机器学习引论

彭玺
pengxi@scu.edu.cn
www.pengxi.me

四川大学-机器学习引论

提纲

一．Review

二．k-means clustering

三．k-medoids clustering

四．Mixture of Gaussian

四川大学-机器学习引论

提纲

一．Review

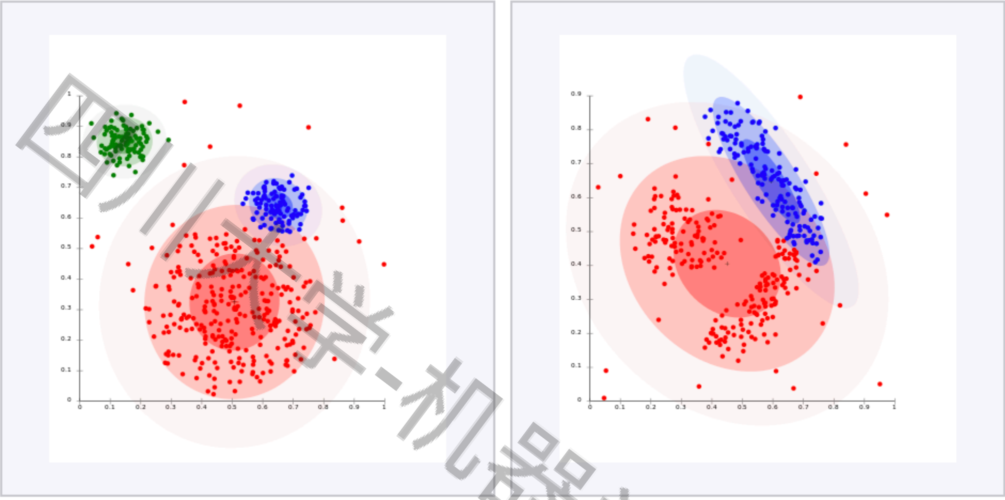
二．k-means clustering

三．k-medoids clustering

四．Mixture of Gaussian

四川大学-机器学习引论

一、Review



Problem Statement :

- Given a set of data points, group them into multiple clusters so that:
  - points within each cluster are similar to each other
  - points from different clusters are dissimilar

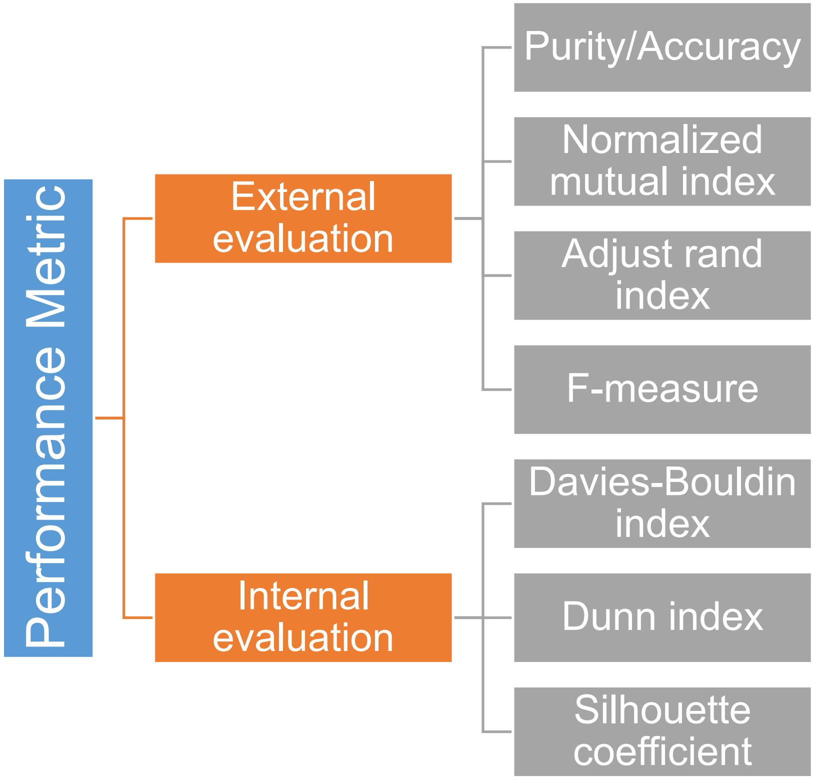
Challenges 1 (key problem of clustering analysis): The major difficulty is that the label is unknown so that the within-/between- class scatter is unavailable.

Challenges 2 (high-dimensional clustering analysis): Usually, points are in a high-dimensional space, and similarity is defined using a distance measure

  - Euclidean, Cosine, Jaccard, edit distance, …

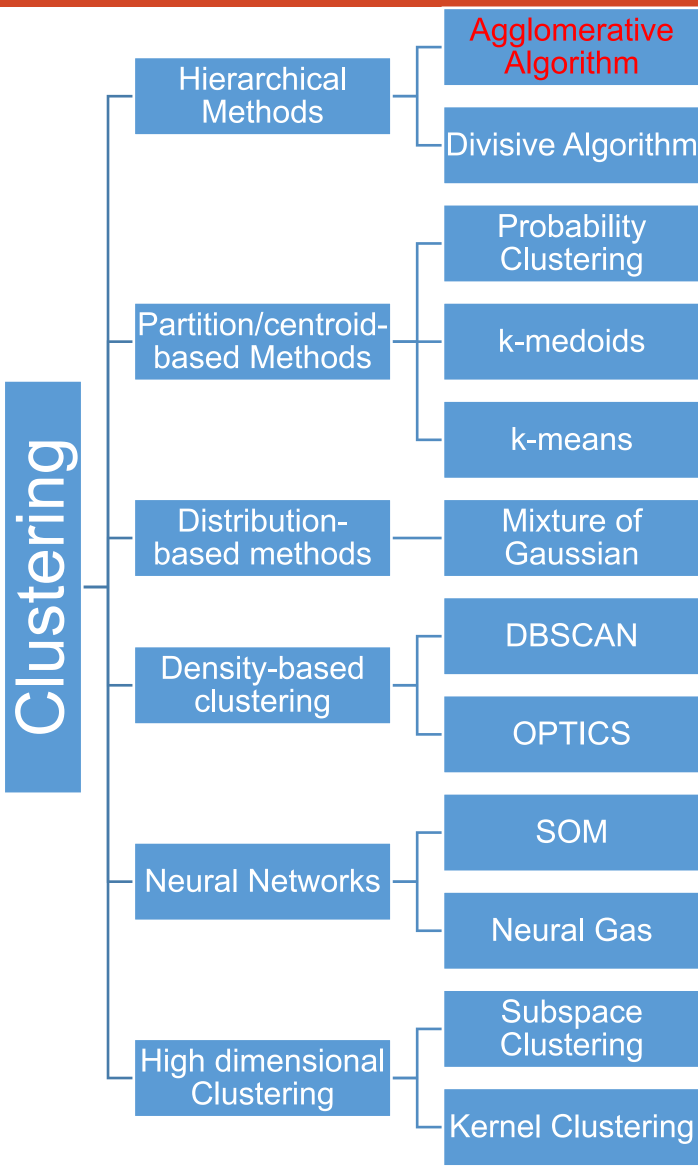
四川大学-机器学习引论

一、Review



四川大学-机器学习引论

一、Review



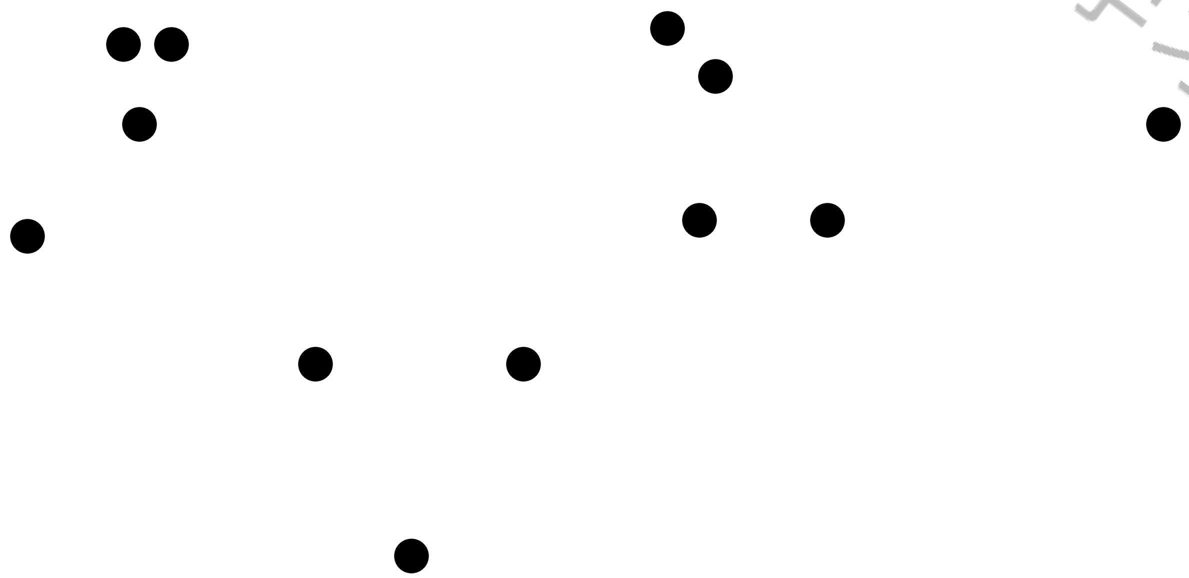
- Agglomerative (Bottom-up)
  - Compute all pair-wise pattern-pattern similarity coefficients
  - Place each of *n* patterns into a class of its own
  - Merge the two most similar clusters into one
    - Replace the two clusters into the new cluster
    - Re-compute inter-cluster similarity scores w.r.t. the new cluster
  - Repeat the above step until there are *k* clusters left (*k* can be 1)

四川大学机器学习引论

一、Review

四川大学-机器学习引论

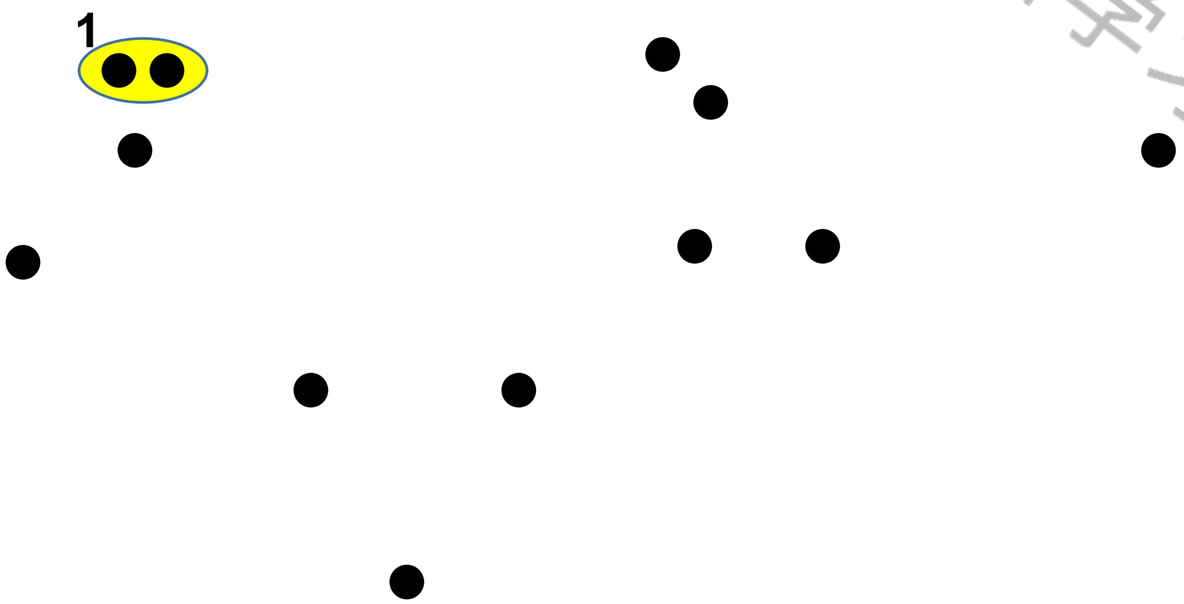
Agglomerative (Bottom up)



一、Review

Agglomerative (Bottom up)

