K-means

Caio Corro

1 Ré-écriture de la dispersion intra-clusters

Soit $X \subseteq \mathbb{R}^m$ un ensemble de points et $\pi = \{C^{(1)}, ..., C^{(k)}\}$ une partition de X. Pour simplifier les notations, on notera :

$$n = |X|\,,$$
$$n_i = |C^{(i)}|.$$

Le centroïde du des données X, que l'on écrit $\bar{\boldsymbol{x}}$, est défini comme :

$$\bar{\boldsymbol{x}} = \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} \boldsymbol{x}\,.$$

Le centroïde du cluster i, que l'on écrit $\boldsymbol{m}^{(i)}$, est défini comme :

$$\boldsymbol{m}^{(i)} = \frac{1}{|C^{(i)}|} \sum_{\boldsymbol{x} \in C^{(i)}} \boldsymbol{x}\,.$$

La dispersion intra-clusters est définie comme :

$$\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C^{(i)}} \|\boldsymbol{x} - \boldsymbol{m}^{(i)}\|_2^2\,.$$

1. Montrer que :
$$\sum_{\boldsymbol{x}, \boldsymbol{x}' \in C^{(i)}} (\boldsymbol{x} - \boldsymbol{m}^{(i)})^\top (\boldsymbol{x}' - \boldsymbol{m}^{(i)}) = 0$$
2. Montrer que :
$$\frac{1}{|C^{(i)}|} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in C^{(i)}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 = 2 \sum_{\boldsymbol{x} \in C^{(i)}} \|\boldsymbol{x} - \boldsymbol{m}^{(i)}\|_2^2$$
3. Utiliser le résultat précédent pour reformuler la dispersion intra-clusters. Comment pouvez-vous interpréter cette nouvelle formulation ?