UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

K-medias, y amigos

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 20) 04.abril.2022

K-medias

Suponga que tenemos un conjunto de datos $\mathbb{X} = \{\mathbf{x}_i\}_{i=1}^n$, con $\mathbf{x}_i \in \mathbb{R}^d$, y se quiere agrupar estos elementos en k conjuntos o categorías distintas.

- Definimos distancias entre observaciones $d(\mathbf{x}_i, \mathbf{x}_j)$ (*e.g.* distancia euclideana).
- Por otro lado, elegimos un representante para cada uno de los k grupos o categorías a construir (*e.g.* el centroide $\mathbf{c}_k \in \mathbb{R}^d$ del grupo).

El método de **k-medias** (ó *k-means*) elige al azar k representantes, y asigna cada dato $\mathbf{x}_i$ al representante más cercano, según la distancia $d(\cdot, \cdot)$.

K-medias

Algoritmo: (K-medias)

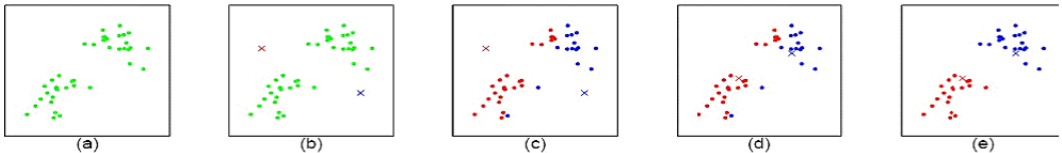
1. Para cada $g = 1, 2, 3, \ldots, k$, elegimos de manera aleatoria el centroide $\mathbf{c}_g \in \mathbb{R}^d$ del grupo g.
2. Repetir hasta convergencia:
   - Asignar cada $\mathbf{x}_i$ al representante más cercano según la métrica d, esto es:

$$h(\mathbf{x}_i) = g(i) \in C, \quad \text{con } g(i) = \text{argmin}_{1 \leq g \leq k} \; d(\mathbf{c}_g, \mathbf{x}_i).$$

   - Recalcular los representantes como los centroides de los datos asociados a cada grupo

$$\mathbf{c}_g = \frac{1}{|\{i : g(i) = g\}|} \sum_{i:g(i)=g} \mathbf{x}_i, \quad g = 1, 2, \ldots, k.$$

K-medias



(a) Conjunto de datos $\mathbb{X}$.
(b) Se eligen aleatoriamente los centroides $\mathbf{c}_g$, $k = 2$.
(c) Cada $\mathbf{x}_i$ se asocia con su centroide más cercano.
(d) Se recalculan los centroides $\mathbf{c}_g$, $k = 2$.
(e) Cada $\mathbf{x}_i$ se etiqueta con su centroide más cercano.
(f) Repetir (b) a (e) hasta convergencia ...

K-medias

K-medias minimiza una función de costo de una manera particular:
Sea $\mathbf{c}_g$ el representante del grupo g, $g(i)$ el grupo de $\mathbf{x}_i$, definimos

$$J(\mathbb{X}) = \min_{g} \min_{\mathbf{c}_g} \sum_{g} \sum_{i:g(i)=g} ||\mathbf{x}_i - \mathbf{c}_g||^2 \tag{1}$$

En k-medias J se minimiza en dos pasos desacoplados:

- Fijando g, y minimizando sobre $\mathbf{c}_g$.
- Fijando $\mathbf{c}_g$, y minimizando sobre g.

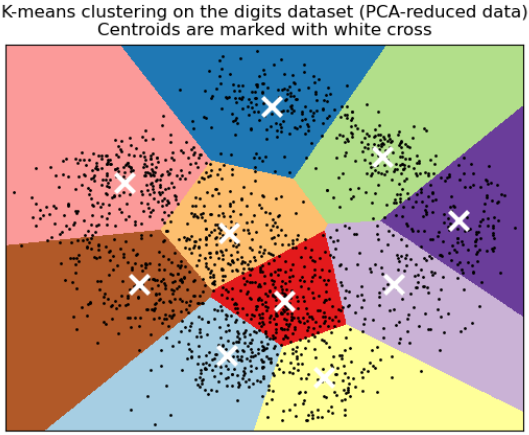
Interpretación:

$$||\mathbf{x}_i - \mathbf{c}_g||^2 = ||\mathbf{x}_i - decoder(encoder(\mathbf{x}_i))||^2. \tag{2}$$

Se trata de minimizar la varianza dentro de cada grupo. Entre mayor es k menor la varianza (2).