1st iteration



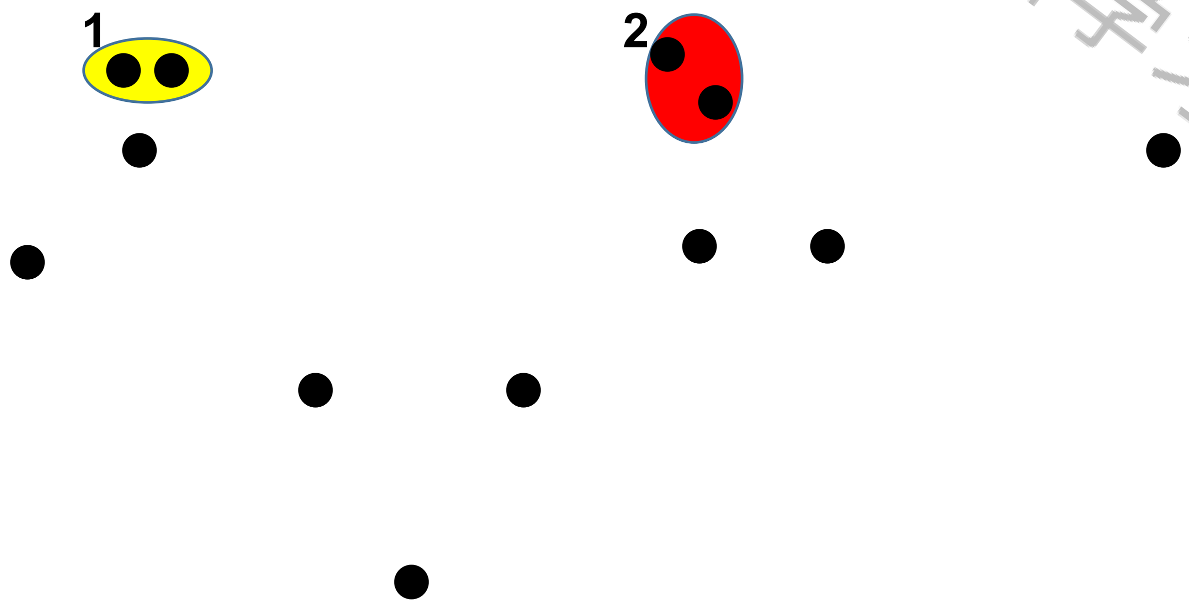
四川大学-机器学习引论

一、Review

四川大学-机器学习引论

Agglomerative (Bottom up)

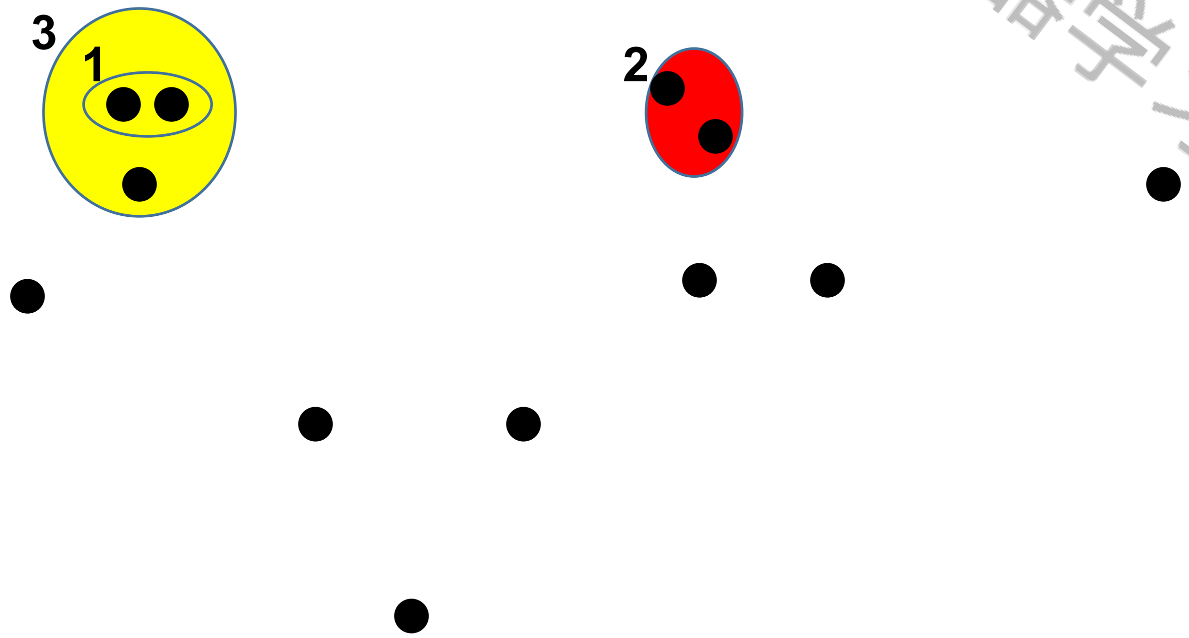
2^{nd} iteration



一、Review

Agglomerative (Bottom up)

3rd iteration

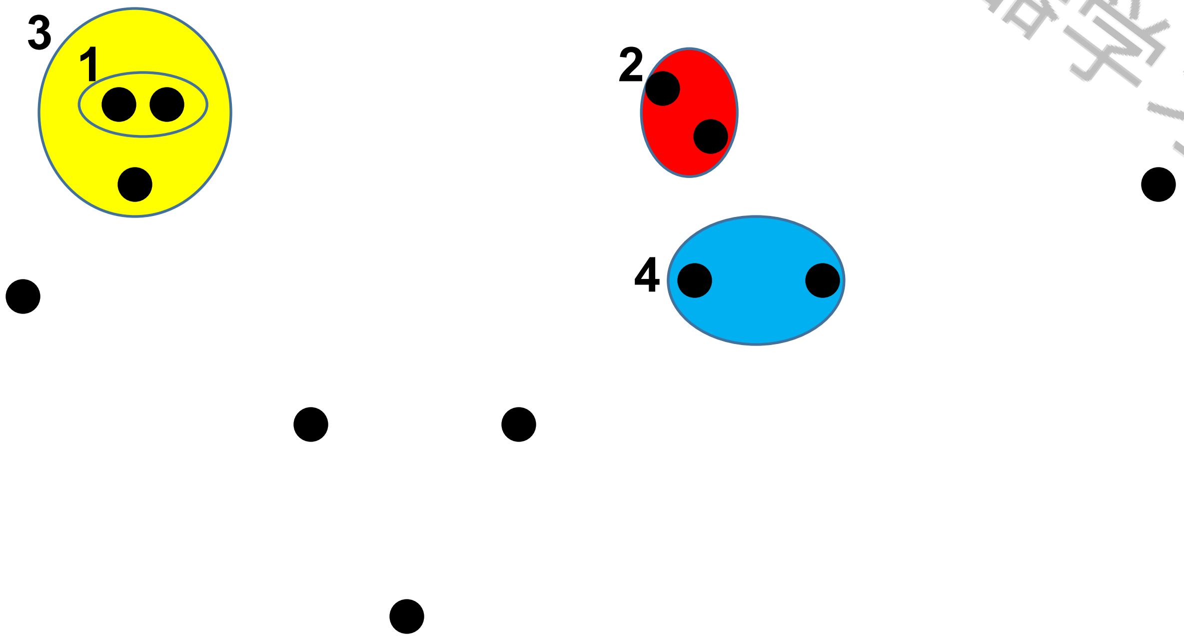


四川大学-机器学习引论

一、Review

Agglomerative (Bottom up)

4th iteration

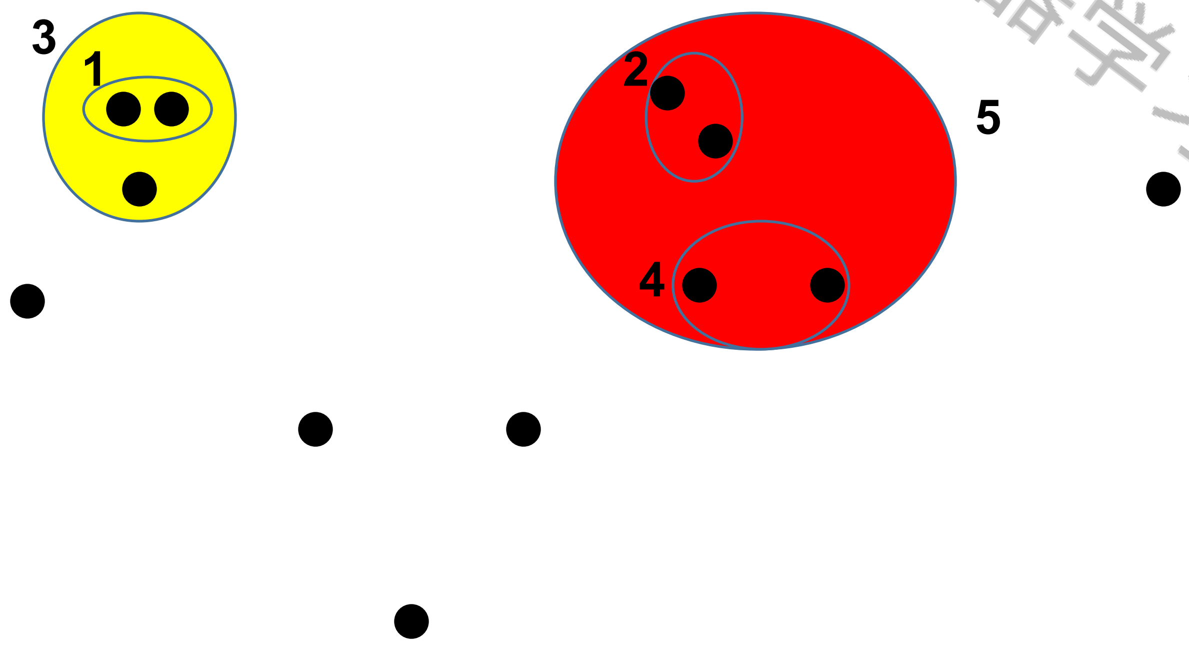


四川大学-机器学习引论

一、Review

Agglomerative (Bottom up)

5th iteration

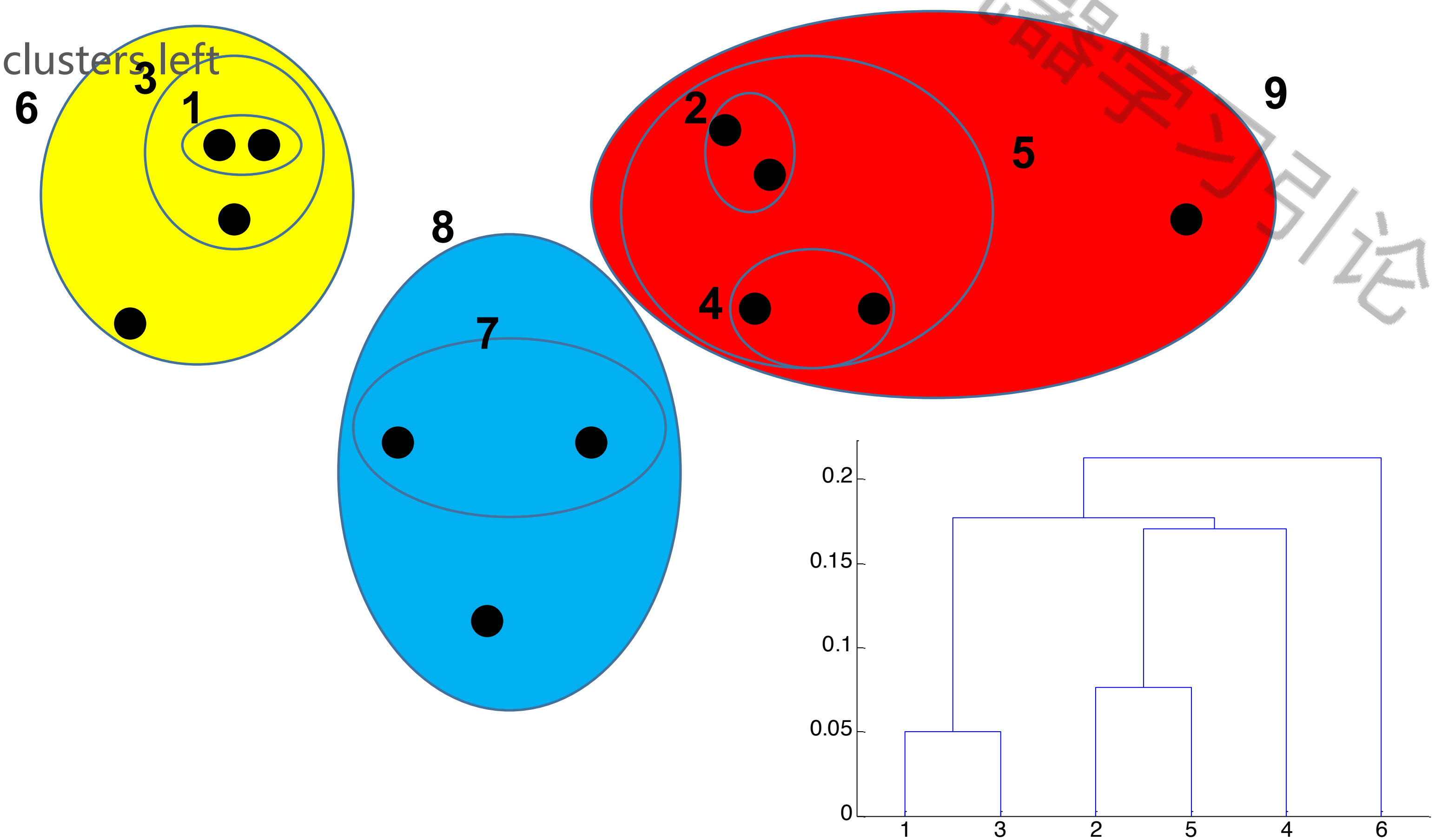


四川大学-机器学习引论

一、Review

Agglomerative (Bottom up)

