Aproximación al reajuste automático de centroides mediante la heurística de Lloyd para resolver el problema de las K-Medias

Xavier Colomé Abril (jcolome0 @ uoc.edu)
(2012) Estudiante Universitat Oberta de Catalunya. www.uoc.edu
Avinguda del Tibidabo, 39, 08035 Barcelona.

Abstracto

En este artículo se describe el uso práctico del algoritmo de Lloyd, utilizado para resolver el problema de las K-Medias, relacionado con las técnicas de reconocimiento de patrones y en concreto asociado con los sistemas de estimación, agrupación y clasificación de entornos en los que se pretende aplicar un aprendizaje no supervisado.

Palabras claves: K-Medias, Lloyd, Aprendizaje no supervisado.

Introducción

El algoritmo de Lloyd es una técnica de agrupamiento que permite unificar *n* objetos en *k* clases. Al agrupar los objetos en distintos grupos o clases obtenemos un método de predicción que podemos aplicar tras encontrar un nuevo elemento. Así mismo podemos establecerlo en una de las clases conocidas, esperando que sus características sean similares a las del resto de elementos de dicha clase.

Este algoritmo comienza dividiendo los elementos de entrada en las distintas clases iniciales, ya sea mediante el uso de datos heurísticos o mediante el azar, para proceder posteriormente con el cálculo del punto medio o centro de gravedad de cada conjunto.

De esta forma se construye una nueva partición, asociando cada elemento con el centro de gravedad más cercano y recalculando de nuevo los centroides.

El algoritmo de Lloyd es pues un método de aprendizaje no supervisado debido a que nos permite obtener información sobre los objetos sin que aportemos de forma explícita un conocimiento previo sobre los mismos.

Modelado Práctico

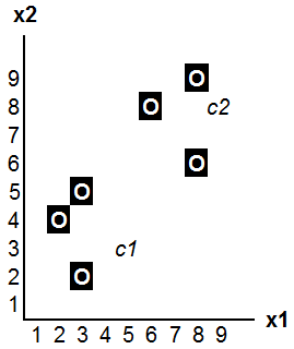
Con el objeto de llevar a cabo la expresión práctica para este problema, partiremos de un ejercicio resoluble. que describiremos a continuación.

De este modo, dada la siguiente representación gráfica del espacio bidimensional $\Re^2$, en la que encontramos elementos de tipo ◘, se han generado de forma aleatoria dos centroides cuyos vectores son:

$$c_1 = (5,3)^t$$

$$c_2 = (9,8)^t:$$

En la siguiente representación gráfica podemos observar tanto los elementos ◘ como los centroides iniciales.



El objetivo será encontrar los elementos, de menor a mayor orden, según los datos del espacio bidimensional descrito, teniendo en cuenta el mayor peso de x_1 respecto a x_2.

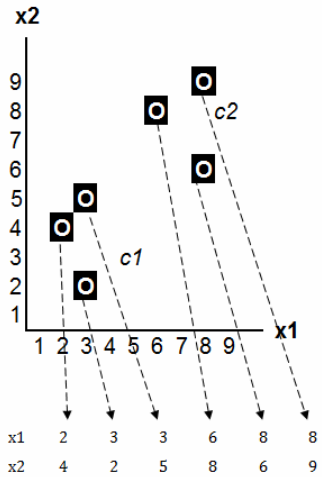
Por otra parte, se deberán ajustar dichos centroides a los objetos ◘ mediante el algoritmo heurístico de Lloyd, teniendo en cuenta un valor arbitrario para $\gamma(k_j) = 0{,}1$ y realizando una sola iteración del algoritmo, calcular los valores del movimiento para c_1 y c_2.

De este modo, la representación gráfica de un espacio bidimensional $\Re^2$, muestra un conjunto de clases u objetos posicionados en el espacio según su naturaleza.

Con las premisas anteriormente descritas, ordenando los objetos de menor a mayor grado de x_1 respecto a x_2, la tabla quedaría completada de la siguiente manera:

x_1	2	3	3	6	8	8
x_2	4	2	5	8	6	9

Esta representación queda definida gráficamente de la siguiente manera y, como podemos observar, x_1 de menor a mayor valor, debido a su mayor peso con respecto a x_2:



De esta forma quedan definidos los siguientes objetos:

$$x^1 = (2,4)^t \qquad x^2 = (3,2)^t \qquad x^3 = (3,5)^t$$
$$x^4 = (6,8)^t \qquad x^5 = (8,6)^t \qquad x^6 = (8,9)^t$$

Como puntos clave de este algoritmo, debemos tener en cuenta que para agrupar *n* datos en *k* grupos o clases, utilizaremos *k* vectores denominados centroides:

$$c_i = \{1,2,\dots,k\} \tag{1}$$

Utilizaremos además una medida de distancia entre objetos y centroides como por ejemplo la Distancia Euclídea:

$$d(x,y) \tag{2}$$

La creación de los centroides se puede realizar de manera aleatoria, teniendo en cuenta que los centroides deben estar suficientemente separados entre sí, para que no tengamos todos los objetos o elementos a distancias similares de todos los centroides.

