K-means and Hierarchical Clustering

Xiaohui Xie

University of California, Irvine

Clustering

Given n data points $X = \{x_1, x_2, \cdots, x_n\}$. Clustering is the partitioning of the set X into subsets (clusters), so that the data in each subset share some “similarity” - according to some defined distance measure.

- Similarity measure between any two points $d(x_i, x_j)$. For example, Euclidean distance: $d(x_i, x_j) = \sqrt{\|x_i - x_j\|^2}$.
- Number of clusters: K
- K subsets (clusters): $S_1, S_2, \cdots, S_K$ where

$$S_i \subset X \quad \forall i \tag{1}$$

$$S_i \bigcap S_j = \phi \quad \forall i \neq j \tag{2}$$

$$\bigcup_{i=1}^{K} S_i = X \tag{3}$$

K-means

- Choose K and randomly guess K cluster Center locations
- Repeat until convergence
  1. Each datapoint finds out which Center it's closest to.
  2. Each Center finds the centroid of the points it owns.

K-means Questions

- What is it trying to optimize?
- Are we sure it will terminate?
- Are we sure it will find an optimal clustering?
- How should we start it?
- How to automatically choose the number of centers?

K-means Error function

- Given n datapoints $X = \{x_1, x_2, \cdots, x_n\}$. Partition them into K clusters: $S_1, S_2, \cdots, S_K$.
- Define an error function:

$$V(S_1, \cdots, S_K, c_1, \cdots, c_K) = \sum_{i=1}^{K} \sum_{x_j \in S_i} d(x_j, c_i)$$

K-means error function

- Assign each datapoint to the closest center. Suppose after the assignment, the updated clusters are $S_1^*, \cdots, S_K^*$. Then,

$$V(S_1^*, \cdots, S_K^*, c_1, \cdots, c_K) \leq V(S_1, \cdots, S_K, c_1, \cdots, c_K)$$

- Each Center finds the centroid of the points it owns.

$$c_i = \arg\min_c \sum_{x_j \in S_i} d(x_j, c)$$

For Euclidean distance measure: $c_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$
After the update:

$$V(S_1, \cdots, S_K, c_1^*, \cdots, c_K^*) \leq V(S_1, \cdots, S_K, c_1, \cdots, c_K)$$

- Thus both steps decrease the error function (if there is any update).

Will we find the optimal configuration?

- Not necessarily.
- Can you find a configuration that has converged, but does not have the minimum error?

Trying to find good optima

- Carefully choose the starting Centers.
- Run K-means multiple times each from a different start configuration.
- Many other ideas.

How to choose the number of Centers?

- In general a difficult problem. Related to model selection.
- Bayesian information criterion (BIC)

$$BIC = V + \lambda mK \log n$$

  where m is the dimension of the data, or in other words, mK is the total number of free parameters.
- Cross-validation
- Non-parametric Bayesian method

Single linkage hierarchical Clustering

- Initialize: “Every point is its own cluster”.
- Find “most similar” pair of clusters
  - Minimum distance between points in clusters
  - Maximum distance between points in clusters
  - Average distance between points in clusters
- Merge it into a parent cluster
- Repeat ... until you have merged the whole dataset into one cluster.

Pros and Cons of Hierarchical Clustering

- The result is a dendrogram, or hierarchy of datapoints.
- To choose K clusters, just cut the $K - 1$ longest links
- Cons: No real statistical or information theoretical foundation for the clustering.

Probabilistic interpretation of K-means

- Input: n data points: $x = \{x_1, \cdots, x_n\}$.
- Model: A mixture model of K components (clusters). Each component is described by a normal distribution:

$$x \sim N(\mu_i, \sigma_i)$$

for the i^{th} component. In other words, if x belongs to cluster i

$$p(x|x \in S_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\|x-\mu_i\|^2}{2\sigma_i^2}}$$

We will use $\theta_i = (\mu_i, \sigma_i)$ and $\theta = (\theta_1, \cdots, \theta_K)$ to denote the parameters of the model.

- The assignment of the data to each of the clusters: $z = \{z_1, \cdots, z_n\}$ with $z_i \in \{1, 2, \cdots, K\}$.

