TIM245 Lecture 16 5/29/17

Agenda

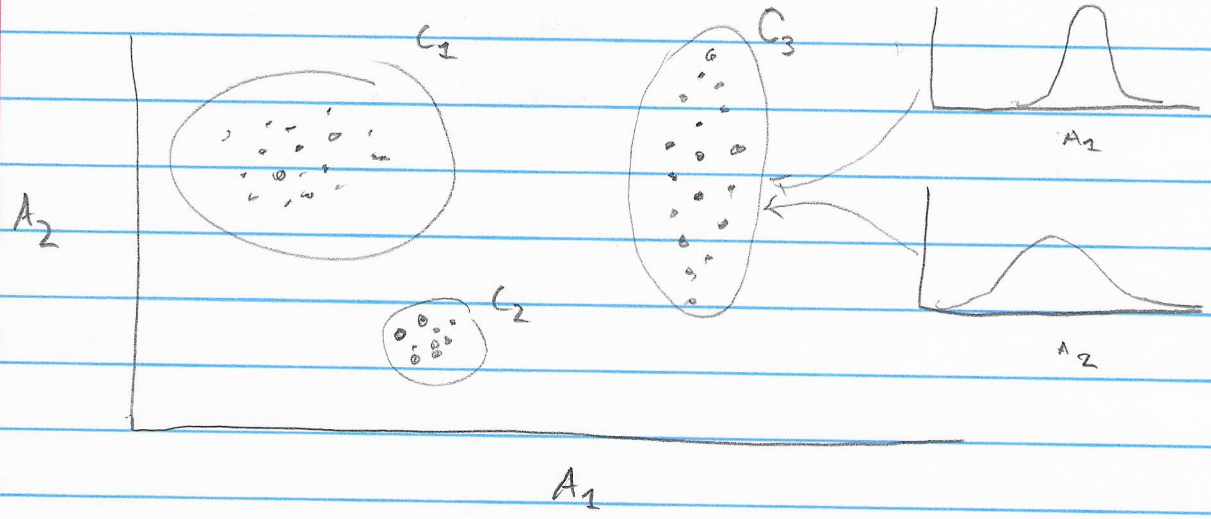
1) Model Based Clustering Methods and the EM Algorithm

2) Evaluation of Clustering Results

3) General Comments on Clustering

① Model Based Clustering Methods and the EM Algorithm

Basic Idea: Clusters are modeled using statistical distributions (Mixture Models)



Each multivariate distribution corresponds to a cluster where the distribution parameters describe the cluster, i.e. central tendency and spread

Each instance has a probability of belonging to a cluster based on the cluster's multivariate distribution

Find the K multivariate distributions that best fit the observed data, i.e. maximize the likelihood of the observed data

Expectation Maximization (EM) Algorithm

1) Select an initial set of parameters for each of the k statistical distributions, typically multivariate normal

2) For each instance, calculate the probability of membership to the k distributions or clusters

$$P(C_L \mid X_i)$$

(This is the expectation step)

3) Given the membership probabilities, find the new parameter estimates that maximize the expected likelihood

$$L(\theta \mid X)$$

(This is the maximization step)

4) Repeat until the parameters do not change

Advantages

- Can model non-spherical shapes, e.g. ellipses
- Provides a concise way of interpreting clustering results using the distribution parameters
- Easy to add domain knowledge about attributes

Disadvantages

- Slow
- Has trouble with small clustes
- Has trouble with noise and outliers
- K is still fixed

② Evaluation of Clustering Results

Clustering results should be validated using evaluation metrics that are both internal and external to the given data set

Internal Metrics

Measure the goodness of fit of the clusters

External Metrics

Measure the utility or usefulness of of the clusters

Two different ways of measuring goodness of fit

Cohesion

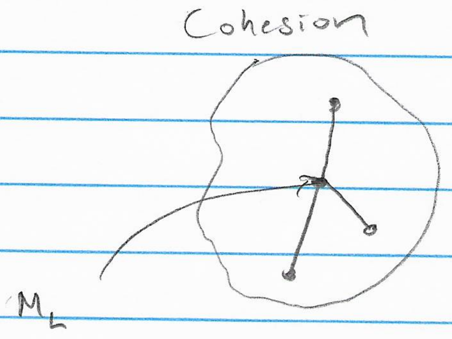
How closely related are the instances in the same cluster?