La idea consiste en establecer a cada objeto el centroide más cercano, calculando el valor de los centroides como la media de los datos que tenga asignados, de forma que, en sucesivas iteraciones, los centroides recalculados se aproximarán cada uno a su región de referencia.

Los pasos del algoritmo son:

1. *Inicialización:* Tras disponer de los puntos de los elementos, se calculan los centroides de salida iniciales mediante una función de aleatoriedad (en nuestro caso).

2. *Asignación:* Cálculo de las distancias Euclídeas entre cada objeto y los centroides. Determinación del centroide más próximo al objeto.

3. *Actualización:* Se actualizarán los valores de los centroides para acercarse al conjunto de objetos que representará. Se repetirá los pasos 2 y 3 hasta la estabilización de los centroides, no obstante, en nuestro caso hemos establecido una única iteración.

Siguiendo estos pasos, iniciaremos el algoritmo de forma didáctica:

1. *Inicialización:* Tras disponer de los objetos en el plano bidimensional, definimos cada centroide:

$$c_1 = (5,3)^t = (j_1, j_2)^t$$

$$c_2 = (9,8)^t = (k_1, k_2)^t$$

De esta forma podremos proceder con el cálculo de la distancia entre cada objeto y cada centroide en el siguiente paso iterativo.

2. *Asignación:* Realizaremos un caso práctico para los cálculos de la norma Euclídea para el primer objeto $x^1 = (2,4)^t = (x_1, x_2)^t$:

La Distancia Euclídea d_1 para el primer objeto x^1, respecto al primer centroide c_1 se calcula con la expresión de la norma Euclídea:

$$d_1 = dE(x^1, c_1) = \sqrt{(j_1 - x_1)^2 + (j_2 - x_2)^2} \tag{3}$$

La Distancia Euclídea d_2 para el primer objeto x^1, respecto al segundo centroide c_2 se calcula con la expresión de la norma Euclídea:

$$d_2 = dE(x^1, c_2) = \sqrt{(k_1 - x_1)^2 + (k_2 - x_2)^2} \tag{4}$$

Los cálculos se definen para el objeto:

$$x^1 = (2,4)^t = (x_1, x_2)^t$$

Las Distancias Euclídeas d_1 y d_2 del objeto x^1 respecto a los centroides c_1 y c_2 son:

Para $c_1 = (5,3)^t = (j_1, j_2)^t$ *:*

$$d_1 = \sqrt{(5-2)^2 + (3-4)^2} = \sqrt{10} = 3{,}16$$

Para $c_2 = (9,8)^t = (k_1, k_2)^t$ *:*

$$d_2 = \sqrt{(9-2)^2 + (8-4)^2} = \sqrt{65} = 8.06$$

3. *Actualización:* De los cálculos observamos que $d_1 < d_2$, que nos indica que el centroide c_1 está más cerca del objeto x^1que del centroide c_2, por lo tanto se debe actualizar el valor del centroide c_1 para acercarse al objeto, que calcularemos mediante la regla de aprendizaje competitivo:

$$c_j(k_j + 1) = c_j(k_j) + \gamma(k_j)[x(k) - c_j(k_j)]$$
$$k_j = k_j + 1 \tag{5}$$

Es decir, el nuevo valor del centroide $c_j(k_j + 1)$ se calcula a través del valor del centroide $c_j(k_j)$ más el valor de $\gamma(k_j)$ por la diferencia entre el objeto y el valor actual del centroide.

Los cálculos para el algoritmo y el caso práctico propuesto, el en que debemos completar una única iteración, quedarían representados de la siguiente manera:

- Primer Centroide: $c_1 = (5,3)^t = \binom{5}{3}$
- Segundo Centroide: $c_2 = (9,8)^t = \binom{9}{8}$
- $\gamma(k_j) = 0{,}1$

<u>Cálculo para el primer objeto</u>:

$$x^1 = (2,4)^t \rightarrow d_1 = 3{,}16 \;\; y \;\; d_2 = 8{,}06$$

Como $d_1 < d_2$ actualizamos el centroide c_1:

$$c_1 = \binom{5}{3} + 0{,}1\left[\binom{2}{4} - \binom{5}{3}\right] =$$
$$= \binom{5}{3} + 0{,}1\binom{-3}{1} =$$
$$= \binom{5}{3} + \binom{-0{,}3}{0{,}1} =$$
$$= \binom{4{,}7}{3{,}1}$$

