School of Computing
UNIVERSITY OF GEORGIA

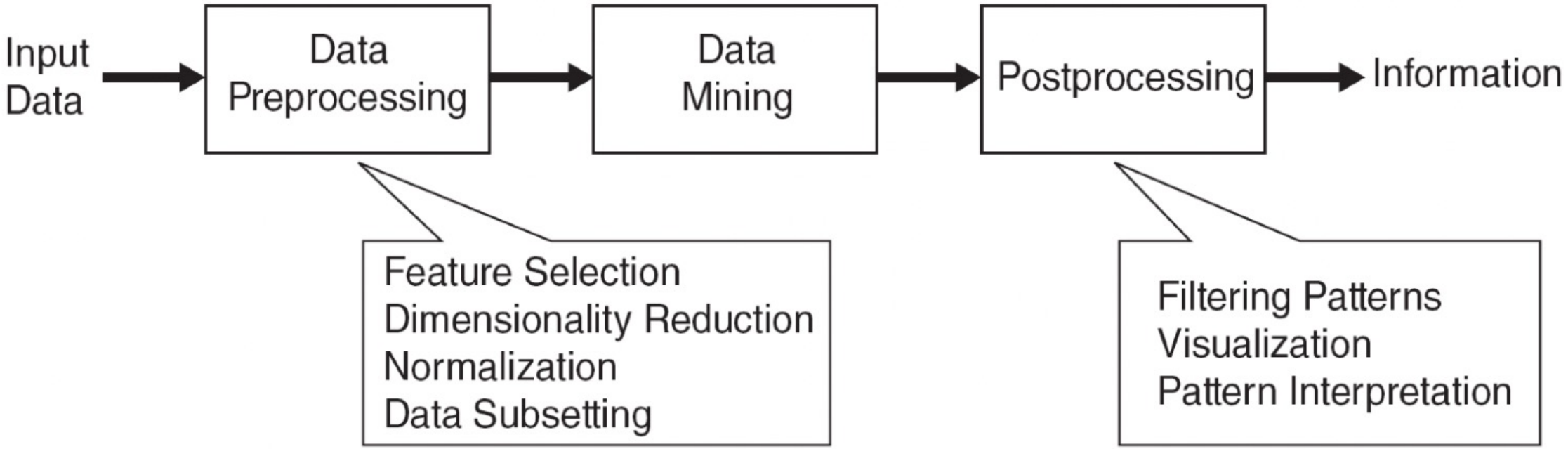
CSCI 4380/6380 DATA MINING

Fei Dou

Assistant Professor
School of Computing
University of Georgia

November 02, 07, 2023

Recap: Data Mining Process



Evaluation Methods

Evaluation Methods

- A pair of data object (O_i,O_j) falls into one of the following categories
  - SS: C_{ij}=1 and P_{ij}=1; (agree)
  - DD: C_{ij}=0 and P_{ij}=0; (agree)
  - SD: C_{ij}=1 and P_{ij}=0; (disagree)
  - DS: C_{ij}=0 and P_{ij}=1; (disagree)

- **Rand index** $$Rand = \frac{|Agree|}{|Agree|+|Disagree|} = \frac{|SS|+|DD|}{|SS|+|SD|+|DS|+|DD|}$$
  - may be dominated by DD

- **Jaccard Coefficient** $$Jaccard\ \ coefficien\ t = \frac{|SS|}{|SS|+|SD|+|DS|}$$

Evaluation Methods

Clustering

	g 1	g 2	g 3	g 4	g 5
g 1	1	1	1	0	0
g 2	1	1	1	0	0
g 3	1	1	1	0	0
g 4	0	0	0	1	1
g 5	0	0	0	1	1

Groundtruth

	g 1	g 2	g 3	g 4	g 5
g 1	1	1	0	0	0
g 2	1	1	0	0	0
g 3	0	0	1	1	1
g 4	0	0	1	1	1
g 5	0	0	1	1	1