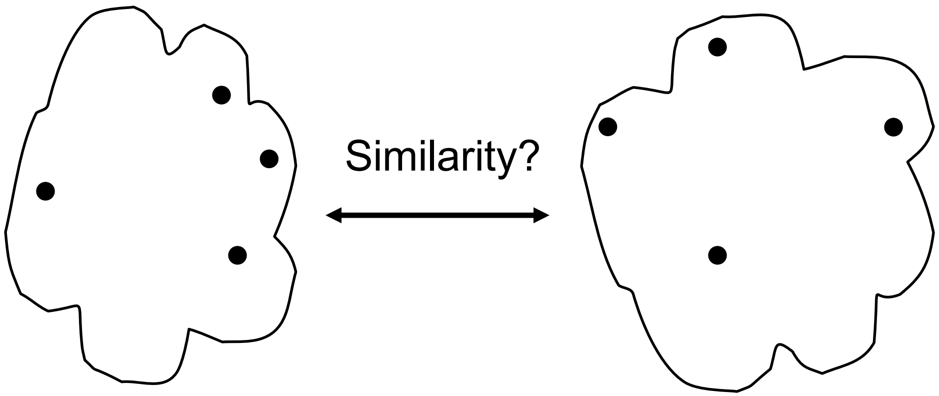
Finally k clusters left



dendrogram

四川大学-机器学习引论

Tip: How to Define Inter-Cluster Similarity



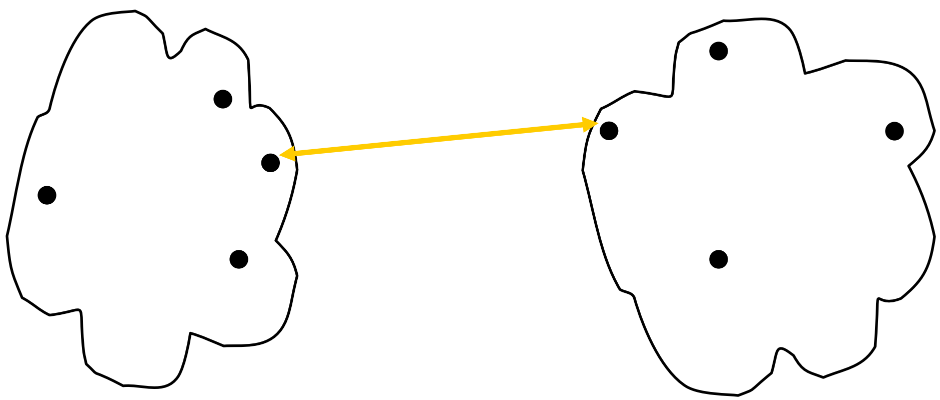
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  – Ward's Method uses squared error

四川大学机器学习引论

Tip: How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

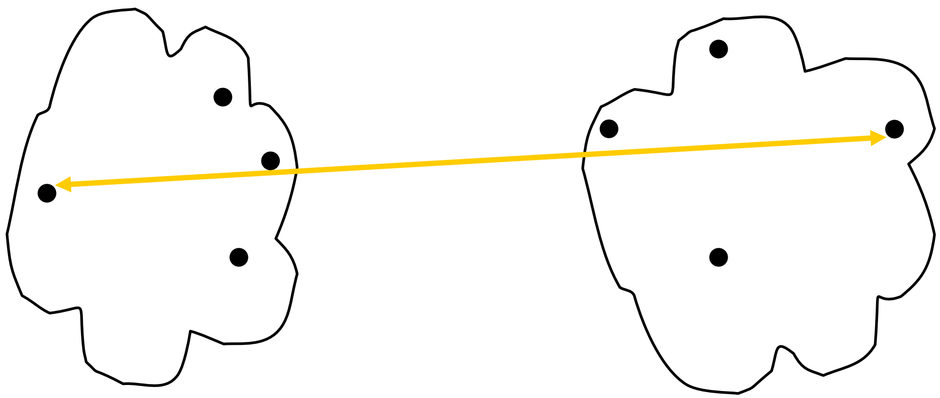
Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

四川大学机器学习引论

Tip: How to Define Inter-Cluster Similarity

四川大学机器学习引论

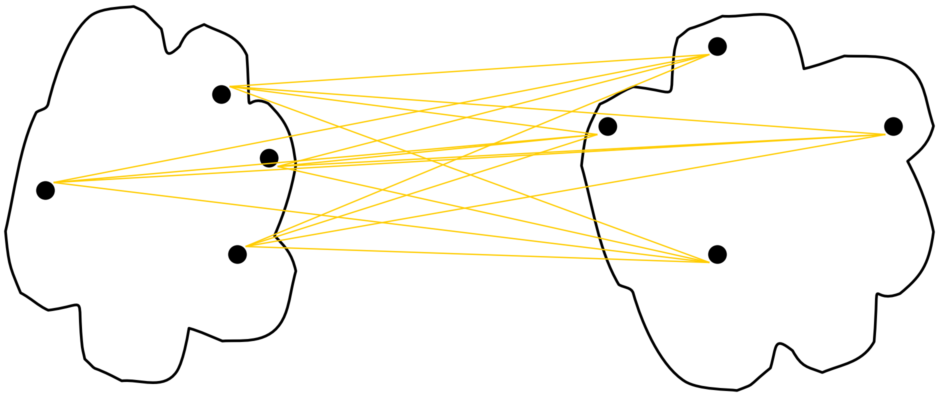


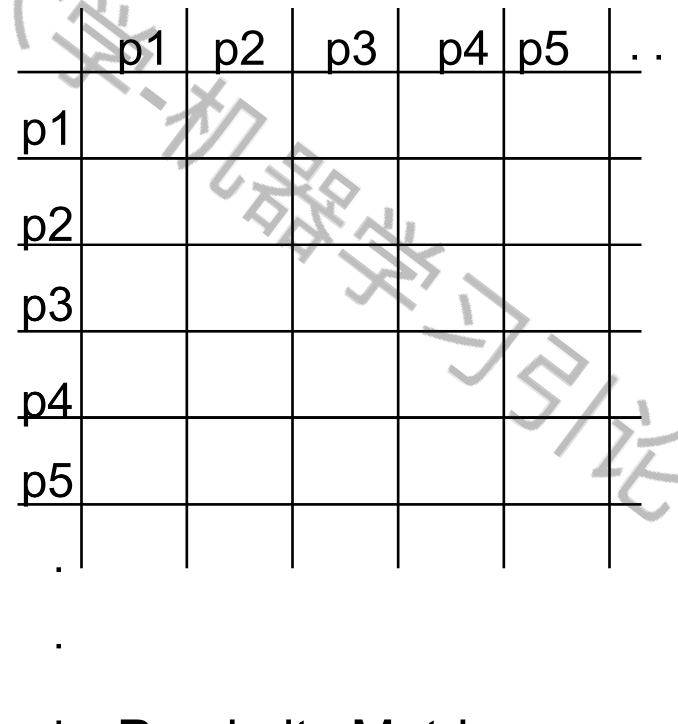
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Tip: How to Define Inter-Cluster Similarity





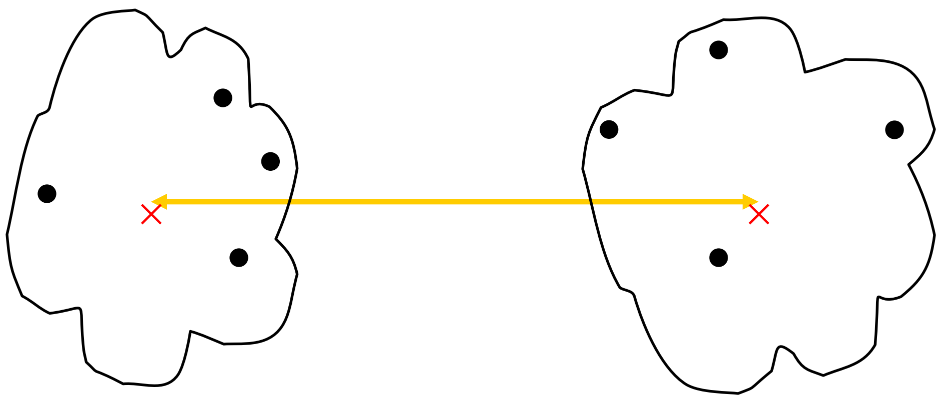
Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

四川大学-机器学习引论

Tip: How to Define Inter-Cluster Similarity

四川大学机器学习引论

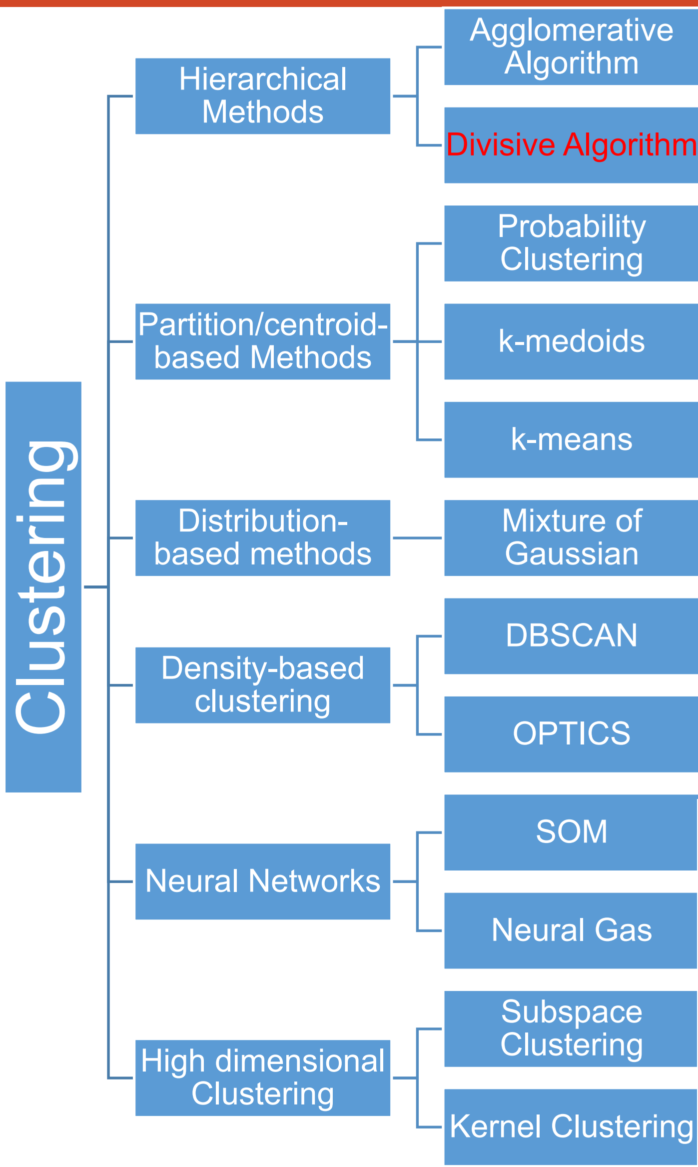


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

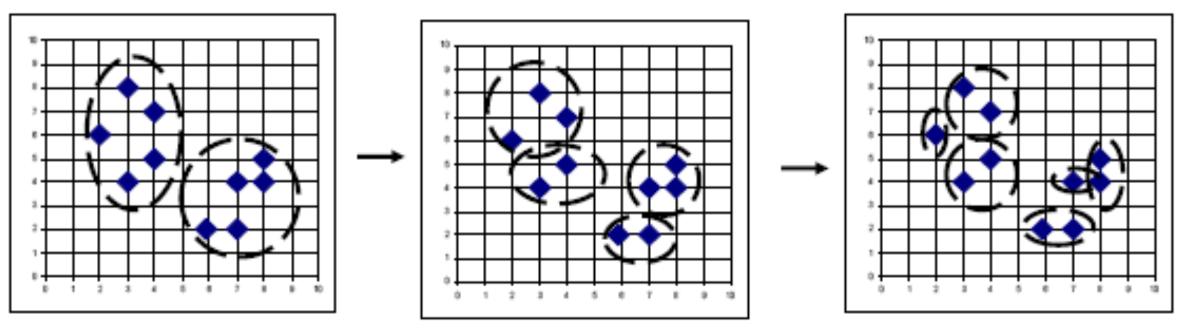
Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

