Clustering and Unsupervised Learning

Learning from unclassified data, Clustering, Hierarchical Aglomerative Clustering, k-means partitioning clustering, EM for soft clustering, Semi supervised learning with EM using labeled and unlabeled data

Unsupervised Learning

- Predicting target variables using predictor variables is the supervised learning example.
- Unsupervised learning addresses the learning without a supervisor.
- A set of observations of input variable X is given. The goal is to infer the properties of the probability densities without the help of a supervisor.
- It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms.

Association Rules

- are used for market basket analysis.
- It determines the joint values of the variables that appear most in the database.
- The primary goal is to find a collection of prototype X-values v1,v2,...,vL for the feature vector X, such that the probability density Pr(v_l) evaluated at each value is relatively large.

Anomaly Detection

- Identify abnormal or rare observations that can raise suspicions
- Time series anomaly detection is the task of finding patterns that do not conform to ‘‘normal’’ behaviour.
- A model learns behaviour or characteristics and detects significant diversion from the normal pattern.
- Regardless of the domain, anomalies translate to meaningful actionable information which experts can act against.

Cluster Analysis

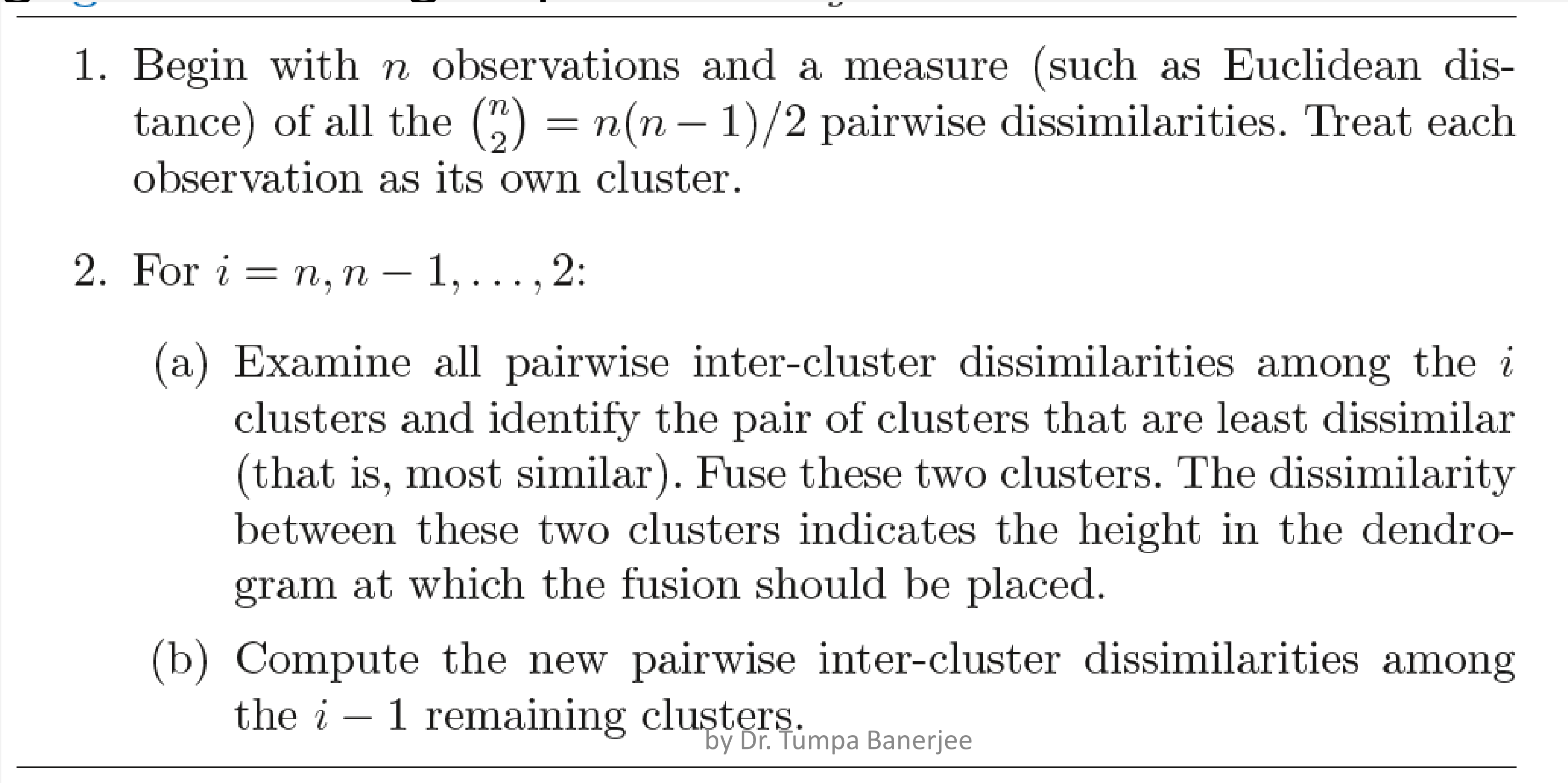
- Grouping a collection of objects into subsets or clusters such that those within each cluster are more closely related to one another than objects assigned to different clusters.
- This involves successively grouping the clusters themselves so that at each level of the hierarchy, clusters within the same group are more similar to each other than those in different groups.
- Cluster analysis is used to form descriptive statistics to ascertain whether or not the data consists of a set of distinct subgroups.