Clustering

Ground truth	Same Cluster	Different Cluster
Same Cluster	9	4
Different Cluster	4	8

$$Rand = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|} = \frac{17}{25}$$

$$Jaccard = \frac{|SS|}{|SS| + |SD| + |DS|} = \frac{9}{17}$$

Evaluation Methods

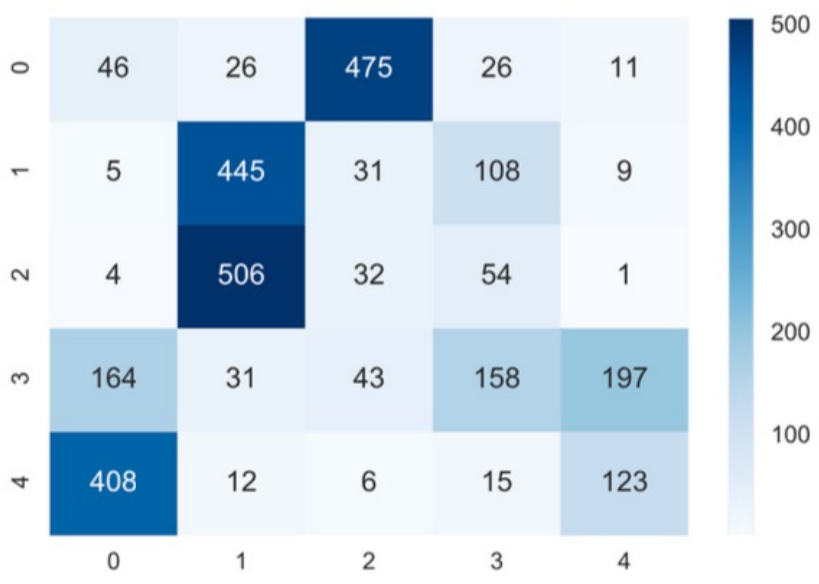
- Classification Accuracy vs. Clustering Accuracy

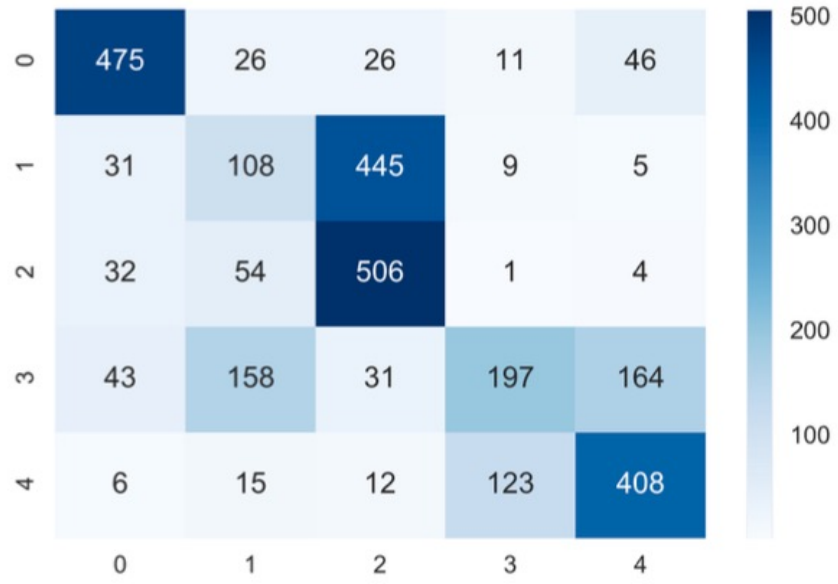
$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i)$$

$$accuracy(y, \hat{y}) = \max_{perm \in P} \frac{1}{n} \sum_{i=0}^{n-1} 1(perm(\hat{y}_i) = y_i)$$

Evaluation Methods

- Classification Accuracy vs. Clustering Accuracy
- Example:





Evaluation Methods

- **Notation**
  - $|C_k \cap P_j|$ the number of objects in both the *k*-th cluster of the clustering solution and *j*-th cluster of the groundtruth
  - $|C_k|$ the number of objects in the *k*-th cluster of the clustering solution
  - $|P_j|$ the number of objects in the *j*-th cluster of the groundtruth

- **Purity** $$Purity = \frac{1}{N}\sum_k \max_j |C_k \cap P_j|$$

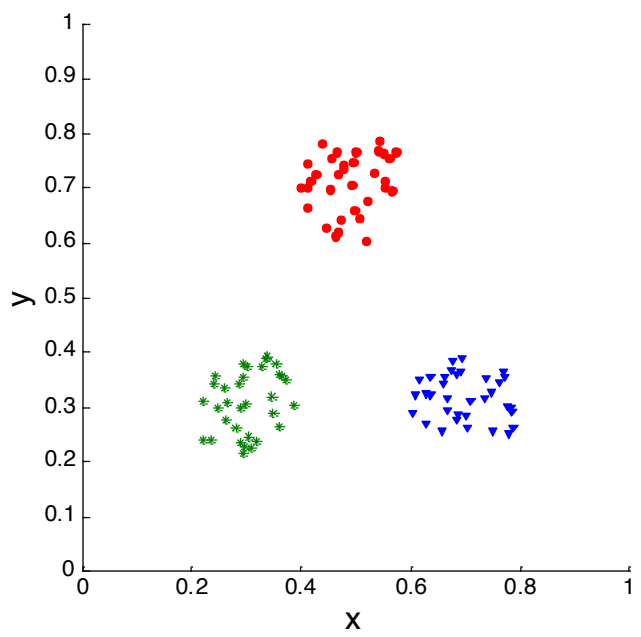
- **Normalized Mutual Information**

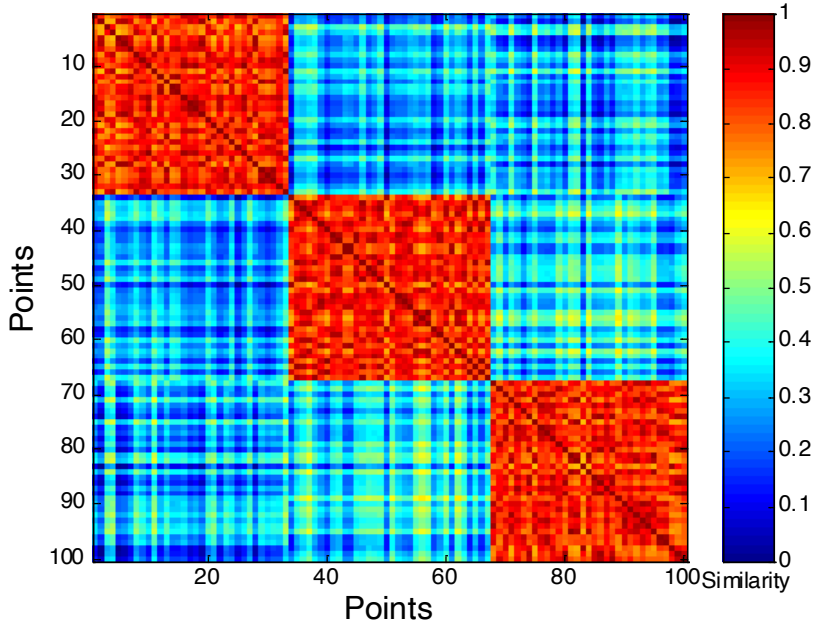
$$NMI = \frac{I(C,P)}{\sqrt{H(C)H(P)}} \qquad I(C,P) = \sum_k \sum_j \frac{|C_k \cap P_j|}{N} \log \frac{N \cdot |C_k \cap P_j|}{|C_k||P_j|}$$

$$H(C) = \sum_k \frac{|C_k|}{N} \log \frac{|C_k|}{N} \qquad H(P) = \sum_j \frac{|P_j|}{N} \log \frac{|P_j|}{N}$$