一、Review

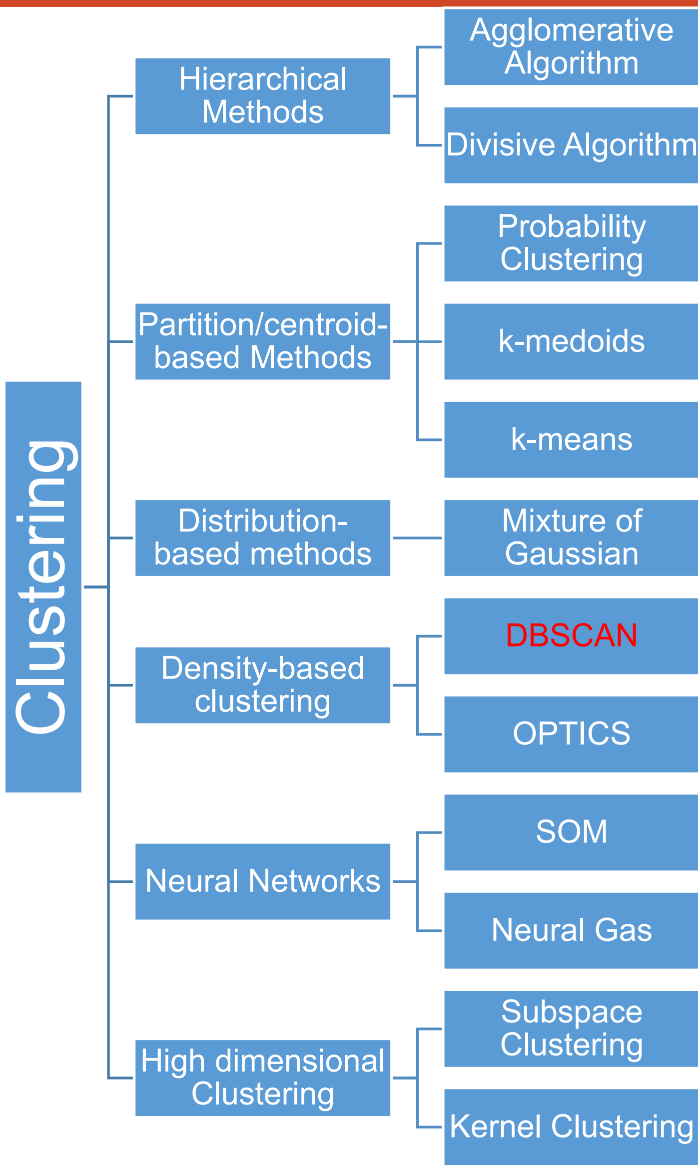


- Divisive (Top-down)
  - Start at the top with all patterns in one cluster
  - The cluster is split using a flat clustering algorithm
  - This procedure is applied recursively until each pattern is in its own singleton cluster



四川大学机器学习引论

一、Review



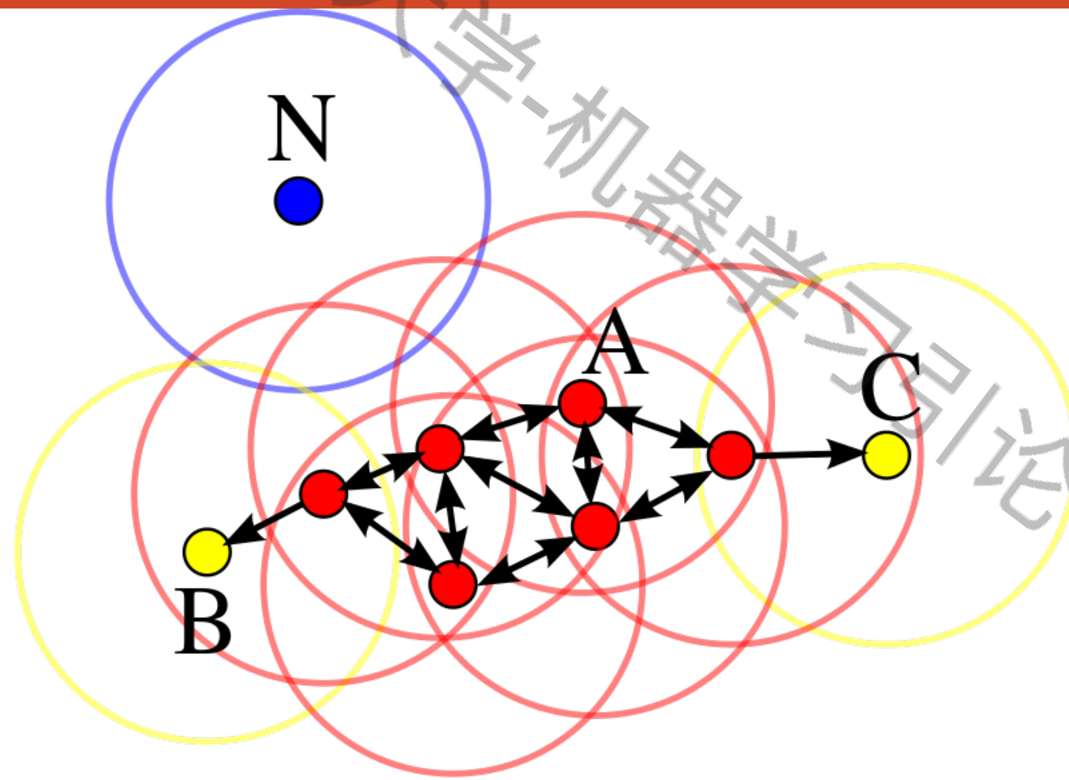
- Important Questions:
  - How do we measure density?
  - What is a dense region?

- Density at point p: number of points within a circle of radius Eps
- Dense Region: A circle of radius Eps that contains at least MinPts points

四川大学机器学习引论

一、Review

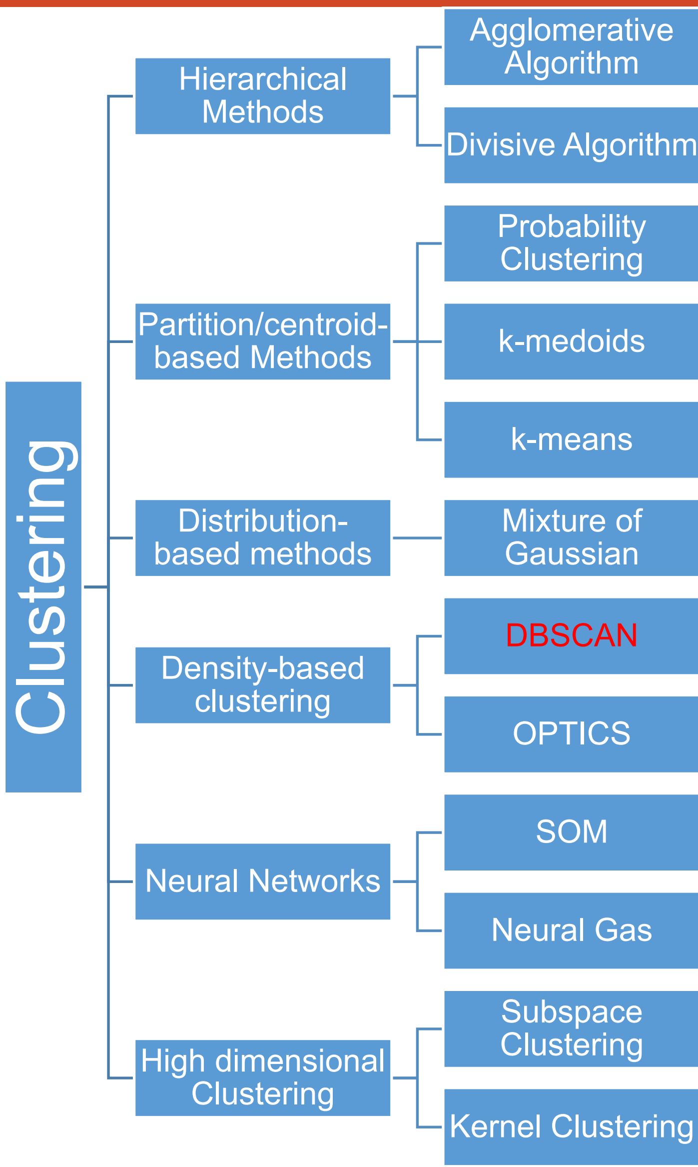
- Clustering
  - Hierarchical Methods
    - Agglomerative Algorithm
    - Divisive Algorithm
  - Partition/centroid-based Methods
    - Probability Clustering
    - k-medoids
    - k-means
  - Distribution-based methods
    - Mixture of Gaussian
  - Density-based clustering
    - DBSCAN
    - OPTICS
  - Neural Networks
    - SOM
    - Neural Gas
  - High dimensional Clustering
    - Subspace Clustering
    - Kernel Clustering



In this diagram, minPts = 4. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

四川大学-机器学习引论

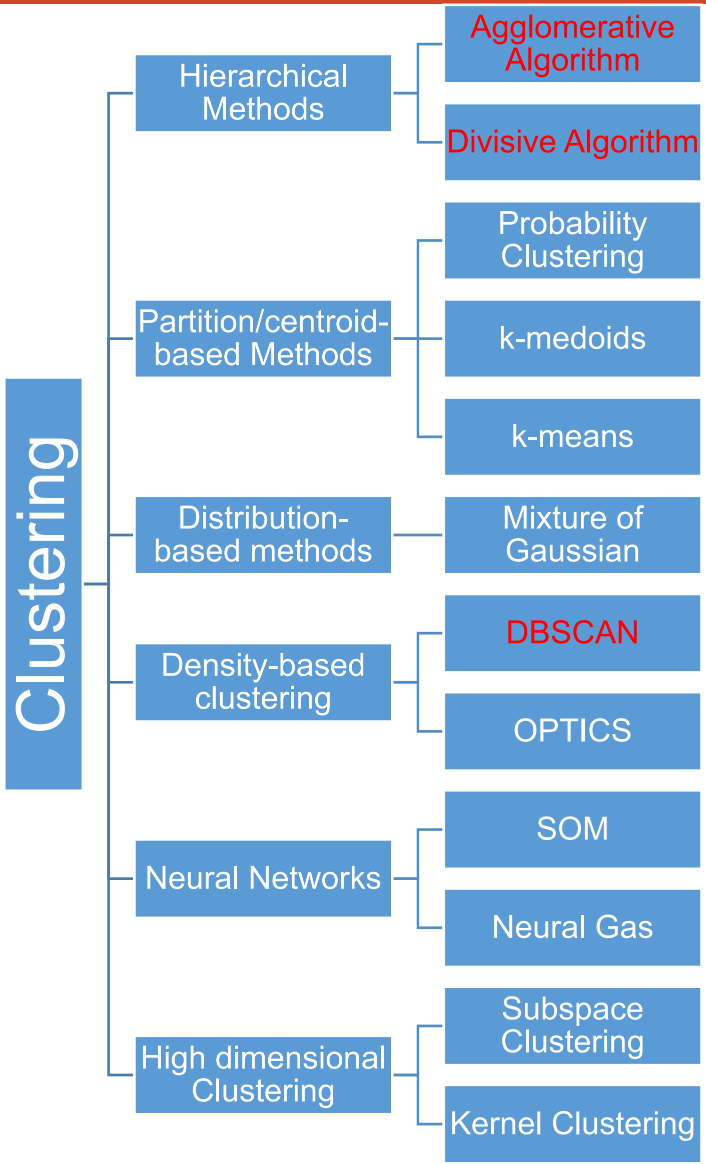
一、Review



- Label points as core, border and noise
- Eliminate noise points
- For every core point p that has not been assigned to a cluster
  - Create a new cluster with the point p and all the points that are density-connected to p.
- Assign border points to the cluster of the closest core point.

四川大学-机器学习引论

一、Review



No objective function is directly minimized, i.e., there are not learning based methods.

四川大学-机器学习引论

提纲

一．Review

二．k-means clustering

三．k-medoids clustering

四．Mixture of Gaussian

四川大学-机器学习引论

二、k-means Clustering

四川大学-机器学习引论

k-means：学习k个means（均值），where each mean corresponds to a cluster center. In other words, k-means achieves clustering by learning/finding k cluster centers.

- Given a set of data points, group them into multiple clusters so that:
  - points within each cluster are similar to each other $\min \sum_{j} \sum_{\mathbf{x}_i \in \mathrm{C}_j} \|\mathbf{x}_i - \mathbf{u}_j\|_2^2$
  - points from different clusters are dissimilar $\max \sum_{i} \sum_{j} \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$

i.e. the cluster assignment (label) and centers are unknown.

二、k-means Clustering

四川大学-机器学习引论

k-means：学习k个means（均值），where each mean corresponds to a cluster center. In other words, k-means achieves clustering by learning/finding k cluster centers.

