uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
REDES DE NEURONAS ARTIFICIALES
Inés M. Galván – José M. Valls

Algoritmo de K-medias

K-medias

- Propuesto por J. MacQueen, 1967
- Algoritmo de agrupación no supervisado mediante el cual el espacio de patrones de entrada se divide en K clusters o regiones
- Es necesario establecer el número de clusters (K)
- Dado un conjunto de patrones $\{X(n)=(x_1(n), x_2(n),.... x_p(n))\}_{n=1,...N}$, se pretende encontrar K centros $C_i=(c_{i1}, c_{12}, ...c_{ip})$ $i=1,\ldots,K$ con el objetivo de minimizar las distancias euclídeas entre los patrones de entrada y el centro más cercano

$$J = \sum_{i=1}^{K} \sum_{n=1}^{N} M_{in} \|X(n) - C_i\|$$

donde *N* es el número de patrones, || . || es la distancia euclídea, *X(n)* es el patrón de entrada *n* y M_{in} es la función de pertenencia, que vale 1 si el centro *Ci* es el más cercano al patrón *X(n)*, y 0 en otro caso, es decir:

$$M_{in} = \begin{cases} 1 & \text{si } \|X(n) - C_i\| < \|X(n) - C_s\| \ \forall s \neq i, \ s = 1, 2, ..., K \\ 0 & \text{en otro caso} \end{cases}$$

K-medias

Dado el número de clases K, el conjunto de patrones de entrada, los pasos son:

1. Se inicializan aleatoriamente los centros de los K clusters (centroides)
2. Se asignan N_i patrones de entrada a cada cluster i del siguiente modo
   - El patrón X(n) pertenece al cluster i si

$$\|X(n) - C_i\| < \|X(n) - C_s\|$$

$$\forall\ s \neq i \text{ con } s = 1, 2, ..., K.$$

   - Por tanto, cada cluster tendrá asociado un determinado número de patrones de entrada, aquellos más cercanos a su centroide (cada centroide determina una región de Voronoi (http://www.pi6.fernuni-hagen.de/GeomLab/VoroGlide/index.html.en)
3. Se calcula la nueva posición de los centroides como la media de todos los patrones que pertenecen al cluster, es decir:

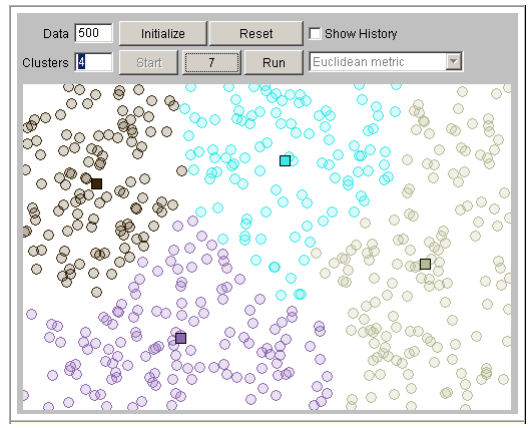
$$c_{ij} = \frac{1}{N_i} \sum_{n=1}^{N} M_{in} x_j(n) \text{ para } j = 1, 2, ...p\ ,\ i = 1, 2, ..., K$$

4. Se repiten los pasos 2 y 3 hasta que las nuevas posiciones de los centroides no se modifiquen respecto a su posición anterior, es decir hasta que:

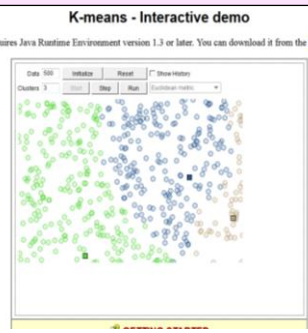
$$\|C_i^{nuevo} - C_i^{anterior}\| < \varepsilon\ \forall\ i = 1, 2, ..., K$$

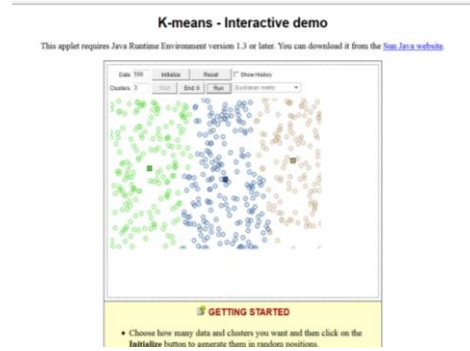
K-medias

- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

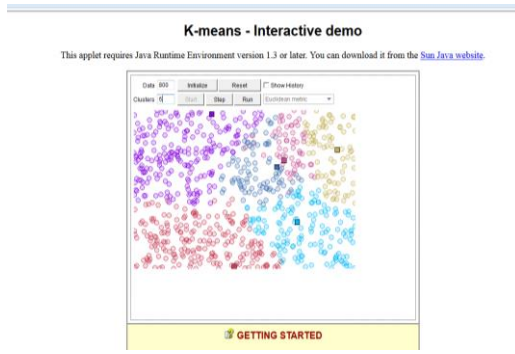


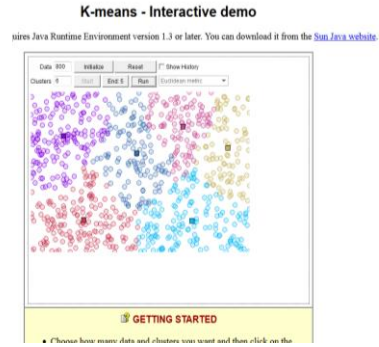
K-medias





K-medias





K-medias

- El algoritmo de K-medias es un método fácil de implementar y usar
- Suele ser un algoritmo bastante eficiente en problemas de "clusterización", pues converge en pocas iteraciones hacía un mínimo de la función J, aunque podría tratarse de un mínimo local.
- Alta dependencia de los valores iniciales asignados a cada centroide (mínimos locales)

K-medias

- Comparación con los mapas de Kohonen:
  - En los mapas de Kohonen no hace falta definir el número de clusters
  - En lo mapas de Kohonen, la función Qerror es igual a J, por tanto ambos algoritmos intentan minimizar la distancia a los centros (o neuronas)
  - Mientras más clusters o más neuronas, menor será el valor de J o Qerror en entrenamiento, pero no tiene por qué ser en test. Es necesario evaluar la función J o Qerror en test
  - En el algoritmo de k-medias no interviene el concepto de vecindario para modificar los centros. No definen estructuras topológicas