Hierarchical Clustering

- Clustered can be nested each other.
- Two approach of hierarchical clustering are: bottom up or agglomerative clustering and top down clustering.
- Both the approach use dissimilarity matrix in use.
- In bottom up approach most similar groups are merged together at each step.
- In top down approach groups are split using various different criteria.
- Both the approach do not have any objective function to optimize.

Agglomerative Clustering

- Start with N groups, each group initially contain only one datapoint.
- Merge two similar groups into one.



1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

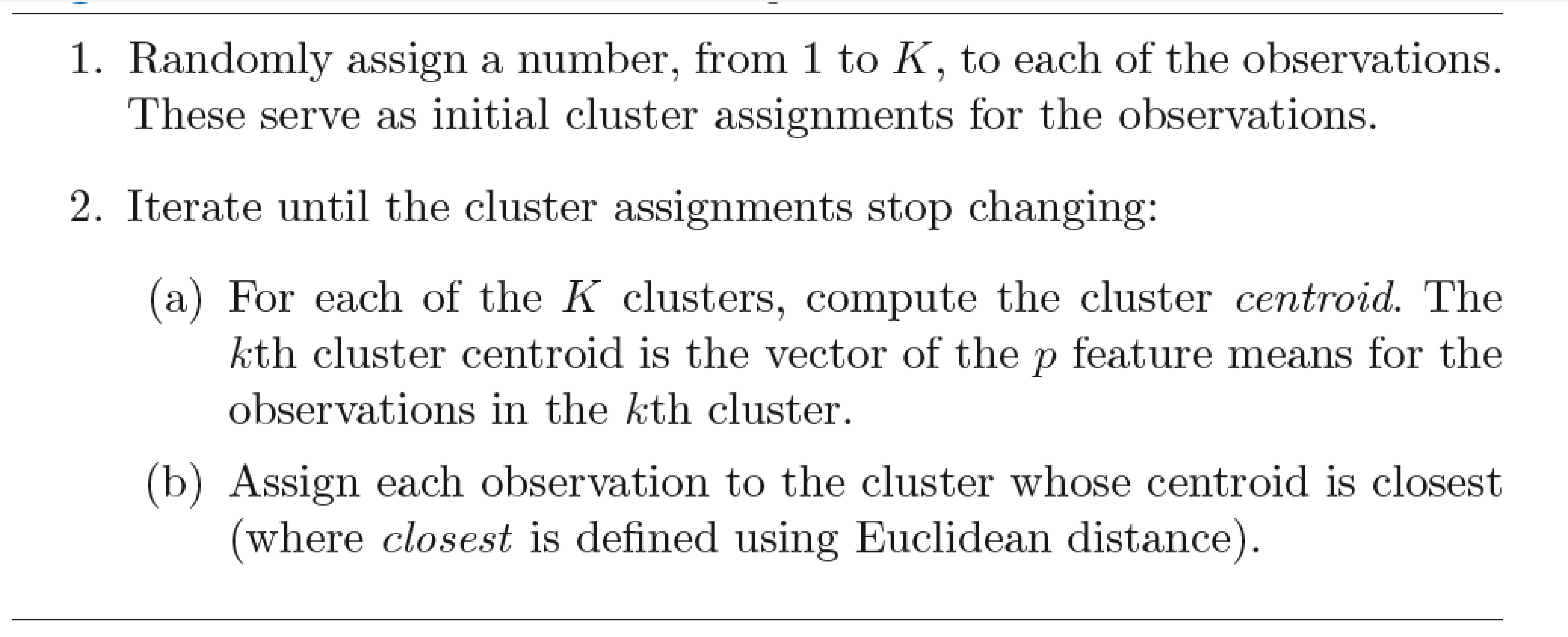
Types of Linkage in Clustering

- The various types of linkages describe distinct methods for measuring the distance between two sub-clusters of data points, influencing the overall clustering outcome.
- Single linkage: For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.
- Complete Linkage: For two clusters R and S, the complete linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.
- Average Linkage: For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated.