- Given a set of data points, group them into multiple clusters so that:
  - points within each cluster are similar to each other $\min \sum_j \sum_{\mathbf{x}_i \in \mathrm{C}_j} \|\mathbf{x}_i - \mathbf{u}_j\|_2^2$
  - points from different clusters are dissimilar $\max \sum_i \sum_j \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$

Solution：Iteratively learning clustering assignment and cluster centers so that

二、k-means Clustering

四川大学-机器学习引论

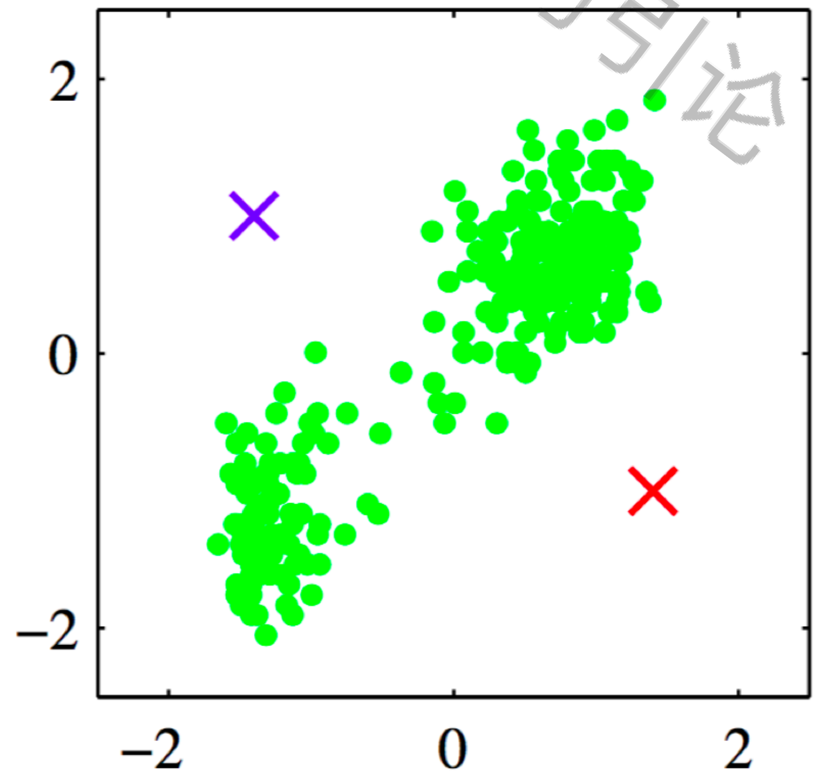
k-means：k个means（均值），clearly, each mean corresponds to a cluster center. In other words, k-means achieves clustering by learning/finding k cluster centers.

Solution: Iteratively learning clustering assignment and cluster centers so that

$$\min \sum_{j} \sum_{\mathbf{x}_i \in \mathrm{C}_j} \|\mathbf{x}_i - \mathbf{u}_j\|_2^2 \quad \max \sum_{i} \sum_{j} \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$$

Example:

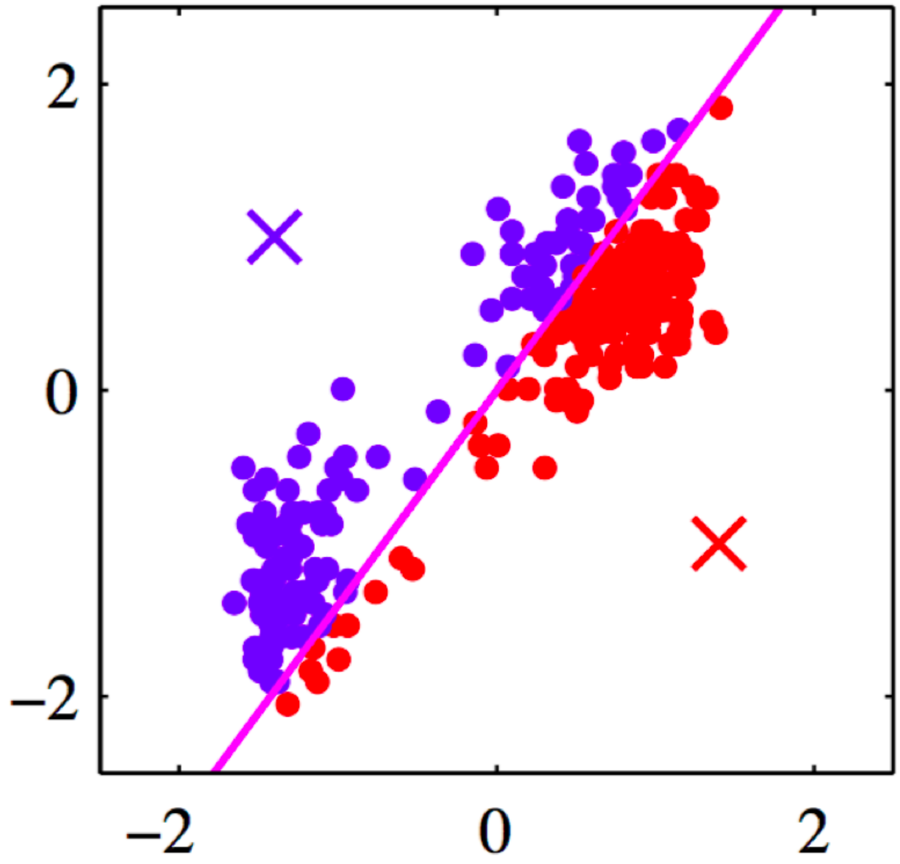
Iter1: randomly choose two points as cluster centers



二、k-means Clustering

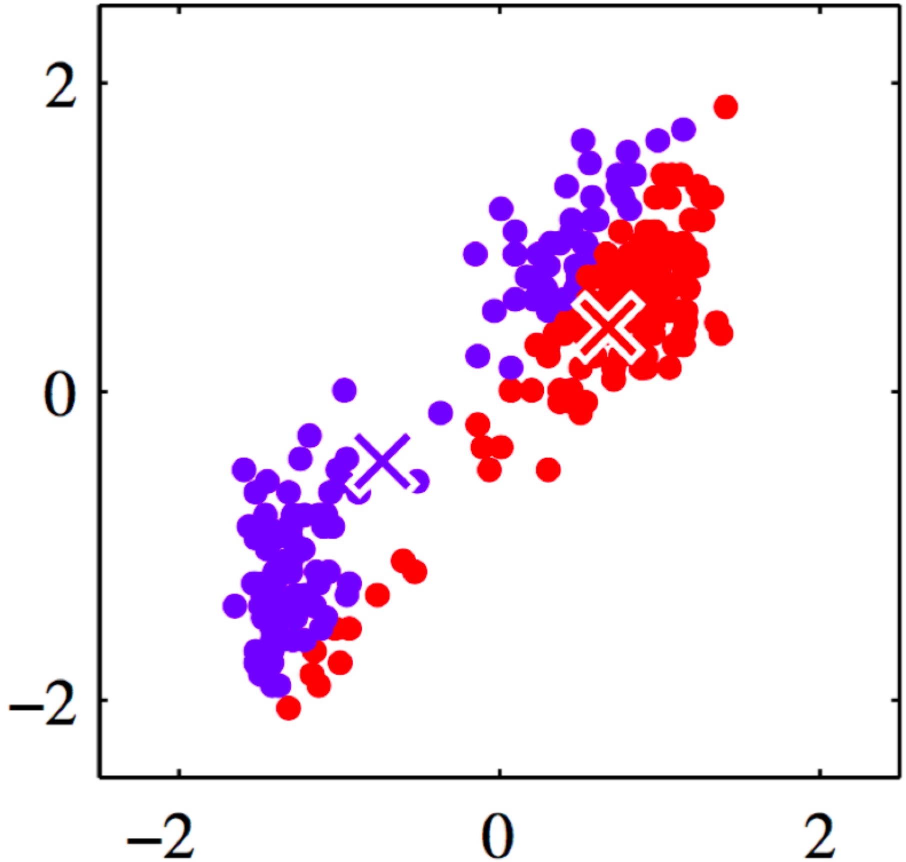
四川大学-机器学习引论

Iter1: Assign each point to closest center.



二、k-means Clustering

Iter2: Compute new class centers by $\mu_j = \frac{\sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i}{|\mathcal{C}_j|}$

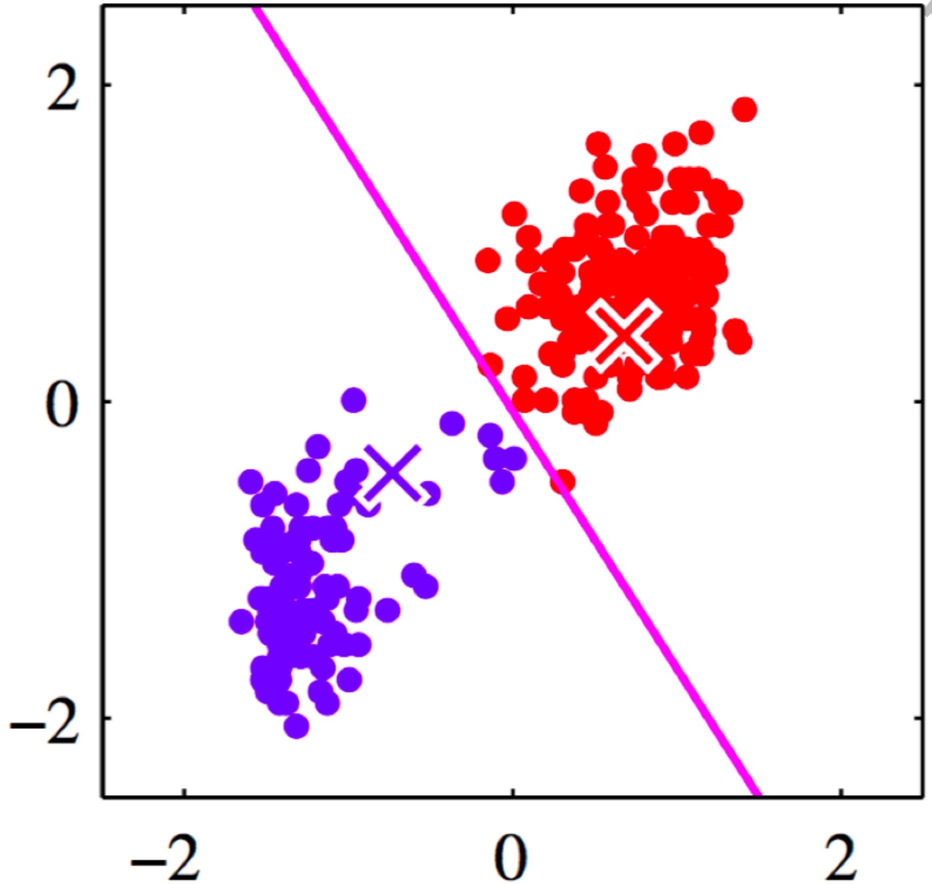


四川大学机器学习引论

二、k-means Clustering

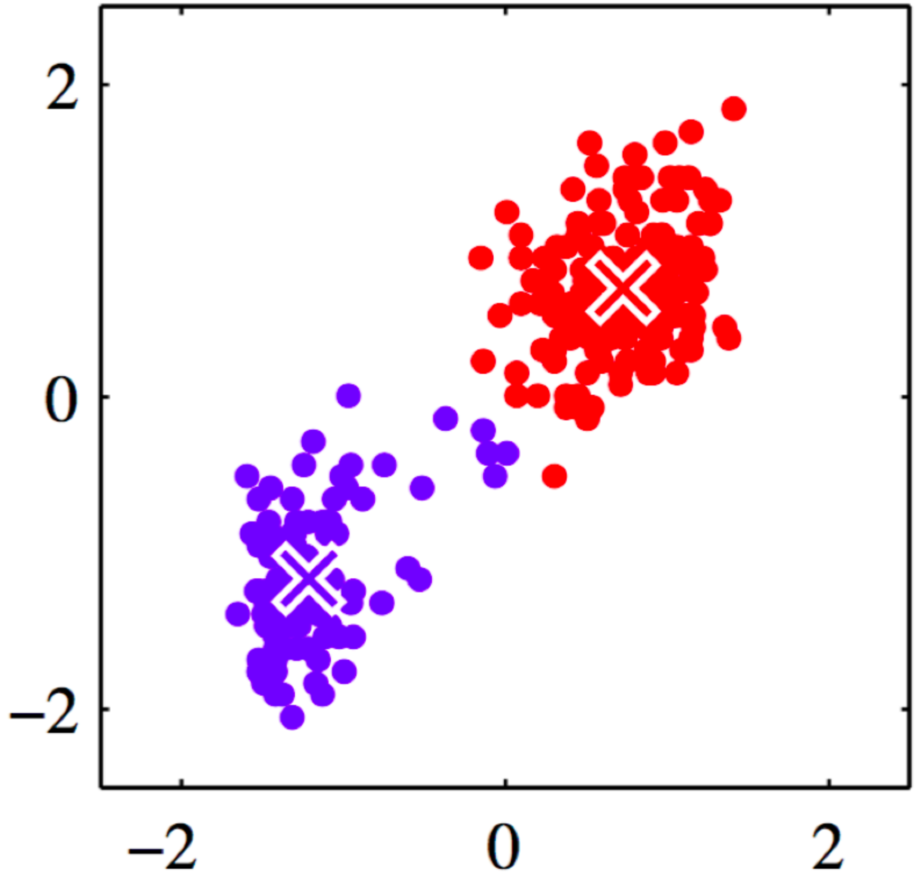
四川大学-机器学习引论

Iter2: Assign points to closest center.



