ML Week

0x05 K-Means

Jeff Abrahamson

2–5 novembre 2015

The Problem

Have points $d = \{d_1, \ldots, d_n\}$.

Have number of clusters k.

Want: an assignment of points to clusters

The Algorithm

1. Assign points to clusters at random
2. Repeat until stable:
    1. Compute centroids of each cluster
    2. Assign points to nearest centroid

Cost function

$$\text{cost} = \sum_i \sum_j \left| x_j - \mu_i \right|$$

Silhouette coefficient

Points $d = \{d_1, \ldots, d_n\}$

Clusters $K = \{k_1, \ldots, d_k\}$.

Cluster k_{d_i} is the cluster of d_i.

Silhouette coefficient

Points $d = \{d_1, \ldots, d_n\}$

Clusters $K = \{k_1, \ldots, d_k\}$.

Cluster k_{d_i} is the cluster of d_i.

Let a_i be the average dissimilarity of d_i to all points in its cluster.

Let b_i be the least average dissimilarity of d_i to any cluster other than k_{d_i}

Silhouette coefficient

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

Silhouette coefficient

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

$$s_i = \begin{cases} 1 - a_i/b_i & \text{if } a_i < b_i \\ 0 & \text{if } a_i = b_i \\ b_i/a_i - 1 & \text{if } a_i > b_i \end{cases}$$

Silhouette coefficient

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

$$s_i = \begin{cases} 1 - a_i/b_i & \text{if } a_i < b_i \\ 0 & \text{if } a_i = b_i \\ b_i/a_i - 1 & \text{if } a_i > b_i \end{cases}$$

So $s_i \in [-1, 1]$

Silhouette coefficient

s_i near 1 $\iff$ d_i well clustered

s_i near 0 $\iff$ d_i on the border between two clusters

s_i near -1 $\iff$ d_i well clustered

Silhouette coefficient

Consider $\overline{s_i}$ over $i \in k_j$ for cluster k_j

Silhouette coefficient

Consider $\overline{s_i}$

Questions?

`purple.com/talk-feedback`