K-medias

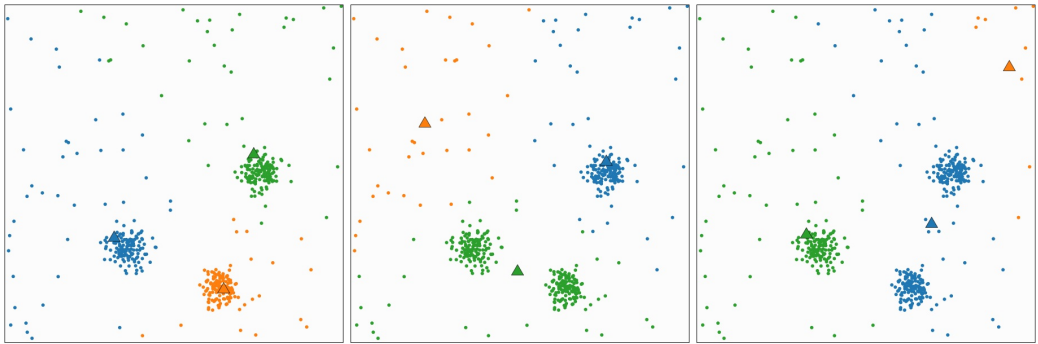
Al final, el algoritmo de k-means induce una partición de Voronoi sobre el espacio $\mathbb{R}^d$:



K-medias

https://stanford.edu/class/engr108/visualizations/kmeans/kmeans.html

Como la elección inicial de los $\mathbf{c}_g$ es aleatoria, es posible obtener diferentes resultados:



Distintos resultados con *k*-means en el mismo conjunto de datos

K-medias

Observaciones:

- Se sugiere replicar k-means varias veces (con diferentes inicializaciones). Esto permite que aparezcan diferentes estructuras de agrupamiento.

- El diagrama de Voronoi es útil, ya pensando en k-means como un método predictivo.

Existen diversas variantes:

- k-medianas.
- k-medioides.
- Fuzzy k-means.

K-medianas

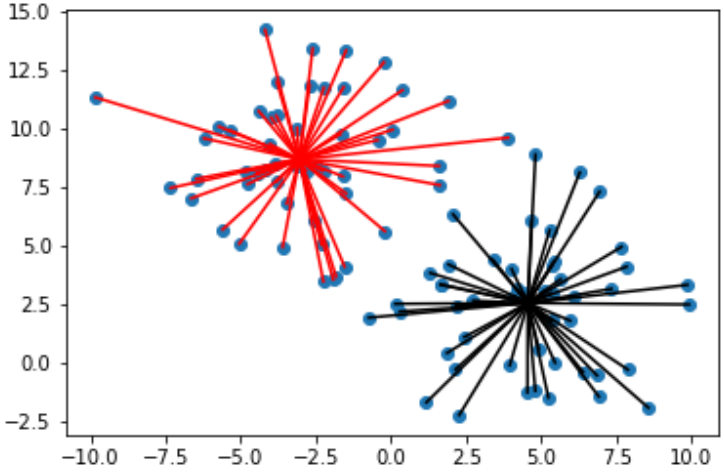
Es una variante de k-medias. En lugar de utilizar el centroide $\mathbf{c}_g$ como representante de cada grupo g, se utiliza la mediana.

Esta mediana se calcula, componente a componente, usando la distancia Manhattan (norma $||\cdot||_1$). De este modo, cada uno de los atributos o variables es una observación dentro del conjunto de datos.