二、k-means Clustering

Iter3: Compute cluster centers by $\mu_j = \frac{\sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i}{|\mathcal{C}_j|}$

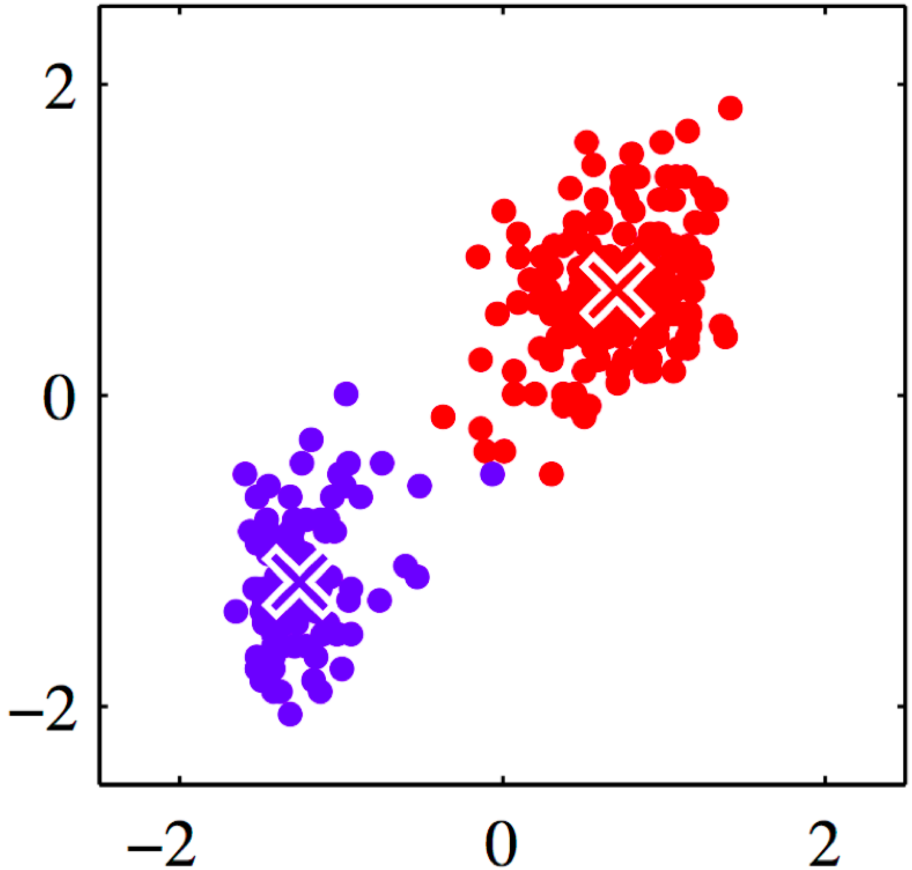


四川大学-机器学习引论

二、k-means Clustering

四川大学-机器学习引论

Iterate until convergence (reach to the max iteration number or the loss is smaller than a given threshold).



二、k-means Clustering

四川大学-机器学习引论

- Dataset $\mathcal{D} = \{x_1, \ldots, x_n\} \in \mathbf{R}^d$
- Goal (version 1): Partition data into k clusters.
- Goal (version 2): Partition $\mathbf{R}^d$ into k regions.
- Let $\mu_1, \ldots, \mu_k$ denote cluster centers.

二、k-means Clustering

四川大学-机器学习引论

- Dataset $\mathcal{D} = \{x_1, \ldots, x_n\} \in \mathbf{R}^d$
- Goal (version 1): Partition data into k clusters.
- Goal (version 2): Partition $\mathbf{R}^d$ into k regions.
- Let $\mu_1, \ldots, \mu_k$ denote cluster centers.
- For each x_i, use a **one-hot encoding** to designate membership:

$$r_i = (0, 0, \ldots, 0, 0, 1, 0, 0) \in \mathbf{R}^k$$

- Let

$$r_{ic} = 1(x_i \text{ assigned to cluster } c).$$

- Then

$$r_i = (r_{i1}, r_{i2}, \ldots, r_{ik}).$$

二、k-means Clustering

- Find cluster centers and cluster assignments minimizing

$$J(r, \mu) = \sum_{i=1}^{n} \sum_{c=1}^{k} r_{ic} \|x_i - \mu_c\|^2.$$

四川大学机器学习引论

二、k-means Clustering

四川大学-机器学习引论

- Find cluster centers and cluster assignments minimizing

$$J(r, \mu) = \sum_{i=1}^{n} \sum_{c=1}^{k} r_{ic} \|x_i - \mu_c\|^2.$$

- Is objective function convex?
- What's the domain of J?

二、k-means Clustering

四川大学-机器学习引论

- Find cluster centers and cluster assignments minimizing

$$J(r, \mu) = \sum_{i=1}^{n} \sum_{c=1}^{k} r_{ic} \|x_i - \mu_c\|^2.$$

- Is objective function convex?
- What's the domain of J?
- $r \in \{0, 1\}^{n \times k}$, which is not a convex set...
- So domain of J is not convex $\implies$ J is not a convex function
- We should expect local minima.

二、k-means Clustering

四川大学机器学习引论

- For fixed r (cluster assignments), minimizing over μ is easy:

$$
\begin{aligned}
J(r,\mu) &= \sum_{i=1}^{n}\sum_{c=1}^{k} r_{ic}\|x_i-\mu_c\|^2 \\
&= \sum_{c=1}^{k}\underbrace{\sum_{i=1}^{n} r_{ic}\|x_i-\mu_c\|^2}_{=J_c} \\
J_c(\mu_c) &= \sum_{\{i|x_i \text{belongs to cluster } c\}} \|x_i-\mu_c\|^2
\end{aligned}
$$

- J_c is minimized by

$$\mu_c = \text{mean}\left(\{x_i \mid x_i \text{ belongs to cluster } c\}\right)$$

二、k-means Clustering

四川大学-机器学习引论

- For fixed μ (cluster centers), minimizing over r is easy:

$$J(r, \mu) = \sum_{i=1}^{n} \sum_{c=1}^{k} r_{ic} \|x_i - \mu_c\|^2$$

- For each i, exactly one of the following terms is nonzero:

$$r_{i1}\|x_i - \mu_1\|^2, r_{i2}\|x_i - \mu_2\|^2, \ldots, r_{ik}\|x_i - \mu_k\|^2$$

- Take

$$r_{ic} = 1(c = \arg\min_j \|x_i - \mu_j\|^2)$$

- That is, assign x_i to cluster c with minimum distance

$$\|x_i - \mu_c\|^2$$

二、k-means Clustering

四川大学-机器学习引论

- We will use an **alternating minimization** algorithm:
  1. Choose initial cluster centers $\mu = (\mu_1, \ldots, \mu_k)$.
     - e.g. choose k randomly chosen data points
  2. Repeat
     1. For given cluster centers, find optimal cluster assignments:

$$r_{ic}^{\text{new}} = 1(c = \arg\min_j \|x_i - \mu_j\|^2)$$

     2. Given cluster assignments, find optimal cluster centers:

$$\mu_c^{\text{new}} = \arg\min_{m \in \mathbf{R}^d}; \sum_{\{i \mid r_{ic}=1\}} \|x_i - \mu_c\|^2$$

二、k-means Clustering

四川大学-机器学习引论

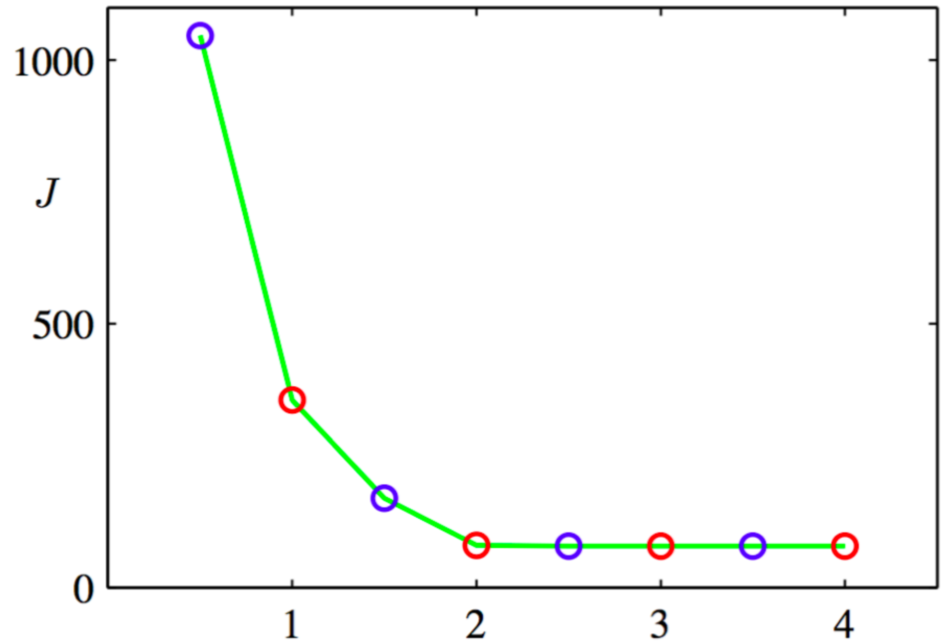
Convergence:

- Note: Objective value never increases in an update.
  - (Obvious: worst case, everything stays the same)
- Consider the sequence of objective values: $J_1, J_2, J_3, \ldots$
  - monotonically decreasing
  - bounded below by zero
- Therefore, k-**Means objective value converges** to $\inf_t J_t$.
- **Reminder**: This is convergence to a **local** minimum.
- Best to repeat k-means several times, with different starting points

二、k-means Clustering

四川大学-机器学习引论

- Blue circles after **"E" step**: assigning each point to a cluster
- Red circles after **"M" step**: recomputing the cluster centers



二、k-means Clustering

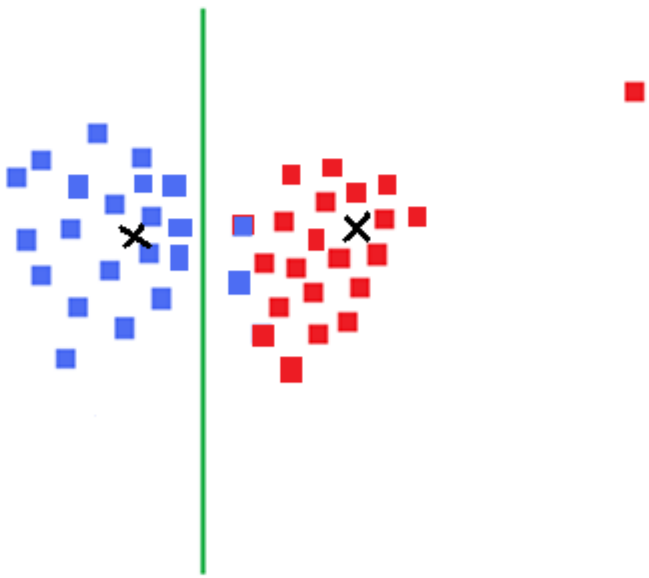
四川大学-机器学习引论

- Disadvantages
  - Dependent on initialization
    - Select random seeds with at least $D_{\min}$
    - Or, run the algorithm many times

