ST720 Data Science

Unsupervised Learning

Seung Jun Shin (sjshin@korea.ac.kr)

Department of Statistics, Korea University

Clustering I

- Generate groups of observations (or variables) based on their similarity.
- Given $\mathbf{x}_1, \cdots, \mathbf{x}_n, i = 1, \cdots, n$,
  - Euclidean Distance:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\| = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$$

  - Manhattan Distance:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} |x_{ij} - x_{i'j}|$$

Clustering II

- Standardization Required.
  - Mean-Variance

$$x_{ij} \to x_{ij}^* = \frac{x_{ij} - m_j}{s_j}, \quad j = 1, \cdots, p.$$

where m_j and s_j denote the sample mean and SD, respectively.

  - Min-Max

$$x_{ij} \to x_{ij}^* = \frac{x_{ij} - l_j\}}{u_j - l_j}, \quad j = 1, \cdots, p.$$

where l_j and u_j denote the sample minimum and maximum, respectively.

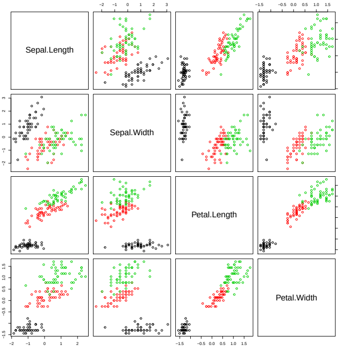
K-means Clustering I

- K-means Clustering: Assume that the number of clusters is given by K,
  1. (Initialization) Randomly select k observation and let them be the centers (means) of K clusters, respectively.
  2. Assign cluster to every observation based on the distance from the cluster center.
  3. Update cluster means (centers).
  4. Repeat Steps 2-3 until convergence (membership of all observations remain unchanged) .

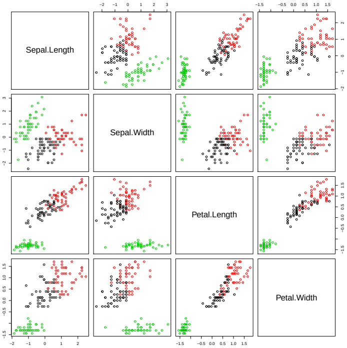
K-means Clustering II

- Minimizes the within cluster sum of squares (for a given K).
- Computationally efficient.
- Suitable for continuous variables.
- k is assumed to be known.
- Returns local solution.

K-means Clustering III



(a) True Class



(b) Estimated Cluster

Figure: K-means Clustering to iris data.

Hierarchical Clustering I

- Types:
  - Agglomeration: from n groups to a single group.
  - Division: from a single group to n groups.
- Visualization via Dendrogram is useful.

Hierarchical Clustering II

- Toy example: Input (Distance Matrix)

	1	2	3	4	5
1	0				
2	7	0			
3	1	6	0		
4	9	3	8	0	
5	8	5	7	4	0

- Step 1

	(1,3)	2	4	5
(1,3)	0			
2	6	0		
4	8	3	0	
5	7	5	4	0

Hierarchical Clustering III

- Step 2

(1,3)	0		
(2,4)	6	0	
5	8	4	0
	(1,3)	(2,4)	5

- Step 3

(1,3)	0	
(2,4,5)	6	0
	(1,3)	(2,4,5)

- Step 4: Cluster (1,2,3,4,5)

Hierarchical Clustering IV

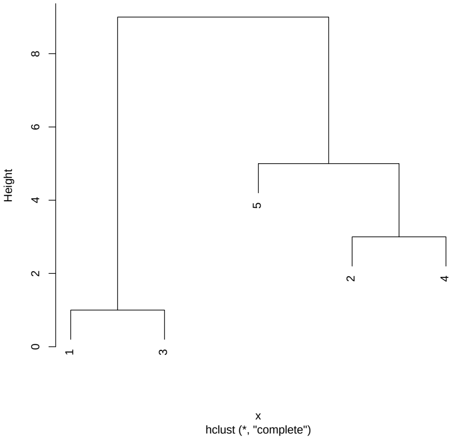
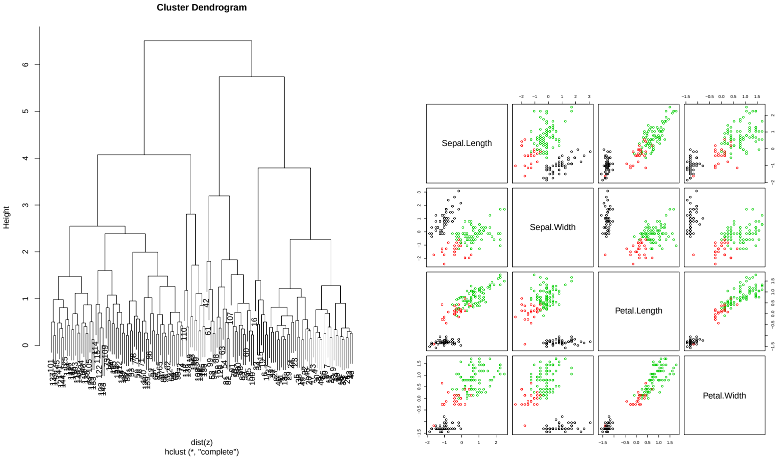