Algoritmo: (K-medianas)

1. Para cada $j = 1, 2, 3, \ldots, k$, elegimos de manera aleatoria el representante $\mathbf{c}_j \in \mathbb{R}^d$.
2. Repetir hasta convergencia:
   - Asignar cada $\mathbf{x}_i$ al representante más cercano según la métrica d.
   - Recalcular los representantes como las medianas, componente a componente, de los datos asociados a cada grupo.

K-medianas



K-medioides

Es otra variante de k-medias. En lugar de utilizar el centroide $\mathbf{c}_g$ como representante de cada grupo g, se utiliza el dato más cercano al centroide $\mathbf{c}_g$:

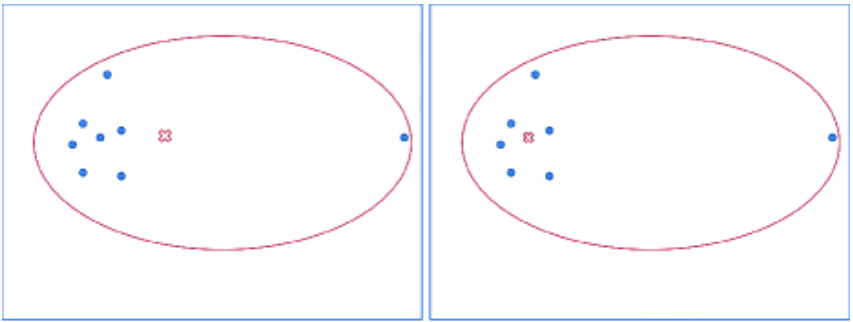
$$\mathbf{c}_g = \mathbf{x}_j, \quad \text{con } j = \text{argmin}_\ell \left|\left| \mathbf{x}_\ell - \frac{1}{|g(i)|} \sum_{i:g(i)=g} \mathbf{x}_i \right|\right|^2$$

De este modo, cada uno de los representante se elige dentro del mismo conjunto de datos.

Algoritmo: (K-medioides)

1. Para cada $j = 1, 2, 3, \ldots, k$, elegimos de manera aleatoria el representante $\mathbf{c}_j \in \mathbb{R}^d$.
2. Repetir hasta convergencia:
   - Asignar cada $\mathbf{x}_i$ al representante más cercano según la métrica d.
   - Recalcular los representantes como el dato más cercano a $\mathbf{c}_g$.

K-medioides



Fuzzy K-medias

Agrupamiento difuso:

El agrupamiento difuso (*fuzzy clustering*) es una clase de algoritmos de agrupamiento donde, en lugar de asignar un único grupo a cada dato $\mathbf{x}_i$, cada elemento tiene un grado de pertenencia (difuso) a cada uno de los grupos.

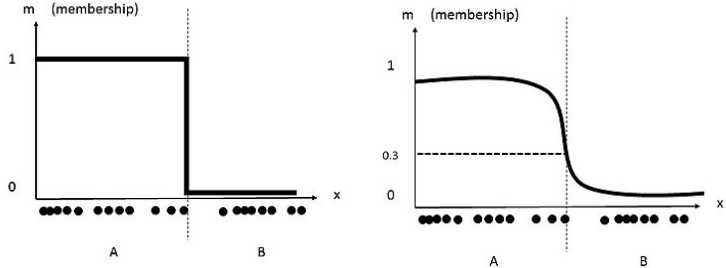
- surge de la necesidad de resolver una deficiencia del agrupamiento exclusivo (agrupación inequívoca).
- implementaciónes a partir del surgimiento de la lógica difusa (Zadeh, 1965).
- Se representa la similitud entre un elemento $\mathbf{x}_i$ y un grupo g por una función, llamada función de pertenencia $\mathbf{w}_i : \mathbf{x}_i \rightarrow [0, 1]^k$, que toma valores entre cero y uno.

