura 1: Iteración 1.

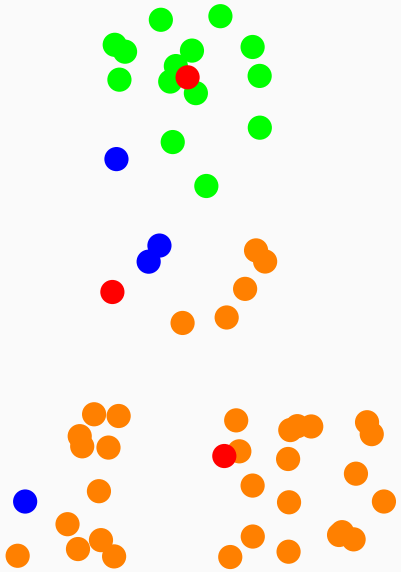

Figura 2: Iteración 2.

Algoritmo K-means. Ejemplo

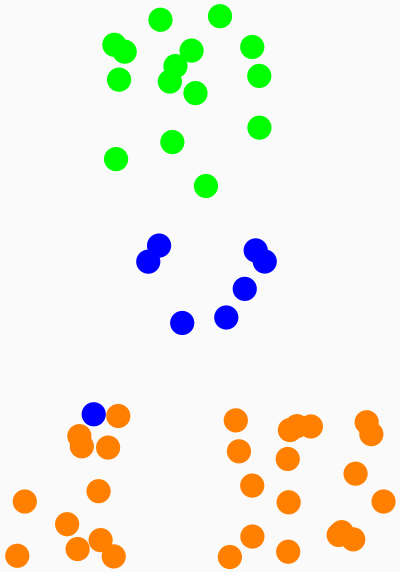

Figura 3: Iteración 3.

Algoritmo K-means. Ejemplo

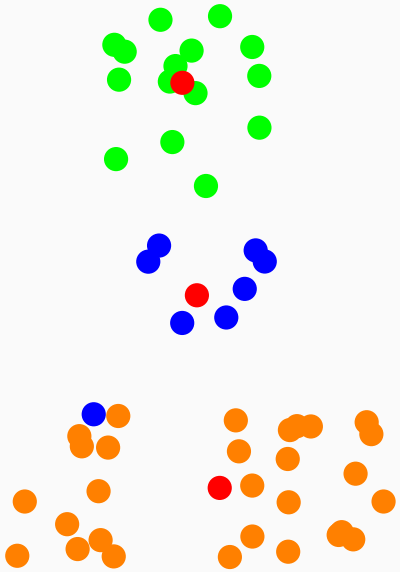

Figura 3: Iteración 3.

Algoritmo K-means. Ejemplo

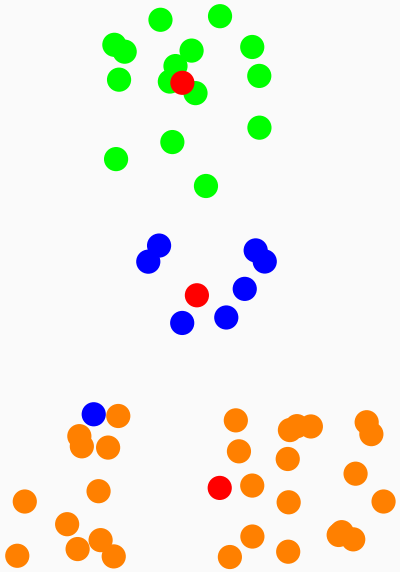

Figura 3: Iteración 3.

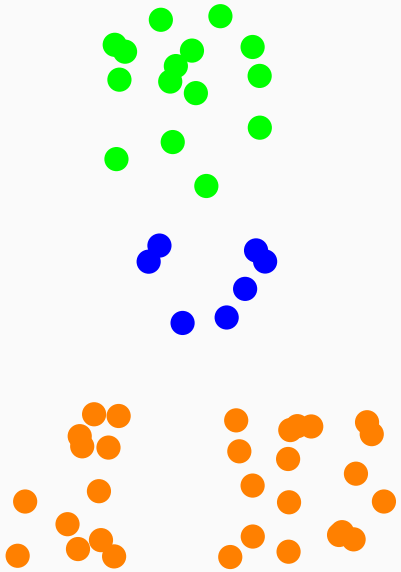

Figura 4: Iteración 4.

Algoritmo K-means. Ejemplo

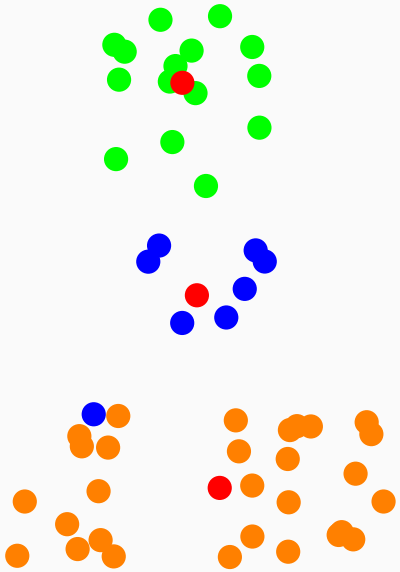

Figura 3: Iteración 3.