二、k-means Clustering

四川大学-机器学习引论

- Disadvantages
  - Dependent on initialization
  - Sensitive to outliers



提纲

一．Review

二．k-means clustering

三．k-medoids clustering

四．Mixture of Gaussian

四川大学-机器学习引论

三、k-medoids clustering

四川大学-机器学习引论

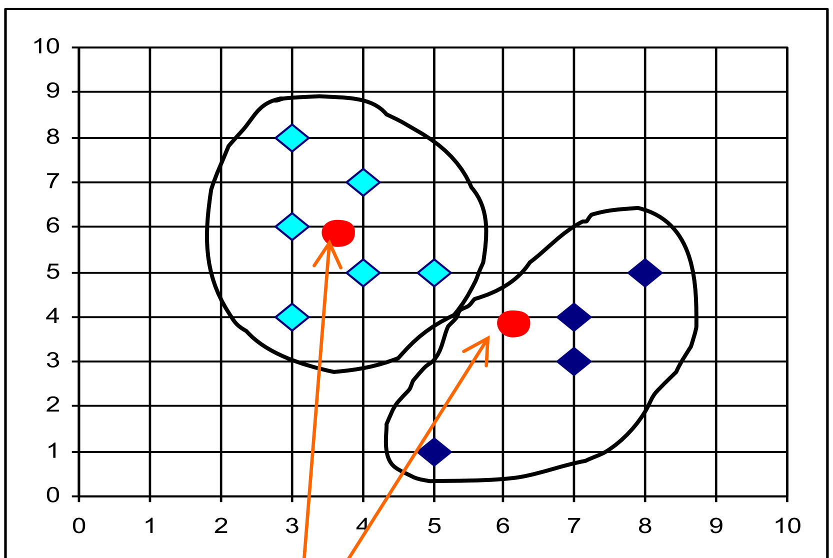
- *k*-means (MacQueen'67): Each cluster is represented by center of cluster
  - Sensitive to noise/outlier

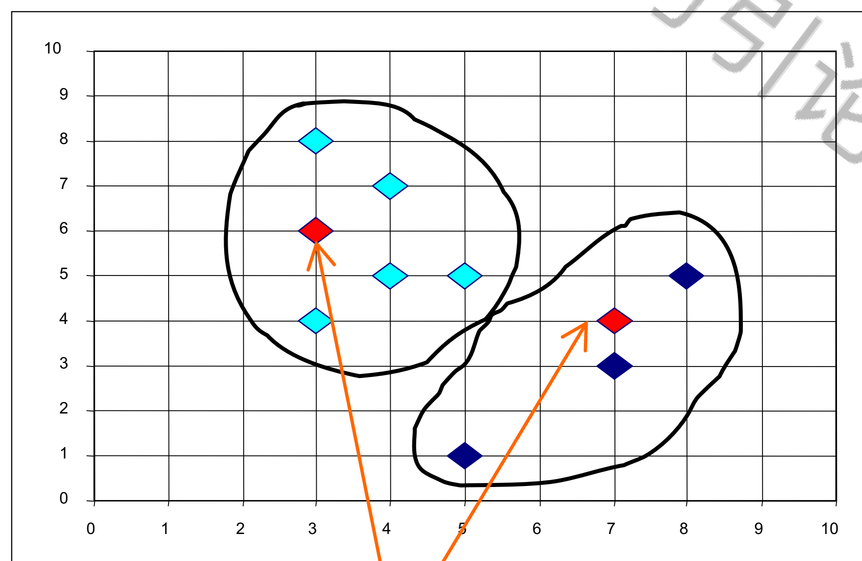
- *k*-medoids (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects (medoid) in cluster
  - Robust to noise/outlier
  - keep the physical meaning of the dataset
  - Higher computational cost than *k*-means

三、k-medoids clustering

四川大学-机器学习引论

- ***k*-medoids: Find *k representative objects*, called *medoids***





k-means

质心

k-medoids

中值，需要遍历整个数据集以得到medoid

提纲

一 . Review

二 . k-means clustering

三 . k-medoids clustering

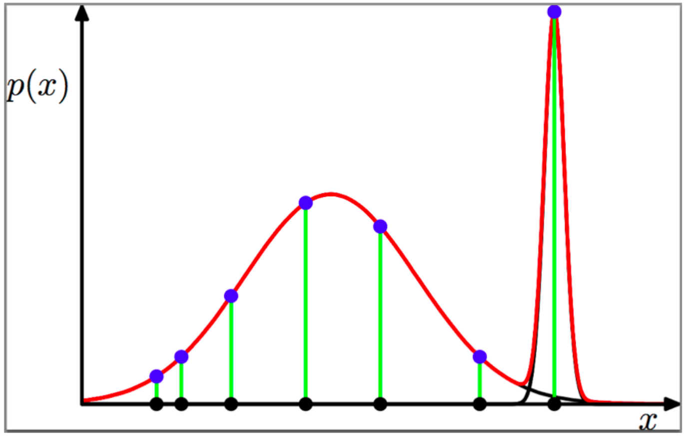
四 . Mixture of Gaussian

四川大学-机器学习引论

四、Mixture of Gaussian

四川大学-机器学习引论

Universal Approximation: any distribution could be represented by a MOG, namely, any data set is a MOG and each cluster corresponds to a Gaussian distribution.



1-dimensional

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{2\pi\sigma^2}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu))$$

四、Mixture of Gaussian

四川大学-机器学习引论

Definition

A probability density $p(x)$ represents a **mixture distribution** or **mixture model**, if we can write it as a **convex combination** of probability densities. That is,

$$p(x) = \sum_{i=1}^{k} w_i p_i(x),$$

where $w_i \geqslant 0$, $\sum_{i=1}^{k} w_i = 1$, and each p_i is a probability density.

- In our Gaussian mixture model, X has a **mixture distribution**.
- More constructively, let S be a set of probability distributions:
  1. Choose a distribution randomly from S.
  2. Sample X from the chosen distribution.
- Then X has a mixture distribution.

四、Mixture of Gaussian

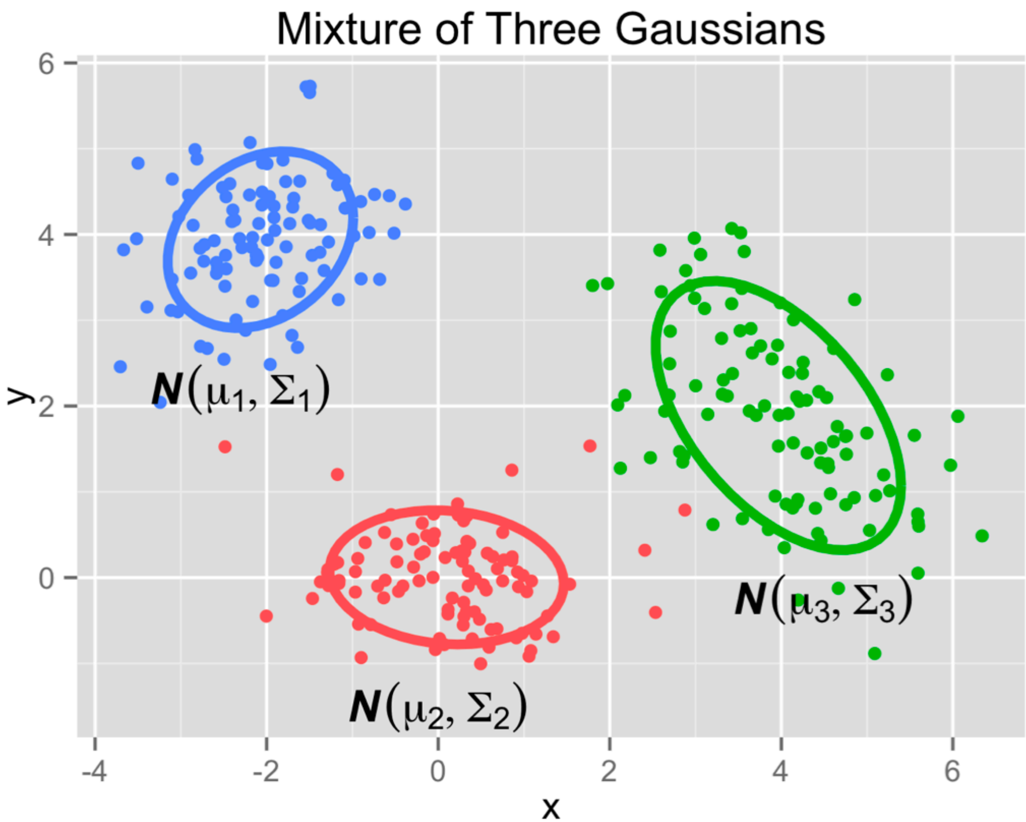
四川大学机器学习引论

- Let's consider a **generative model** for the data.
- Suppose
  1. There are k clusters.
  2. We have a probability density for each cluster.
- Generate a point as follows
  1. Choose a random cluster $z \in \{1, 2, \ldots, k\}$.
     - $Z \sim \text{Multi}(\pi_1, \ldots, \pi_k)$.
  2. Choose a point from the distribution for cluster Z.
     - $X \mid Z = z \sim p(x \mid z)$.

四、Mixture of Gaussian

四川大学-机器学习引论

1. Choose $Z \in \{1,2,3\} \sim \text{Multi}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.
2. Choose $X \mid Z = z \sim \mathcal{N}\left(X \mid \mu_z, \Sigma_z\right)$.



四、Mixture of Gaussian

$$\begin{aligned}\text{Cluster probabilities:} \quad & \pi = (\pi_1, \ldots, \pi_k) \\ \text{Cluster means:} \quad & \mu = (\mu_1, \ldots, \mu_k) \\ \text{Cluster covariance matrices:} \quad & \Sigma = (\Sigma_1, \ldots \Sigma_k)\end{aligned}$$

- The model likelihood for $\mathcal{D} = \{x_1, \ldots, x_n\}$ is

Since we only observe X, we need the marginal distribution

$$\boxed{\begin{aligned} p(x) &= \sum_{z=1}^{k} p(x, z) \\ &= \sum_{z=1}^{k} \pi_z \mathcal{N}(x \mid \mu_z, \Sigma_z) \end{aligned}}$$

$$\begin{aligned} L(\pi, \mu, \Sigma) &= \prod_{i=1}^{n} p(x_i) \\ &= \prod_{i=1}^{n} \sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z). \end{aligned}$$

- As usual, we'll take our objective function to be the log of this:

$$J(\pi, \mu, \Sigma) = \sum_{i=1}^{n} \log \left\{ \sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z) \right\}$$

四川大学-机器学习引论

$$\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu))$$

四、Mixture of Gaussian

四川大学机器学习引论

- Let's start by considering the MLE for the Gaussian model.
- For data $\mathcal{D} = \{x_1, \ldots, x_n\}$, the log likelihood is given by

$$\sum_{i=1}^{n} \log \mathcal{N}(x_i \mid \mu, \Sigma) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu).$$

- With some calculus, we find that the MLE parameters are

$$\begin{aligned}\mu_{\mathrm{MLE}} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \Sigma_{\mathrm{MLE}} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_{\mathrm{MLE}})(x_i-\mu_{\mathrm{MLE}})^T\end{aligned}$$

- For GMM, If we knew the cluster assignment z_i for each x_i,
  - we could compute the MLEs for each cluster.

四、Mixture of Gaussian

四川大学-机器学习引论

- Denote the probability that observed value x_i comes from cluster j by

$$\gamma_i^j = \mathbb{P}(Z = j \mid X = x_i).$$

- The **responsibility** that cluster j takes for observation x_i.
- Computationally,

$$\begin{aligned}
\gamma_i^j &= \mathbb{P}(Z = j \mid X = x_i). \\
&= p(Z = j, X = x_i) / p(x) \\
&= \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)}
\end{aligned}$$

- The vector $(\gamma_i^1, \ldots, \gamma_i^k)$ is exactly the **soft assignment** for x_i.
- Let $n_c = \sum_{i=1}^{n} \gamma_i^c$ be the number of points "soft assigned" to cluster c.