Fuzzy K-medias

Básicamente, a cada dato $\mathbf{x}_i$, el calsificador difuso asigna un vector de coeficientes

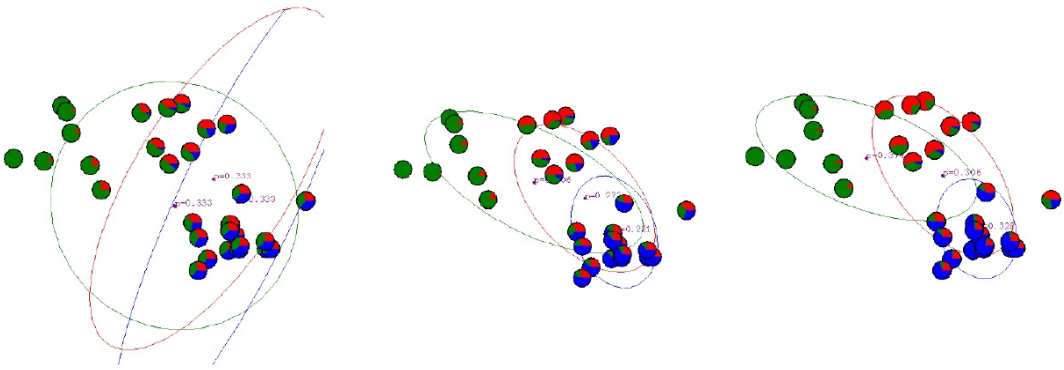
$$h(\mathbf{x}_i) = \mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ik}),$$

con $w_{ij} = \mathbb{P}(\mathbf{x}_i \in g_j)$.



Diferencias entre (a) *hard-clustering*, y (b) *fuzzy-clustering*.

Fuzzy K-medias



Segmentación o classificación obtenida con *fuzzy-clustering*.

Fuzzy K-medias

Para cada dato $\mathbf{x}_i \in \mathbb{R}^d$, consideramos un vector de coeficientes $\mathbf{w}_i = (w_{i1}, \ldots, \mathbf{w}_{ik}) \in \mathbb{R}^k$, con $w_{ij} \geq 0$, $\forall\, j = 1, 2, \ldots, k$.

Con este esquema, el centroide de un grupo g se calcula como el promedio ponderado de sus elementos:

$$\mathbf{c}_j = \frac{\sum_{i:g(i)=j} w_{ij}^m \mathbf{x}_i}{\sum_{i:g(i)=j} w_{ij}^m},$$

$m > 0$ es un hiperparámetro que controla el suavizamiento. A mayor m, mayor difusividad.
Los pesos w_{ij} se recalculan como

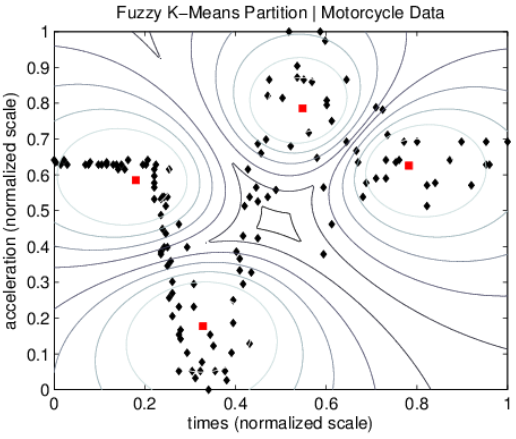
$$w_{ij} = \frac{1}{\sum_{\ell=1}^{k} \left(\frac{||\mathbf{x}_i - \mathbf{c}_j||}{||\mathbf{x}_i - \mathbf{c}_\ell||} \right)^{\frac{2}{m-1}}}.$$

Fuzzy K-medias

Algoritmo: (Fuzzy k-medias) Dado un conjunto de datos $\mathbb{X} = \{\mathbf{x}_i\}$, y un conjunto de clases $C = \{g_1, g_2, \ldots, g_k\}$, el algoritmo construye una matriz $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n \times k}$ con los grados de pertenencia.