Figure: Dendrogram

Hierarchical Clustering V



(a) Dendrograms

(b) Estimated Clusters

Figure: Hierarchical Clustering to iris data, `hclust()` function.

Gaussian Mixture Model I

- Gaussian mixture distribution:

$$\mathbf{x}_1, \cdots, \mathbf{x}_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

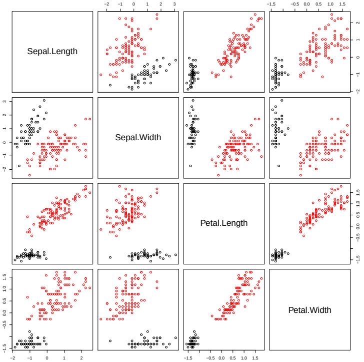
- K: Number of Cluster.
- π_k: Proportion of the kth cluster.
- $f_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: Density of the observations in the kth Cluster

$$N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

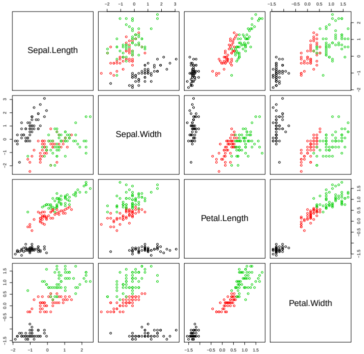
Gaussian Mixture Model II

- MLE is used (EM Algorithm)
- To determine K, model selection criterion can be used.
- Group membership naturally follows after the model parameter estimation.

Gaussian Mixture Model III



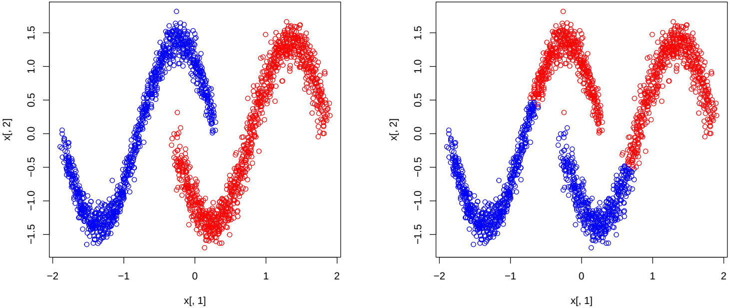
(a) $k = 2$ Selected by BIC



(b) $k = 3$ prespecified

Figure: Gaussian Mixture

DBScan I

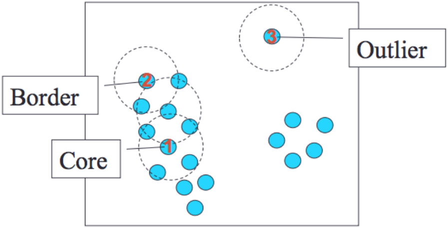


(a) Data (b) *k*-means

Figure: Motivating Example

DBScan II

- Density-Based spatial clustering of applications with noise.
- Definition:
  - Core/High Density Point: at least `MinPt` points are within its ϵ-neighborhood.
  - Border Point: Not a core point, but lies on the ϵ-neighborhood of a core point.
  - Noise Point: neither core nor border point.



MinPts = 5

DBScan III

- $\mathbf{x}_i$ is Directly Density-Reachable (DDR) from $\mathbf{x}_j$.

$\rightarrow$ $\mathbf{x}_j$ is a core point and $\mathbf{x}_i$ is in the $\epsilon-$neighborhood of $\mathbf{x}_j$.

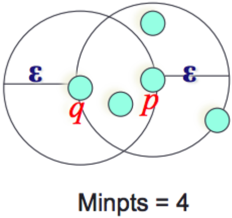


Figure: q is DDR from p, but p is not DDR from q since q is not a core point. That is DDR is asymmetric relation.

