K-means Clustering
Lecture notes for Cmput466/551 30/Mar/05
S Wang

K-means: one of the most popular iterative descent clustering method.

Given a set of observations $(x_1, \cdots, x_N)$, a prespecified number of clusters $K < N$ is postulated, and each observation x_i is assigned to one and only one cluster which is denoted as $C(i)$.

Assume we are using squared Euclidean distance $d(x_i, x_{i'}) = ||x_i - x_{i'}||$ to denote *dissimilarity* of pair of observations $x_i, x_{i'}$.

For a cluster assignment C, define its loss function as

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'}) \tag{1}$$

This criterion characterizes the extent to which observations assigned to the same cluster tend to be close to one another. It is referred to as *within-cluster* point scatter.

Similarly we can define *between-cluster* point scatter,

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d(x_i, x_{i'}) \tag{2}$$

This will tend to be large when observations assigned to different clusters are far apart.

Define the *total* point scatter,

$$T = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d(x_i, x_{i'}) \tag{3}$$

which is a constant given the data, independent of cluster assignment.

$$\begin{aligned} T &= \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d(x_i, x_{i'}) \\ &= \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \Big(\sum_{C(i')=k} d(x_i, x_{i'}) + \sum_{C(i')\neq k} d(x_i, x_{i'}) \Big) \\ &= W(C) + B(C) \end{aligned}$$

Thus one has

$$W(C) = T - B(C)$$

So minimizing $W(C)$ is equivalent to maximizing $B(C)$.

The *within-cluster* point scatter can be written as

$$\begin{aligned} W(C) &= \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x_{i'}||^2 \\ &= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \hat{x}_k||^2 \end{aligned} \tag{4}$$

where $\hat{x}_k$ is the mean vector associated with the kth cluster, and $N_k = \sum_{i=1}^{N} I(C(i) = k)$. Thus, the criterion is minimized by assigning the N observations to the K clusters in such a way that with each cluster the average dissimilarity of the observations from the cluster mean, as define by the points in that cluster, is minimized.

Thus the optimal assignment is

$$C^* = \min_C \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \hat{x}_k||^2 \tag{5}$$

First noting that for any set of observations S

$$\hat{x}_S = \arg\min_m \sum_{i\in S} ||x_i - m||^2 \tag{6}$$

Hence we can obtain C^* by solving the enlarged optimization problem

$$\min_{C,\{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2 \tag{7}$$

This can be minimized by an alternating optimization procedure as the following:

K-means clustering

1. For a given cluster assignment C, the total cluster variance (7) is minimized with respect to $\{m_1, \cdots, m_K\}$ yielding the means of the currently assigned clusters (8).
2. Given a current set of means $\{m_1, \cdots, m_K\}$, (7) is minimized by assigning each observation to the closest (current) cluster mean. That is,

$$C(i) = \arg\min_m C = \arg \min_{1\leq k\leq K} ||x_i - m_k||^2 \tag{8}$$

3. Steps 1 and 2 are iterated until the assignments do not change.

Each steps 1 and 2 reduces the value of (7), and (7) is bounded below by 0, so that convergence is assured.

Principal Component Analysis (PCA)

PCA: a dimensionality reduction method.

Given a set of observations $(x_1, \cdots, x_N), x \in \Re^p$, find best hyperplane of rand q to represent the data.

$$\hat{x} = \mu + V_q\lambda, \qquad q < p \tag{9}$$

where $\mu \in \Re^p$ a location vector, V_q, a $p \times q$ orthonormal matrix,

$$v_i^T v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$v_i, i = 1, \cdots, q$: orthogonal unit vectors, $\lambda \in \Re^q$: parameters, $V_q\lambda$: a subspace of $\Re^p$.

Reconstruction error:

$$\sum_{i=1}^{N} ||x_i - \hat{x}_i||^2 \tag{10}$$

Choose $u, \{\lambda_i\}, V_q$ to minimize the reconstruction error,

$$\min_{u,\{\lambda_i\},V_q} \sum_{i=1}^{N} ||x_i - \mu - V_q\lambda_i||^2 \tag{11}$$

We can partially optimize for u and λ_i's to obtain

$$\begin{aligned} \hat{\mu} &= \bar{x} \text{ sample mean} && (12) \\ \hat{\lambda}_i &= V_q^T(x_i - \bar{x}) && (13) \end{aligned}$$

This leaves us to find the orthonormal normal matrix V_q:

$$\min_{V_q} \sum_{i=1}^{N} ||x_i - \bar{x} - V_q V_q^T (x_i - \bar{x})||^2 \tag{14}$$

For convience, we assume that $\bar{x} = 0$ (otherwise we simplely replace the observations by their centered versions $\tilde{x}_i = x_i - \bar{x}$).

Let $H_q = V_q V_q^T$, projection matrix, maps each point x_i onto its rank q reconstruction $H_q x_i$, orthogonal projection of x_i onto the subspace spanned by $\{v_i\}, i = 1, \cdots, q$.

Stack $x_1, \cdots, x_N$ to form an $N \times p$ matrix A

$$A_{N\times p} = U_{N\times p} D_{p\times p} V_{p\times p}^T \tag{15}$$

U: $N \times p$ orthogonal matrix, $U^T U = I_p$, V: $p \times p$ orthogonal matrix, $V^T V = I_p$, D diagonal matrix, $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ singular values. u_i: left singular vectors, v_i right singular vectors.

Columns of UD: principal components of A.

Optimal $\hat{\lambda}_i, i = 1, \cdots, q$:

$$\underline{\hat{\lambda}} = U_q D_q \tag{16}$$

Theorem 1 *(The Eckart and Young theorem) Let the SVD of A be $A = \sum_{k=1}^{p} d_k u_k v_k^T$ with $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. Let $\hat{A}_q$ denote the truncated sum, $\hat{A}_q = \sum_{k=1}^{q} d_k u_k v_k^T$ q integer, $1 \leq q \leq p - 1$, then*

$$\min_{B \text{ of rank} \leq q} ||A - B||_F = \sqrt{\sum_{k=q+1}^{p} d_k^2} \tag{17}$$

and a minimizer is $B = \hat{A}_q$. The minimizer of $||A - B||_F$ is unique iff $d_q > d_{q+1}$.

$||A||_F \doteq \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{p} |a_{ij}|^2} = \sqrt{\text{trace}(A^T A)}$: Frobenius norm of a matrix, square root of the sum of squares of all elements in the matrix.

References

[1] T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Spring-Verlag. 2001.