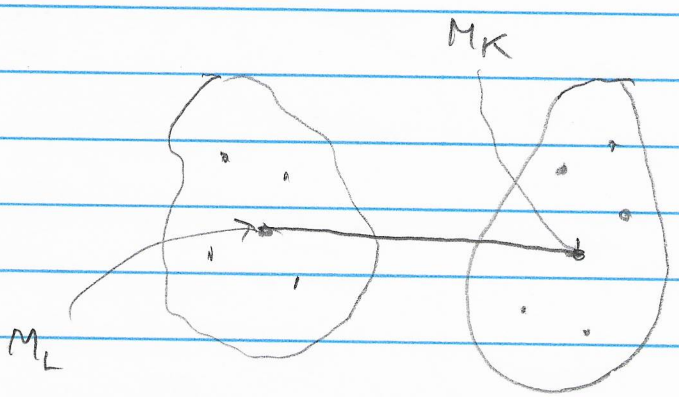
Separation

How distinct or well separated is a cluster from the other clusters?

Cohesion and Separation are measured differently depending on the type of Clustering.

Partitioning Methods : distances are measured (K-means , EM) from the cluster centroids



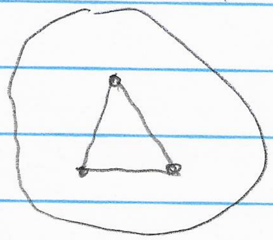


$$\text{Cohesion}\ (C_L) = \sum_{X_i \in C_L} (X_i - M_L)^2$$

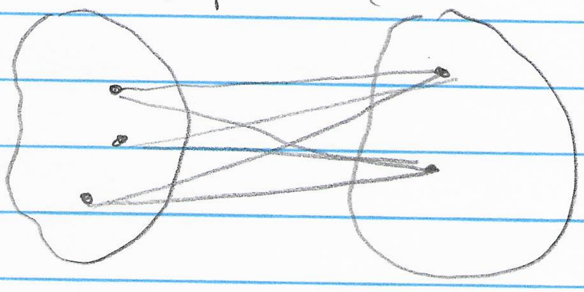
$$\text{Separation}\ (C_L, C_K) = (M_L - M_K)^2$$

Graph-Based Methods : distances are measured
(Hierarchical, DB-Scan) between points

Cohesion



Separation



$$\text{Cohesion}(C_L) = \sum_{\substack{x_i \in C_L \\ y \in C_L}} (x_i - y_i)^2$$

$$\text{Separation}(C_L, C_k) = \sum_{\substack{x \in C_L \\ y \in C_k}} (x - y)^2$$

Silhoutte Coefficient : combines cohesion and separation into a single metric

Process for Computing

1) Calculate the average distance to all other instances in the same cluster. Let this value be a_i

2) Calculate the average distance to all other clusters. Let b_i denote the minimum distance.

3) The silhouette coefficient for the ith instance is

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

ranges from -1 to 1, negative values typically indicate a poor clustering

4) Compute for $i = 1, 2, \ldots, n$ and average the results

③ General Comments on Clustering

1) Make sure that there is non-random structure in the data and that the algorithm assumptions match the data.
    - Exploratory Data Analysis
    - Domain Knowledge
    - Experiment with different algorithms and check the results using internal measures cohesion, separation, and silhouette coefficient

2) Clustering is an iterative process. Involve domain experts early in the process to provide external evaluation and guidance.

3) Clustering results are typically messy and will need post-processing before they can be used.

4) Always plot the data. TSNE and other visualization techniques can be useful when plotting high-dimensional data