DBScan IV

- $\mathbf{x}_i$ is Density-Reachable (DR) from $\mathbf{x}_j$:

$\rightarrow$ There is a sequence of points from $\mathbf{x}_j = p_1, p_2, \cdots, p_n = \mathbf{x}_i$ such that p_{i+l} is directly density-reachable from p_i.

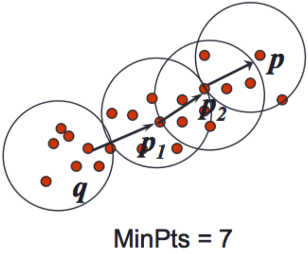


Figure: p is DR from q, but q is not DR from p. That is, DR is also asymmetric.

DBScan V

- $\mathbf{x}_i$ and $\mathbf{x}_j$ are Density-Connected (DC)

$\rightarrow$ here is a point such that both, $\mathbf{x}_i$ and $\mathbf{x}_j$ are density-reachable from the point.

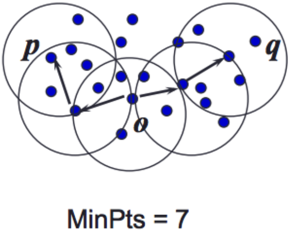


Figure: *p* and *q* are DC and DC is symmetric.

DBScan VI

- Cluster is defined by a set of points C satisfying
  - $\forall \mathbf{x}_i, \mathbf{x}_j$: if $\mathbf{x}_i \in C$ and $\mathbf{x}_j$ is density-reachable from $\mathbf{x}_i$ then $\mathbf{x}_j \in C$. (Maximality)
  - $\forall \mathbf{x}_i, \mathbf{x}_j \in C$, $\mathbf{x}_i$ is density-connected to $\mathbf{x}_j$. (Connectivity)
- `dbscan` package available.

DBScan VII

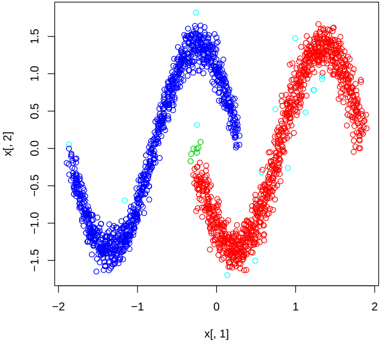


Figure: DBscan applied to the synthetic data.

DBScan VIII

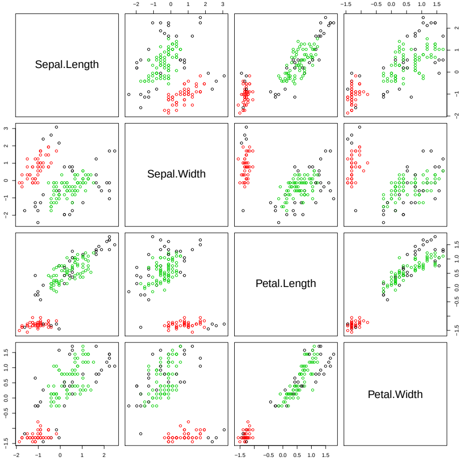


Figure: DBscan applied to Iris data.

MDS I

- Classical MDS (`cmdscale`) solves

$$\underset{\mathbf{z}_1,\cdots,\mathbf{z}_n}{\operatorname{argmin}} \sum_{i \neq j} \Big(\underbrace{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}_{d_{ij}} - \langle \mathbf{z}_i, \mathbf{z}_j \rangle \Big)^2$$

(Turns out to be equivalent to PCA)

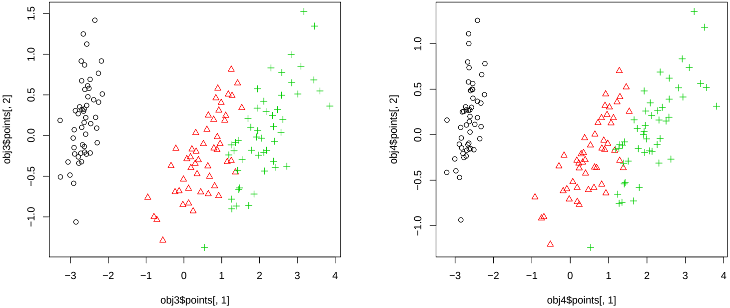
- Sammon mapping (`sammon{MASS}`)

$$\underset{\mathbf{z}_1,\cdots,\mathbf{z}_n}{\operatorname{argmin}} \sum_{i \neq j} \frac{(d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|)^2}{d_{ji}^2}$$

(A weighted version)

- Non-metric scaling: orders (of distances) is used only (`isoMDS{MASS}`)

MDS II



(a) Metric (Sammon) MDS

(b) Nonmetric MDS

Figure: Two versions of MDS ($k = 2$) applied to Iris Data.