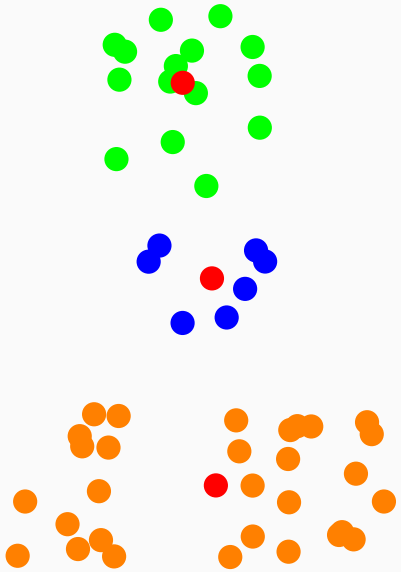

Figura 4: Iteración 4.

Algoritmo K-means. Observaciones

- En gran medida, el comportamiento del algoritmo K-means depende de los centroides iniciales elegidos.
- En general, suministra un agrupamiento que es localmente óptimo.
- La ejecución del algoritmo desde centroides iniciales distintos suministra agrupamientos distintos.

Algoritmo K-means. Observaciones

- En gran medida, el comportamiento del algoritmo K-means depende de los centroides iniciales elegidos.
- En general, suministra un agrupamiento que es localmente óptimo.
- La ejecución del algoritmo desde centroides iniciales distintos suministra agrupamientos distintos.

Algoritmo K-means. Observaciones

- En gran medida, el comportamiento del algoritmo K-means depende de los centroides iniciales elegidos.
- En general, suministra un agrupamiento que es localmente óptimo.
- La ejecución del algoritmo desde centroides iniciales distintos suministra agrupamientos distintos.

Calidad del clustering (i)

- Dados dos agrupamientos, ¿cuál es mejor?
- Si la distancia empleada es la euclídea, suele usarse como medida de calidad del agrupamiento la suma de errores al cuadrado (SSE):

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

donde

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

y m_i es el número de puntos que pertenecen al cluster C_i.

Calidad del clustering (i)

- Dados dos agrupamientos, ¿cuál es mejor?
- Si la distancia empleada es la euclídea, suele usarse como medida de calidad del agrupamiento la suma de errores al cuadrado (SSE):

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

donde

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

y m_i es el número de puntos que pertenecen al cluster C_i.

Calidad del clustering (ii)

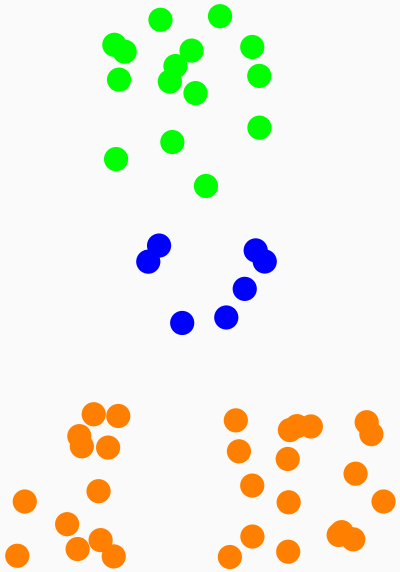

Figura 5: SSE = 23,772

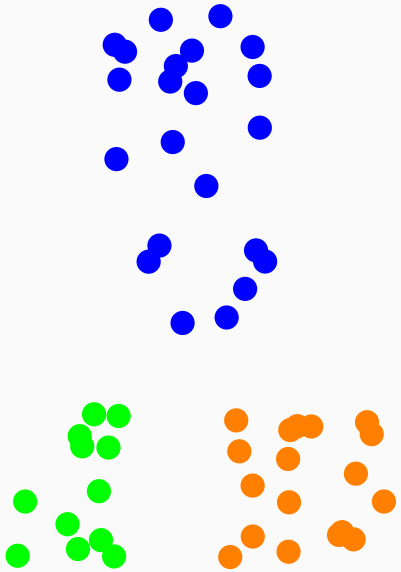

Figura 6: SSE = 16,761

Algoritmo K-means. Centroides iniciales

- Es habitual que los centroides iniciales se generen aleatoriamente.
- Este procedimiento puede suministrar agrupamientos de baja calidad.
- Para incrementar la calidad del agrupamiento final, se puede ejecutar varias veces el algoritmo K-means desde diferentes centroides iniciales. El agrupamiento con menor suma de errores al cuadrado, *SSE*, será el elegido.
- Otra alternativa consiste en seleccionar, iterativamente, como nuevo centroide al punto más alejado de los centroides actuales. Para ello, se parte de un centroide seleccionado al azar y se ejecuta el paso anterior hasta que se obtienen K centroides.

