36-462 Data Mining Recitation Notes

Week 4

Li Liu

Department of Statistics

Carnegie Mellon University

(Feb 4, 2013)

Abstract

In this recitation we will review K-means, K-medoids and hierarchial clustering. I will also give a brief introduction of model based clustering.

1 Clustering

1.1 What is clustering?

(a) Clustering: dividing data subjects into clusters so that data subjects are similar with each other in the same cluster, and dissimilar to the subjects in different clusters.

(b) Unlike classification, clustering is unsupervised learning. We have no training data.

1.2 Why clustering

(a) As a tool to summary or discover

(b) As a preprocessing for other algorithms

1.3 Major clustering approaches

(a) Centroid based clustering

(b) Hierarchical clustering

(c) Model based clustering

(d) Others:spectral clustering, density based clustering,....

2 Centroid based clustering

2.1 Key idea

The idea of K-means or K-medoids is to minimize within-cluster scatter (dissimilarity). The definition of within-cluster scatter is:

$$W = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{C(i)=k, C(j)=k} d_{ij}, \tag{1}$$

where K is the number of clusters, d_{ij} is the dissimilarities between subjects i and j. $C(i) = k$ means subject i is assigned to cluster k. n_k is the number of points in the group k.

2.2 K-means algorithm

Observations are $X_1, ...X_n$. If we use Euclidean distance $||X_i - X_j||_2^2$ as the dissimilarity measure, then K-means can be implemented as:

(a) Give initial values for cluster centers $c_1, ...c_K$.

(b) For each i, find the cluster center c_k closet to X_i, and let $C(i) = k$.

(c) For each k, let $c_k = \bar{X}_k$.

2.3 K-medoids and K-medians approaches

(a) K-medoids approach chooses data points as centers.

(b) K-medians approach chooses the medians as centers. This has the effect of minimizing distance over all clusters with respect to the L_1 distance metric.

(c) K-medoids and K-medians are more robust to noise and outliers as compared to K-means.

2.4 How to choose K

(a) CH index

$$\mathrm{CH}(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}, \tag{2}$$

$$\text{where} \quad B(K) = \sum_{k=1}^{K} n_k ||\bar{X}_k - \bar{X}||_2^2,$$

$$W(K) = \sum_{k=1}^{K} \sum_{C(i)=k} ||X_i - \bar{X}_k||_2^2.$$

(b) Gap statistics

$$\mathrm{Gap}(K) = \log W(K) - \log W_{\text{unif}}(K), \tag{3}$$

where $W_{\text{unif}}(K)$ is the within-cluster variation we'd see if we had points distributed uniformly.

3 Hierarchical clustering:

3.1 Important concepts

(a) Dendogram: A tree where each node represents a group, each leaf node is a singleton and each internal node has two children nodes.

(b) Linkages: The way to measure the dissimilarity between two groups.

3.2 Two types of hierarchical clustering

(a) Agglomerative (bottom-up): start with all points in their own group

(b) Divisive (top down): start with all points in one cluster

3.3 Different types of linkages

(a) Single linkage: the dissimilarity between groups G and H is the smallest dissimilarity between two points in opposite groups.

(b) Complete linkage: the dissimilarity between groups G and H is the largest dissimilarity between two points in opposite groups.

(c) Average linkage: the dissimilarity between groups G and H is the average dissimilarity between two points in opposite groups.

(d) Centroid linkage: the dissimilarity between the group averages.

(e) Minimax linkage: the smallest radius of all points in groups G and H. The radius of one point is defined as the distance between this point and the furthest point in the opposite group.

4 (Optional) Model based clustering:

4.1 Basic idea: clustering as probability estimation. It's a soft clustering.

K-means and hierarchical clustering are nonparametric approaches and model based clustering is parametric approach.

4.2 Mixture of normal distribution

$$X \sim \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k), \tag{4}$$

(a) π_k is the probability that an object belongs to cluster k, given no observation information.

(b) μ_k are the cluster center, and Σ_k are the variance.

(c) need to estimate the membership of each subject, π_k, μ_k and Σ_k (can be assumed as diagonal matrix and same for all clusters, or even as given.).

4.3 EM (Expectation-Maximization) algorithm

K-means is a case of EM algorithm.

(a) Give initial value to μ_k, π_k and Σ_k.

(b) E-step: For all subjects and all clusters, estimate the membership value y_{ik}, which is defined as the probability of subject i belongs to cluster k.

$$y_{ik} = \frac{\pi_k p(X_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_k p(X_i; \mu_j, \Sigma_j)} \tag{5}$$

(c) M-step: Estimate μ_k, π_k and Σ_k based on X_i and y_{ik}.

$$\begin{aligned}
\pi_k &= \frac{1}{N} \sum_{i=1}^{N} y_{ik} \\
\mu_k &= \frac{\sum_{i=1} y_{ik} X_i}{\sum_{i=1} y_{ik}} \\
\Sigma_k &= \frac{\sum_{i=1} y_{ik} [X_i - \mu_k][X_i - \mu_k]^T}{\sum_{i=1} y_{ik}}
\end{aligned} \tag{6}$$

(d) repeat E & M steps until convergence.