K-means Clustering

- first specify the desired number of clusters K
- the clusters should have minimum distance between the points within clusters and maximum distance between the clusters.
- the center of the cluster considered as centroid.
- Initialize centroid for k clusters with random k datapoints.
- Assign the datapoints to the cluster with centroid c_k such that the distance of each datapoint to its closet centroid is minimum.
- Compute centroid of each cluster and reassign the datapoints.

K-means Clustering



1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

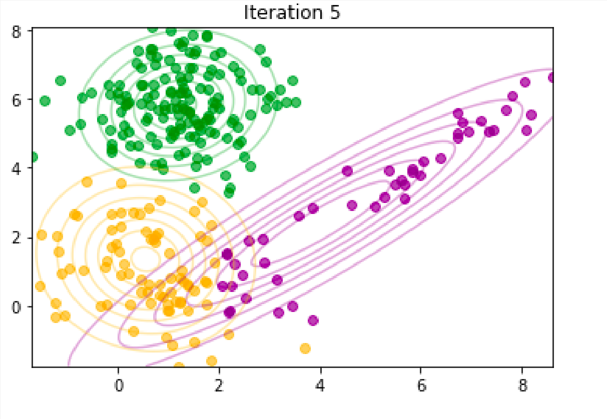
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Types of Clustering

- Hard clustering: Assign each data point to a single, exclusive cluster based on similarity.
- Soft clustering: Assign a degree of membership or probability to each cluster for a data point.

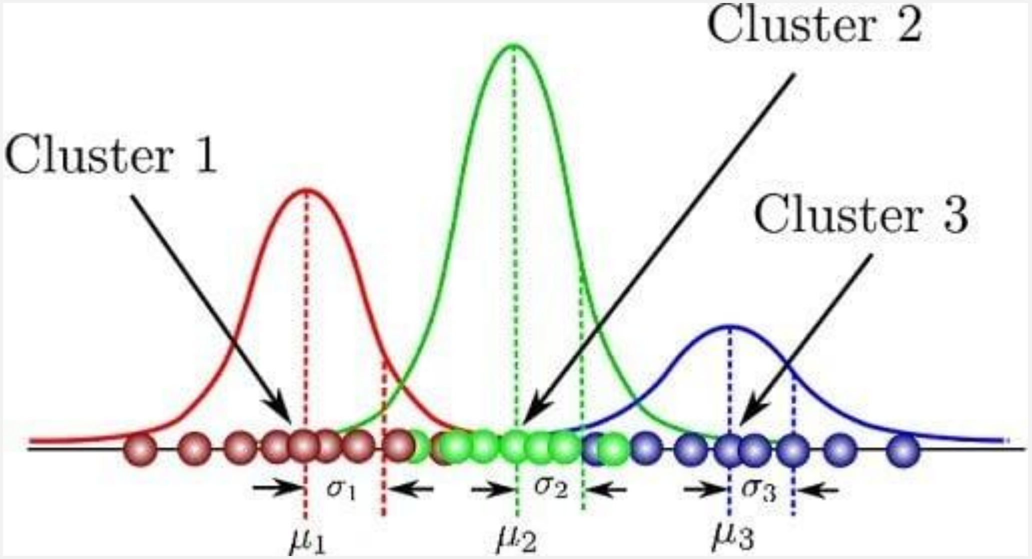
Gaussian Mixture Model

- A Gaussian mixture is a function that is composed of several Gaussians, each identified by k ∈ {1,…, K}, where K is the number of clusters of our data set. Each Gaussian k in the mixture is comprised of the following parameters:
  - A mean μ that defines its center.
  - A covariance Σ that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario.
  - A mixing probability π that defines how big or small the Gaussian function will be.



Gaussian Mixture Model

- The mixing coefficients are themselves probabilities and must meet this condition: $\sum_k \pi_k = 1$
- To achieve this we must ensure that each Gaussian fits the data points belonging to each cluster. This is exactly what maximum likelihood does.



Gaussian Mixture Model

- The Gaussian density function is given by:
- $\mathcal{N}(x|\Sigma,\mu) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$
- We can take log of this equation, then we differentiate this equation with respect to the mean and covariance and then equate it to zero, then we will be able to find the optimal values for these parameters, and the solutions will correspond to the maximum likelihood estimation(MLE) for this setting.

Gaussian Mixture Model Expectation-Maximization (EM)

- Used for optimization problems where the objective function has complexities like the Gaussian Mixture Model(GMM).
- The parameters our model be: $\theta = \{\mu, \Sigma, \pi\}$
- Initialize θ randomly.
- Expectation Step: calculate the posterior probabilities of data points belonging to each centroid using the current parameter values.
- Maximization Step: adjust the parameters to fit the data points assigned to them
- Iterate until convergence