Algoritmo K-means. Centroides iniciales

- Es habitual que los centroides iniciales se generen aleatoriamente.
- Este procedimiento puede suministrar agrupamientos de baja calidad.
- Para incrementar la calidad del agrupamiento final, se puede ejecutar varias veces el algoritmo K-means desde diferentes centroides iniciales. El agrupamiento con menor suma de errores al cuadrado, *SSE*, será el elegido.
- Otra alternativa consiste en seleccionar, iterativamente, como nuevo centroide al punto más alejado de los centroides actuales. Para ello, se parte de un centroide seleccionado al azar y se ejecuta el paso anterior hasta que se obtienen K centroides.

Algoritmo K-means. Centroides iniciales

- Es habitual que los centroides iniciales se generen aleatoriamente.
- Este procedimiento puede suministrar agrupamientos de baja calidad.
- Para incrementar la calidad del agrupamiento final, se puede ejecutar varias veces el algoritmo K-means desde diferentes centroides iniciales. El agrupamiento con menor suma de errores al cuadrado, SSE, será el elegido.
- Otra alternativa consiste en seleccionar, iterativamente, como nuevo centroide al punto más alejado de los centroides actuales. Para ello, se parte de un centroide seleccionado al azar y se ejecuta el paso anterior hasta que se obtienen K centroides.

Algoritmo K-means. Centroides iniciales

- Es habitual que los centroides iniciales se generen aleatoriamente.
- Este procedimiento puede suministrar agrupamientos de baja calidad.
- Para incrementar la calidad del agrupamiento final, se puede ejecutar varias veces el algoritmo K-means desde diferentes centroides iniciales. El agrupamiento con menor suma de errores al cuadrado, SSE, será el elegido.
- Otra alternativa consiste en seleccionar, iterativamente, como nuevo centroide al punto más alejado de los centroides actuales. Para ello, se parte de un centroide seleccionado al azar y se ejecuta el paso anterior hasta que se obtienen K centroides.

Algoritmo K-means. Clusters vacíos

- En ocasiones, el algoritmo K-means suministra clusters vacíos.
- Cuando esto ocurre, debe implementarse algún procedimiento para escoger un nuevo centroide que reemplace al del cluster vacío.
- Algunas estrategias son:
  - escoger como nuevo centroide al punto más alejado del resto de centroides;
  - escoger como nuevo centroide un punto del cluster con mayor suma de errores al cuadrado.

Algoritmo K-means. Clusters vacíos

- En ocasiones, el algoritmo K-means suministra clusters vacíos.
- Cuando esto ocurre, debe implementarse algún procedimiento para escoger un nuevo centroide que reemplace al del cluster vacío.
- Algunas estrategias son:
  - escoger como nuevo centroide al punto más alejado del resto de centroides;
  - escoger como nuevo centroide un punto del cluster con mayor suma de errores al cuadrado.

Algoritmo K-means. Clusters vacíos

- En ocasiones, el algoritmo K-means suministra clusters vacíos.
- Cuando esto ocurre, debe implementarse algún procedimiento para escoger un nuevo centroide que reemplace al del cluster vacío.
- Algunas estrategias son:
  - escoger como nuevo centroide al punto más alejado del resto de centroides;
  - escoger como nuevo centroide un punto del cluster con mayor suma de errores al cuadrado.

Algoritmo K-means. Outliers (i)

- Los *outliers* pueden tener una influencia indebida en la calidad de los clusters encontrados.
- Al aplicar el algoritmo K-means, los *outliers* suelen formar clusters con pocos elementos que no son representativos del conjunto de puntos.
- Para evitar esta influencia no deseada, los *outliers* suelen ser descartados antes de ejecutar el algoritmo.
- Se utilizan para ello técnicas de detección de *outliers*.

Algoritmo K-means. Outliers (i)

- Los *outliers* pueden tener una influencia indebida en la calidad de los clusters encontrados.
- Al aplicar el algoritmo K-means, los *outliers* suelen formar clusters con pocos elementos que no son representativos del conjunto de puntos.
- Para evitar esta influencia no deseada, los *outliers* suelen ser descartados antes de ejecutar el algoritmo.
- Se utilizan para ello técnicas de detección de *outliers*.

Algoritmo K-means. Outliers (i)

- Los *outliers* pueden tener una influencia indebida en la calidad de los clusters encontrados.
- Al aplicar el algoritmo K-means, los *outliers* suelen formar clusters con pocos elementos que no son representativos del conjunto de puntos.
- Para evitar esta influencia no deseada, los *outliers* suelen ser descartados antes de ejecutar el algoritmo.
- Se utilizan para ello técnicas de detección de *outliers*.

