Capitolo 7

Clustering

7.1 Clustering k-means: giustificazione del passo di centratura

La funzione di costo del clustering k-means

$$g(\mu) = \sum_{i=1}^{m} \min_{j=1,\dots,k} \left\| x^{(i)} - \mu^{(j)} \right\|^2$$

può essere riscritta, definendo i k cluster

$$C_j = \{i = 1, \dots, m \,:\, j \in \operatorname*{argmin}_{j'} \left\| x^{(i)} - \mu^{(j')} \right\|^2\}$$

come

$$g(\mu) = \sum_{j=1}^{k} \sum_{i \in C_j} \left\| x^{(i)} - \mu^{(j)} \right\|^2.$$

Si osservi che i cluster C_j dipendono dalla posizione dei k rappresentanti $\mu^{(1)}, \dots, \mu^{(k)}$.

Se però consideriamo i cluster C_j come fissati (independentemente dai $\mu^{(j)}$), allora la posizione ottimale del rappresentante di ciascun C_j è data dal baricentro di C_j. Infatti, fissati i C_j, abbiamo che g è una funzione convessa dei $\mu^{(j)}$ e le derivate parziali

$$\frac{\partial g}{\partial \mu_\ell^{(j)}} = 2 \sum_{i \in C_j} (x_\ell^{(i)} - \mu_\ell^{(j)}), \qquad j = 1, \dots, k, \ell = 1, \dots, d,$$

si annullanno se e solo se

$$\sum_{i \in C_j} x_\ell^{(i)} = \sum_{i \in C_j} \mu_\ell^{(j)} = |C_j| \mu_\ell^{(j)}, \qquad j = 1, \dots, k, \ell = 1, \dots, d,$$

ovvero se e solo se

$$\mu^{(j)} = \frac{1}{|C_j|} \sum_{i \in C_j} x^{(i)}, \qquad j = 1, \dots, k.$$

Ricapitolando, l'algoritmo di Lloyd ripete i seguenti passi:

1. Considera fissati i rappresentanti e ottimizza i cluster, associando ciascun punto $x^{(i)}$ al rappresentante $\mu^{(j)}$ più vicino.
2. Considera fissati i cluster e ottimizza i rappresentanti, centrando ciascun rappresentante $\mu^{(j)}$ nel baricentro di ogni cluster C_j.