Se mantiene el valor original para el segundo centroide:

$$c_2 = \binom{9}{8}$$

<u>Cálculo para el segundo objeto</u>:

$$x^2 = (3{,}2)^t \rightarrow d_1 = 2{,}02 \ y \ d_2 = 8{,}49$$

Como $d_1 < d_2$ actualizamos el centroide c_1:

$$c_1 = \binom{4{,}7}{3{,}1} + 0{,}1\left[\binom{3}{2} - \binom{4{,}7}{3{,}1}\right] = \binom{4{,}53}{2{,}99}$$

Se mantiene el valor original para el segundo centroide:

$$c_2 = \binom{9}{8}$$

<u>Cálculo para el tercer objeto</u>:

$$x^3 = (3{,}5)^t \rightarrow d_1 = 2{,}52 \ y \ d_2 = 6{,}71$$

Como $d_1 < d_2$ actualizamos el centroide c_1:

$$c_1 = \binom{4{,}53}{2{,}99} + 0{,}1\left[\binom{3}{5} - \binom{4{,}53}{2{,}99}\right] = \binom{4{,}38}{3{,}19}$$

Se mantiene el valor original para el segundo centroide:

$$c_2 = \binom{9}{8}$$

<u>Cálculo para el cuarto objeto</u>:

$$x^4 = (6{,}8)^t \rightarrow d_1 = 5{,}07 \ y \ d_2 = 3$$

Como $d_2 < d_1$ actualizamos el centroide c_2:

$$c_2 = \binom{9}{8} + 0{,}1\left[\binom{6}{8} - \binom{9}{8}\right] = \binom{8{,}7}{8}$$

Se mantiene el último valor calculado para el primer centroide:

$$c_1 = \binom{4{,}38}{3{,}19}$$

<u>Cálculo para el quinto objeto</u>:

$$x^5 = (8{,}6)^t \rightarrow d_1 = 4{,}58 \ y \ d_2 = 2{,}12$$

Como $d_2 < d_1$ actualizamos el centroide c_2:

$$c_2 = \binom{8{,}7}{8} + 0{,}1\left[\binom{8}{6} - \binom{8{,}7}{8}\right] = \binom{8{,}63}{7{,}8}$$

Se mantiene el último valor calculado para el primer centroide:

$$c_1 = \binom{4{,}38}{3{,}19}$$

<u>Cálculo para el sexto objeto</u>:

$$x^6 = (8{,}9)^t \rightarrow d_1 = 6{,}84 \ y \ d_2 = 1{,}36$$

Como $d_2 < d_1$ actualizamos el centroide c_2:

$$c_2 = \binom{8{,}63}{7{,}8} + 0{,}1\left[\binom{8}{9} - \binom{8{,}63}{7{,}8}\right] = \binom{8{,}57}{7{,}92}$$

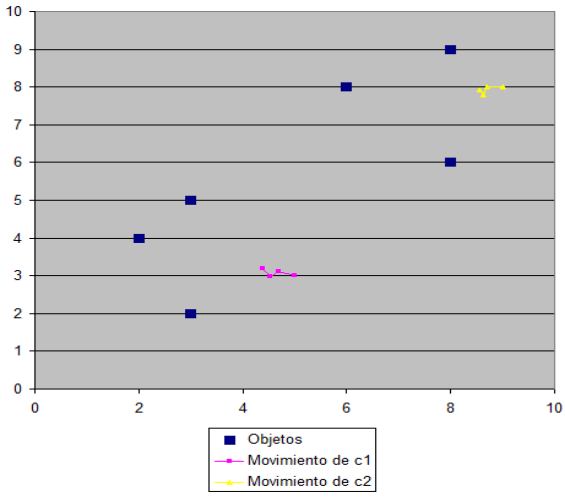
Se mantiene el último valor calculado para el primer centroide:

$$c_1 = \binom{4{,}38}{3{,}19}$$

A efectos prácticos, los movimientos de los centroides se han formalizado en la primera iteración de la siguiente forma:

	x	***y***
Fases de Movimiento del Centroide c_1	5	3
	4,7	3,1
	4,53	2,99
	4,38	3,19
	4,38	3,19
	4,38	3,19
	4,38	3,19
Fases de Movimiento del Centroide c_2	9	8
	9	8
	9	8
	9	8
	8.7	8
	8,63	7,8
	8,57	7,92

Gráficamente la representación de los objetos y del movimiento de los centroides dentro del plano bidimensional sería la siguiente:



Referencias

Ostrovsky, Chaitanya Swamy; Rabani, Yuval; Schulman, Leonard J.; Swamy, Chaitanya. *The Effectiveness of Lloyd-Type Methods for the k-Means Problem.* Dept. of Computer Science, California Univ., Los Angeles, CA.

Pajares, Gonzalo; De la Cruz, Jesús. (2008). *Visión por computador. Imágenes digitales y aplicaciones (2ª edición).* Alfaomega, Ra-Ma Editorial. ISBN 978-84-7897-831-1