1. Para cada $j = 1, 2, 3, \ldots, k$, elegimos de manera aleatoria el centroide $\mathbf{c}_j \in \mathbb{R}^d$. Elegimos pesos aleatorios $\mathbf{w}_i$ (e.g. $\mathbf{w}_i = (1, 1, \ldots, 1)$, $\forall i$.
2. Repetir hasta convergencia:
   - Asignar cada $\mathbf{x}_i$ al representante más cercano según la métrica d.
   - Recalcular los representantes como los centroides ponderados de los datos asociados a cada grupo.
   - Recalcular los grados de pertenencia w_{ij}.

Fuzzy K-medias



Curvas de nivel de la partición difusa.

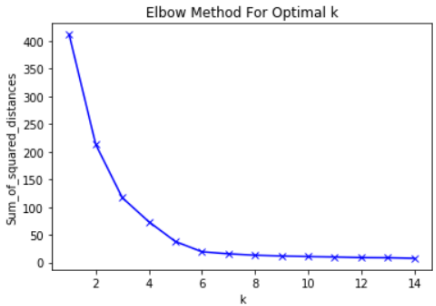
Aplicaciones

Segmentación de imágenes:
https://www.youtube.com/watch?v=yR7k19YBqiwab_channel=Computerphile

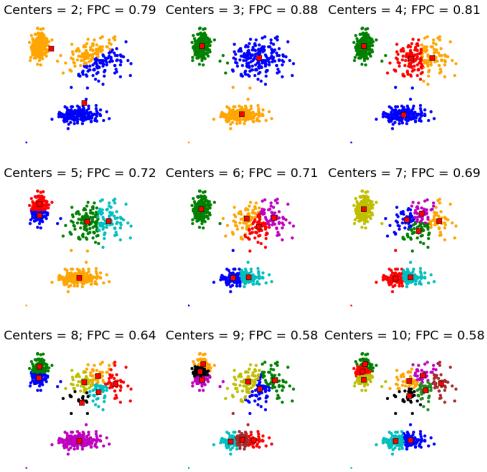
Observaciones

Observaciones:

- Como la distancia euclideana da igual peso a cada dimensión, mejor normalizar los datos (normalizar o estandarizar).
- Existen muchas heuristicas para elegir *k*, *e.g.* el método del “codo”: cómo cambia la suma de variazas por grupo (1) en función de *k*.



Observaciones

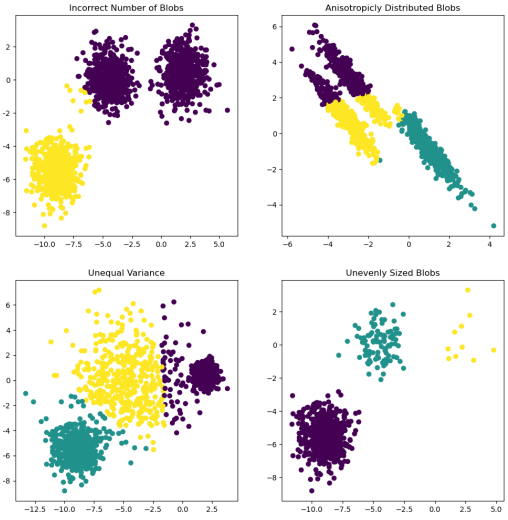


Particiones obtenidas con diferentes valores de k.

Observaciones

- En muchas situaciones, k-medias falla.

  Solución: transformar los datos o usar métodos donde $d(\cdot, \cdot)$ captura la forma local de los datos.



Observaciones

- Muchas veces se usa primero algun método de reducción de dimensión, cuando la dimensionalidad del espacio es muy alta.

  Las distancias (euclideanas) pierden poder discriminativo en dimensiones altas. Se llama la *maldicion de la alta dimensionalidad*.
- Si los datos no son muy continuos, se prefiere tomar como representantes observaciones de la muestra (k-medianas, k-medioides).
- Miles de variantes! (también porque para conjuntos de datos grandes, el algoritmo básico es demasiado costoso).
- Es buena idea correr el algoritmo con diferentes puntos de arranque para evitar óptimos locales.