K means within-point scatter deviation

——Elements of Statistical Learning Chapter 14 Reading Note

Shuai

December 29, 2013

The overall within-point dissimilarity can be replaced with the summation of dissimilarity between each point in the cluster and the cluster mean, meaning:

$$
\begin{aligned}
W(C) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}||x_i - x_{i'}||^2 && (1)\\
&= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x_k}||^2 && (2)
\end{aligned}
$$

where, there are K clusters; $C(i) = k$ means the label given to sample i is k; N_k means the overall number of points in cluster k; x_i means sample i.

In this doc, detailed proof will be presented. During the proof, norm will be just written as square.

For expression 1:

For clarity, we replace $\sum\limits_{C(i)=k}\sum\limits_{C(i')=k}$ with $\sum\limits_{i}\sum\limits_{j}$.

$$
\begin{aligned}
\sum_{C(i)=k}\sum_{C(i')=k}||x_i - x_{i'}||^2 &= \sum_{i}\sum_{j}||x_i - x_{i'}||^2 && (3)\\
&= \sum_{i}\sum_{j}(x_i^2 + x_j^2) - \sum_{i}\sum_{j} x_i x_j && (4)\\
&= N_k\sum_{i} x_i^2 + N_k\sum_{j} x_j^2 - 2\sum_{i}\sum_{j} x_i x_j && (5)\\
&= 2(N_k\sum_{i} x_i^2 + \sum_{i}\sum_{j} x_i x_j) && (6)
\end{aligned}
$$

For expression 2:

For clarity, we replace $\sum\limits_{C(i)=k}$ with $\sum\limits_{i}$.

$$
\begin{aligned}
\sum_{C(i)=k} ||x_i - \bar{x_k}||^2 &= \sum_i ||x_i - \bar{x_k}||^2 && (7)\\
&= \sum_i (x_i^2 + \bar{x_k}^2) - \sum_i x_i \bar{x_k} && (8)\\
&= \sum_i x_i^2 + N_k \bar{x_k}^2 - \sum_i x_i \bar{x_k} && (9)\\
&= \sum_i x_i^2 + N_k \bar{x_k}^2 - 2N_k \bar{x_k}^2 && (10)\\
&= \sum_i x_i^2 - N_k \bar{x_k}^2 && (11)\\
&= \sum_i x_i^2 - \sum_i x_i \sum_j x_j / N_k && (12)\\
& && (13)
\end{aligned}
$$

Since $\sum\limits_i x_i \sum\limits_j x_j = \sum\limits_i \sum\limits_j x_i x_j$, proof is done.