四、Mixture of Gaussian

四川大学-机器学习引论

1. Initialize parameters μ, Σ, π.
2. "E step". Evaluate the responsibilities using current parameters:

$$\gamma_i^j = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)},$$

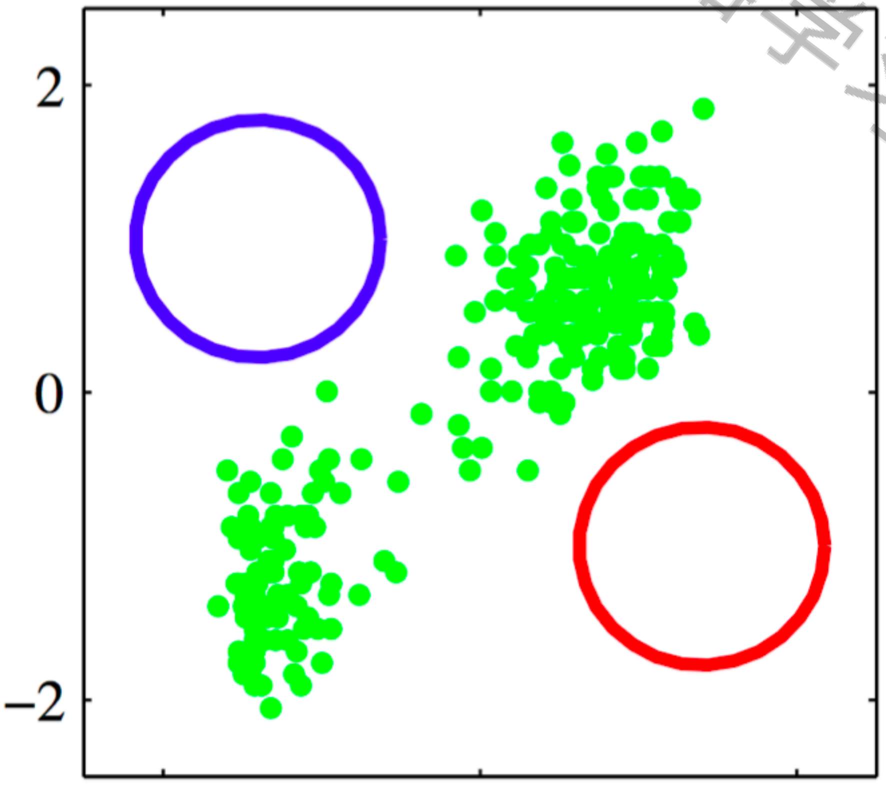
   for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.
3. "M step". Re-estimate the parameters using responsibilities:

$$\begin{aligned}
\mu_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c x_i \\
\Sigma_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c (x_i - \mu_{\text{MLE}})(x_i - \mu_{\text{MLE}})^T \\
\pi_c^{\text{new}} &= \frac{n_c}{n},
\end{aligned}$$

4. Repeat from Step 2, until log-likelihood converges.

四、Mixture of Gaussian

- Initialization

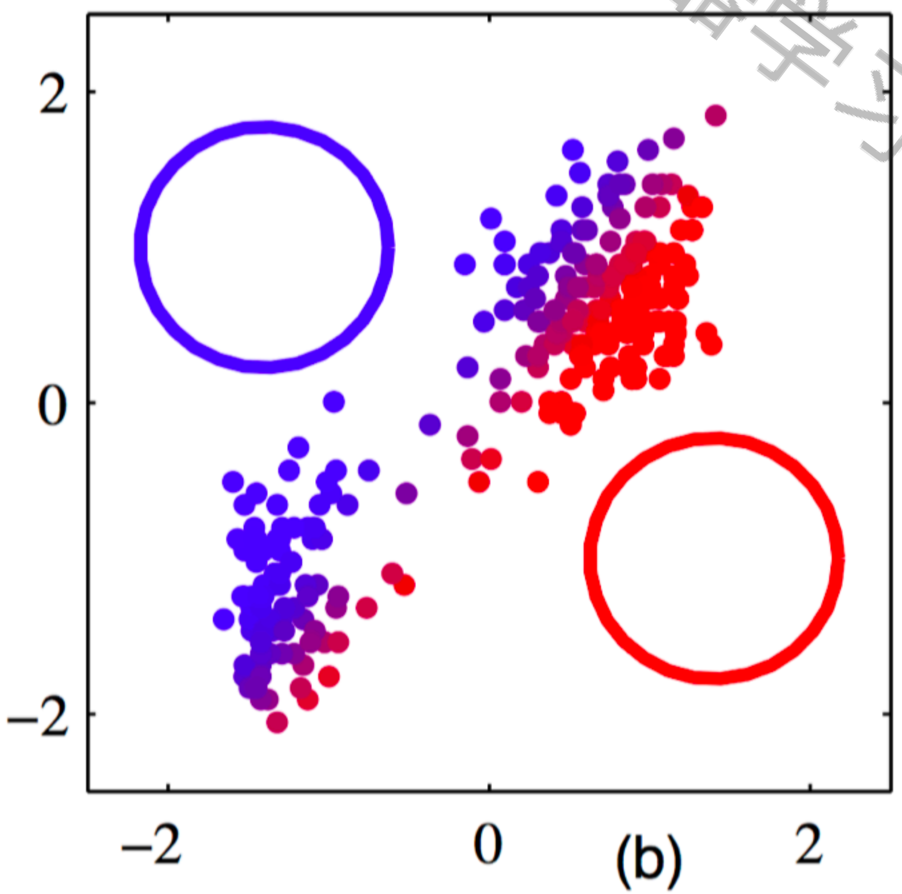


(a)

四川大学-机器学习引论

四、Mixture of Gaussian

- First soft assignment:

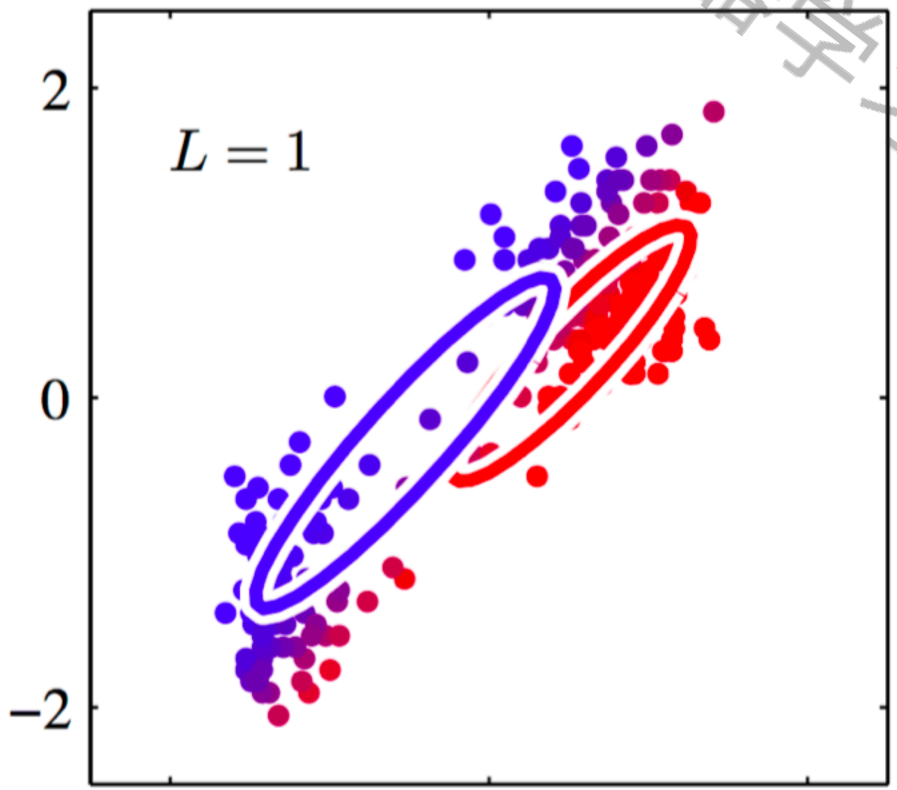


四川大学-机器学习引论

四、Mixture of Gaussian

四川大学-机器学习引论

- First soft assignment:

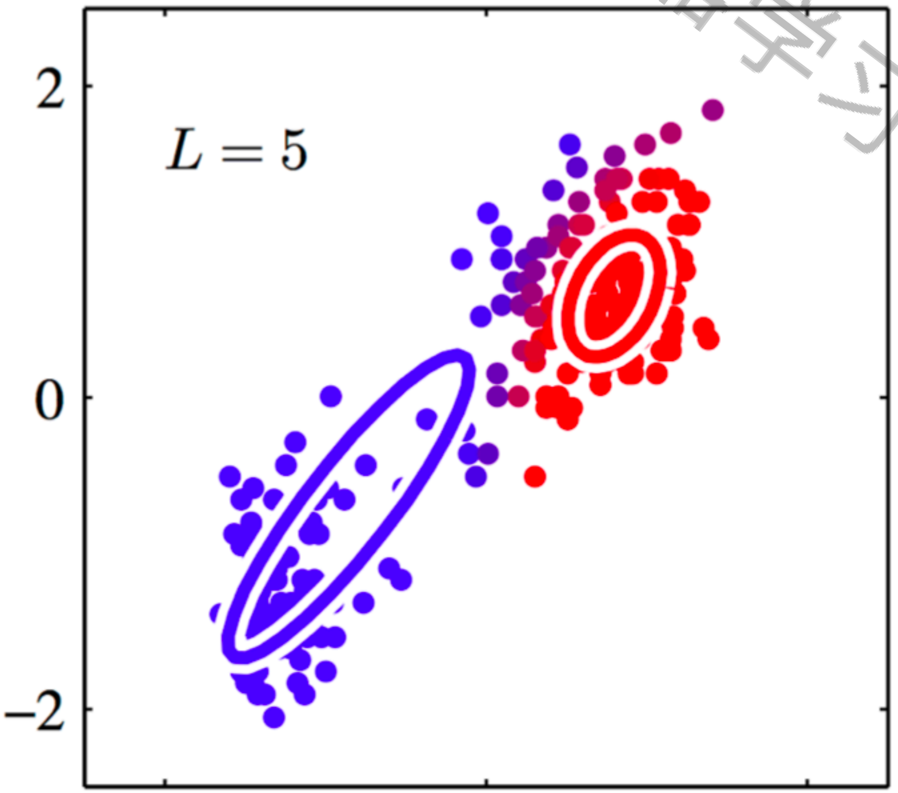


(c)

四、Mixture of Gaussian

四川大学-机器学习引论

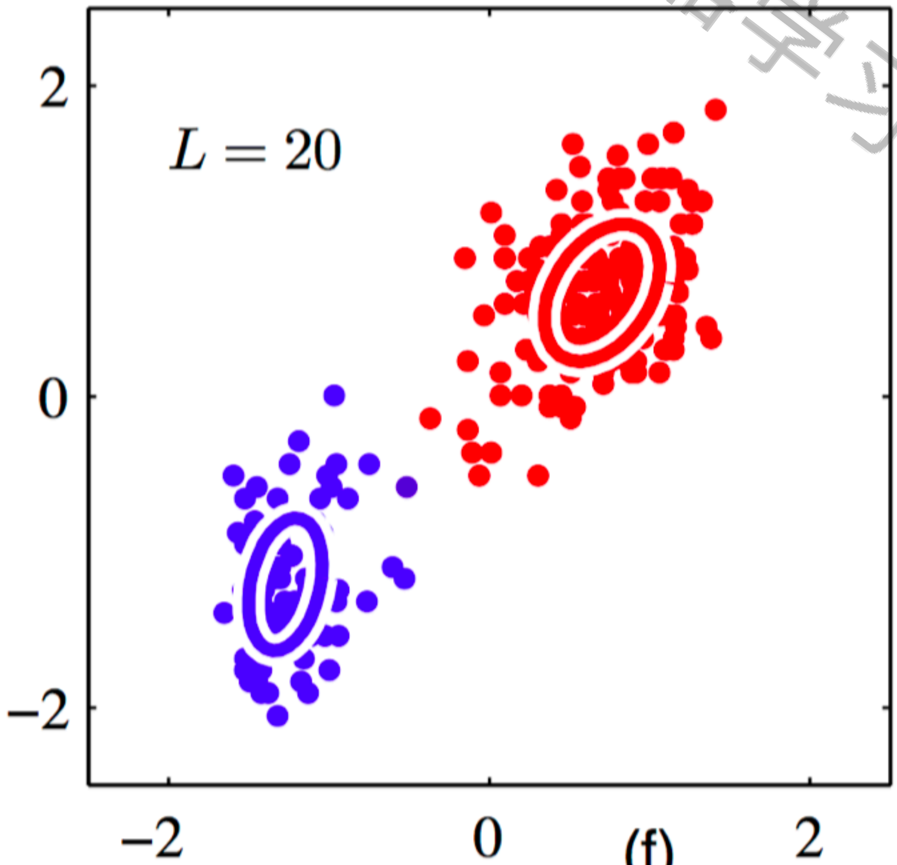
- After 5 rounds of EM:



(e)

四、Mixture of Gaussian

- After 20 rounds of EM:



(f)

四川大学-机器学习引论

四、Mixture of Gaussian

四川大学-机器学习引论

k-means vs. MOG

- EM for GMM seems a little like k-means.
- In fact, there is a precise correspondence.
- First, fix each cluster covariance matrix to be $\sigma^2 I$.
- As we take $\sigma^2 \to 0$, the update equations converge to doing k-means.
- If you do a quick experiment yourself, you'll find
  - Soft assignments converge to hard assignments.
  - Has to do with the tail behavior (exponential decay) of Gaussian.

Test Questions:

四川大学-机器学习引论

- Write k-means in pseudocode?
- Write MOG in pseudocode?
- What the limitations of k-means besides that I given?
- What the limitations of MOG besides that I given?
- What the connections between k-means and MOG?
- What the difference between k-means and MOG?
- What the advantage of k-means/MOG over MOG/k-means?

Further reading:

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," presented at the International Conference on Learning Representations, 2014.

Final Test

四川大学-机器学习引论

For a given data set (mnist test partition), achieving

1. a classification accuracy over 80% using the methods introduced in this course. Report the corresponding F-measure.
2. alternatively, a clustering accuracy over 58% using the methods introduced in this course. Report the corresponding NMI.

Requirements:

- Give the design details and explain why it as does
- Report the mean and std score
- Report the tuned parameters
- Report the hardware and used time cost

Q&A

THANKS!

四川大学-机器学习引论