Probabilistic K-means: inference

- The probability of generating the data given the model and the label:

$$p(x|z,\theta) = \prod_{i=1}^{n} p(x_i|z_i,\theta)$$

where $p(x_i|z_i,\theta) = p(x_i|\mu_{z_i},\sigma_{z_i})$

- The probability of generating the data given the model and the label:

$$p(x|z,\theta) = \prod_{i=1}^{n} p(x_i|z_i,\theta)$$

where $p(x_i|z_i,\theta) = p(x_i|\mu_{z_i},\sigma_{z_i})$

- The probability of generating the data given the model only:

$$p(x|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{K} P(z_i = j) p(x_i|z_i = j,\theta)$$

Probabilistic K-means: EM-algorithm

- ML estimate of the mixture model

$$\theta^* = \operatorname*{argmax}_{\theta} \log p(x|\theta)$$

- EM-algorithm: E-step

$$q(z_i = j) = p(z_i = j|\theta_j, x_i) \quad (4)$$

$$\sim \frac{1}{\sigma_j} e^{-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}} \quad (5)$$

Probabilistic K-means: EM-algorithm

- EM-algorithm: M-step

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{K} q(z_i = j) \log p(x_i|\theta_j) \tag{6}$$

$$\hat{\theta}_j = \operatorname*{argmax}_{\theta_j} \sum_{i=1}^{n} q(z_i = j) \log p(x_i|\theta_j) \tag{7}$$

$$= \operatorname*{argmin}_{\theta_j} \sum_{i=1}^{n} q(z_i = j)[\|x_i - \mu_j\|^2/(2\sigma_j^2) + \log \sigma_j] \tag{8}$$

Probabilistic K-means: EM-algorithm

- If we hold $\sigma_i = \sigma$ fixed for all i and only learn μ_i, then

$$\hat{\mu}_j = \operatorname*{argmin}_{\mu_j} \sum_{i=1}^{n} q(z_i = j) \|x_i - \mu_j\|^2 \tag{9}$$

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} q(z_i = j) x_i}{\sum_{i=1}^{n} q(z_i = j)} \tag{10}$$

- In summary, the probabilistic K-means alternates between
  - E-step: $q(z_i = j) \leftarrow e^{-\frac{\|x_i - \mu_j\|^2}{2\sigma^2}} / Z$
  - M-step: $\mu_j \leftarrow \frac{\sum_{i=1}^{n} q(z_i = j) x_i}{\sum_{i=1}^{n} q(z_i = j)}$

Relationship to the standard K-means – 1

- Under what condition does the EM-algorithm become the standard K-means?
- Note that in the E-step:

$$q(z_i = j) = \frac{e^{-\frac{\|x_i - \mu_j\|^2}{2\sigma^2}}}{\sum_{k=1}^{K} e^{-\frac{\|x_i - \mu_k\|^2}{2\sigma^2}}} \tag{11}$$

$$= \frac{e^{-\frac{\|x_i - \mu_j\|^2 - \|x_i - \mu_m\|^2}{2\sigma^2}}}{1 + \sum_{k=1, k\neq m}^{K} e^{-\frac{\|x_i - \mu_k\|^2 - \|x_i - \mu_m\|^2}{2\sigma^2}}} \tag{12}$$

where $m = \mathrm{argmin}_k \|x_i - \mu_k\|^2$, is the index of the cluster closest to x_i (we assume m is unique).

Relationship to the standard K-means – 2

- Let $\sigma \to 0$,
  - For all $k \neq m$, $\|x_i - \mu_k\|^2 - \|x_i - \mu_m\|^2 > 0$, hence the denominator in above Eq. $\to 1$ as $\sigma \to 0$.
  - The numerator is 1 only when $j = m$, otherwise 0.
- In summary, as $\sigma \to 0$,

$$q(z_i = j) = \begin{cases} 1 & \text{if } j = m \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $m = \operatorname{argmax}_k \|x_i - \mu_k\|^2$ is the index of the cluster closest to x_i.
Note that this is just the standard K-means procedure.