Algoritmo K-means. Outliers (i)

- Los *outliers* pueden tener una influencia indebida en la calidad de los clusters encontrados.
- Al aplicar el algoritmo K-means, los *outliers* suelen formar clusters con pocos elementos que no son representativos del conjunto de puntos.
- Para evitar esta influencia no deseada, los *outliers* suelen ser descartados antes de ejecutar el algoritmo.
- Se utilizan para ello técnicas de detección de *outliers*.

Algoritmo K-means. Outliers (ii)

- Los *outliers* también pueden ser identificados y eliminados en una fase de postprocesamiento.
- Para su identificación se emplean características que suelen poseer los *outliers* como:
  - ser puntos con alta contribución a la suma de errores al cuadrado;
  - pertenecer a clusters con pocos puntos.

Algoritmo K-means. Outliers (ii)

- Los *outliers* también pueden ser identificados y eliminados en una fase de postprocesamiento.
- Para su identificación se emplean características que suelen poseer los *outliers* como:
  - ser puntos con alta contribución a la suma de errores al cuadrado;
  - pertenecer a clusters con pocos puntos.

Algoritmo K-means. Outliers (ii)

- Los *outliers* también pueden ser identificados y eliminados en una fase de postprocesamiento.
- Para su identificación se emplean características que suelen poseer los *outliers* como:
  - ser puntos con alta contribución a la suma de errores al cuadrado;
  - pertenecer a clusters con pocos puntos.

Algoritmo K-means. Cómo reducir la suma de errores al cuadrado

- **Dividir un cluster.** Normalmente, se divide el cluster con mayor error cuadrático medio.
- **Crear un nuevo cluster.** Para crear el nuevo cluster se suele escoger como centroide el punto más alejado de los centroides actuales.
- **Eliminar un cluster.** Tras eliminar el centroide del cluster, los puntos asignados a él son reasignados a otros centroides. En general, se elimina el cluster con mayor error cuadrático medio o el que produce el menor incremento en el error cuadrático medio global.
- **Unir dos clusters.** Pueden unirse los clusters cuyos centroides están más próximos o aquellos con cuya unión se produce el menor incremento en el error cuadrático medio global.

Algoritmo K-means. Cómo reducir la suma de errores al cuadrado

- **Dividir un cluster.** Normalmente, se divide el cluster con mayor error cuadrático medio.
- **Crear un nuevo cluster.** Para crear el nuevo cluster se suele escoger como centroide el punto más alejado de los centroides actuales.
- **Eliminar un cluster.** Tras eliminar el centroide del cluster, los puntos asignados a él son reasignados a otros centroides. En general, se elimina el cluster con mayor error cuadrático medio o el que produce el menor incremento en el error cuadrático medio global.
- **Unir dos clusters.** Pueden unirse los clusters cuyos centroides están más próximos o aquellos con cuya unión se produce el menor incremento en el error cuadrático medio global.

Algoritmo K-means. Cómo reducir la suma de errores al cuadrado

- **Dividir un cluster.** Normalmente, se divide el cluster con mayor error cuadrático medio.
- **Crear un nuevo cluster.** Para crear el nuevo cluster se suele escoger como centroide el punto más alejado de los centroides actuales.
- **Eliminar un cluster.** Tras eliminar el centroide del cluster, los puntos asignados a él son reasignados a otros centroides. En general, se elimina el cluster con mayor error cuadrático medio o el que produce el menor incremento en el error cuadrático medio global.
- **Unir dos clusters.** Pueden unirse los clusters cuyos centroides están más próximos o aquellos con cuya unión se produce el menor incremento en el error cuadrático medio global.

Algoritmo K-means. Cómo reducir la suma de errores al cuadrado

- **Dividir un cluster.** Normalmente, se divide el cluster con mayor error cuadrático medio.
- **Crear un nuevo cluster.** Para crear el nuevo cluster se suele escoger como centroide el punto más alejado de los centroides actuales.
- **Eliminar un cluster.** Tras eliminar el centroide del cluster, los puntos asignados a él son reasignados a otros centroides. En general, se elimina el cluster con mayor error cuadrático medio o el que produce el menor incremento en el error cuadrático medio global.
- **Unir dos clusters.** Pueden unirse los clusters cuyos centroides están más próximos o aquellos con cuya unión se produce el menor incremento en el error cuadrático medio global.

BIBLIOGRAFÍA

Bibliografía

MacQueen, J.

Some methods for classification and analysis of multivariate observations.

In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley, 1967), University of California Press, pp. 281–297.

Licencia

Esta obra está bajo una licencia de Creative Commons.
Reconocimiento - No comercial - Compartir igual