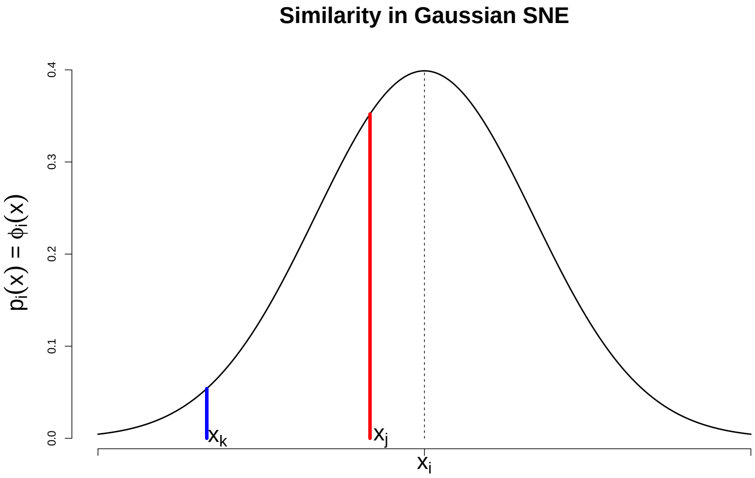
t-SNE I

- DR to 2- or 3-dimensional space. (Visualization)
- A version of Stochastic Neighbor Embedding:
- SNE converts the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities.
- The similarity of $\mathbf{x}_j$ to $\mathbf{x}_i$:

$$p_i(\mathbf{x}_j) = \phi_i(\mathbf{x}_j), \qquad (Conditional Probability)$$

where $\phi_i(\mathbf{x})$ denotes the density of $\mathbf{x} \sim N_p(\mathbf{x}_i, \sigma^2 I)$.

t-SNE II



t-SNE III

- Let $\mathbf{y}_i$ and $\mathbf{y}_j$ denote the low-dim. representations of $\mathbf{x}_i$ and $\mathbf{x}_j$.
- Define $q_i(\mathbf{y}_j)$ similar to $p_i(\mathbf{x}_j)$.
- SNE solves

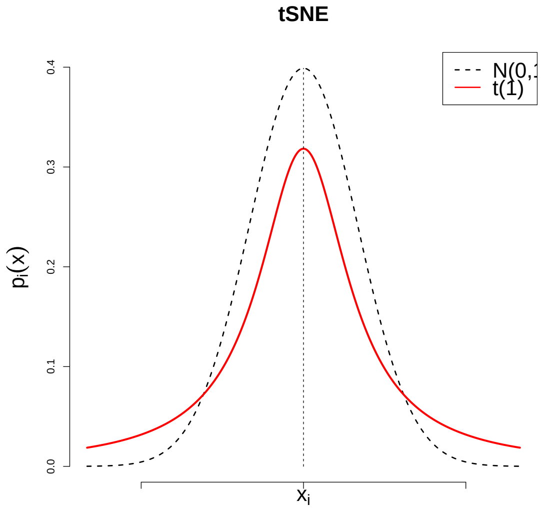
$$\min_{\mathbf{y}_1,\cdots,\mathbf{y}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i(\mathbf{x}_j) \log \frac{p_i(\mathbf{x}_j)}{q_i(\mathbf{y}_j)}, \qquad \text{(KL Divergence)}$$

via the gradient decent algorithm.

t-SNE IV

- As p increases, the pairwise distances between $\mathbf{x}_i$ and $\mathbf{x}_j$ tend to be undistiguishable. (Why?).
- SNE solutions suffer from Crowding Problem.
- Gaussian PDF to measure similarity is not a good choice.
- Let's use heavy tailed distribution $t(1)$! $\rightarrow$ tSNE!

t-SNE V



t-SNE VI

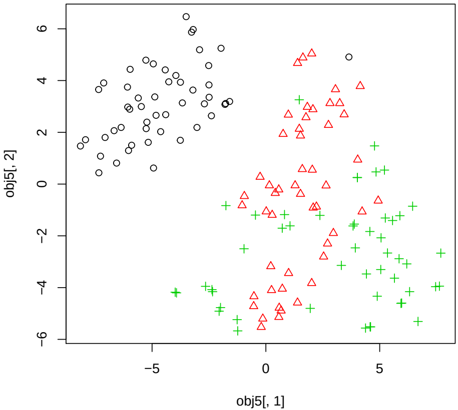


Figure: tSNE applied to Iris Data.