Judging a Clustering Visually by its Similarity Matrix

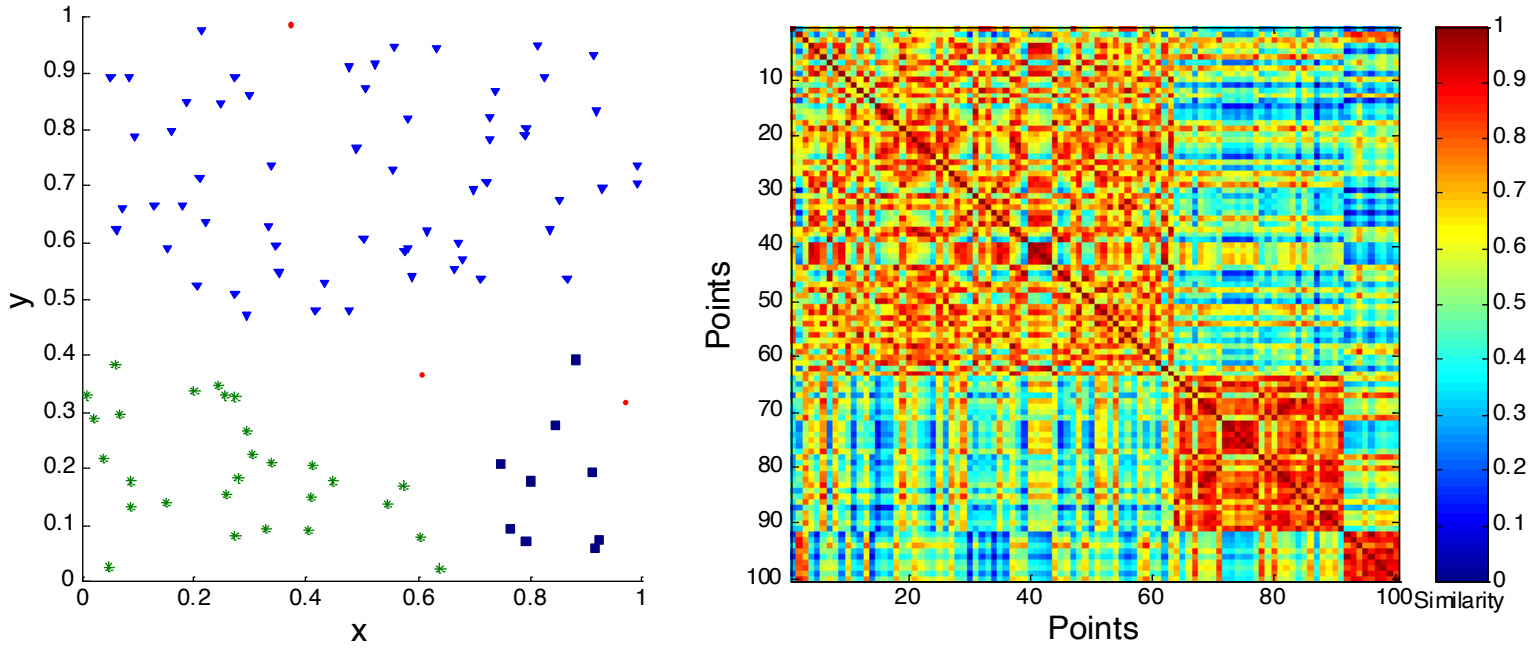
- Visualization/Interpretation
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)
  - Order the similarity matrix with respect to cluster labels and inspect visually.





Judging a Clustering Visually by its Similarity Matrix

- Visualization/Interpretation
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)
  - Order the similarity matrix with respect to cluster labels and inspect visually.
  - Clusters in random data are not so crisp

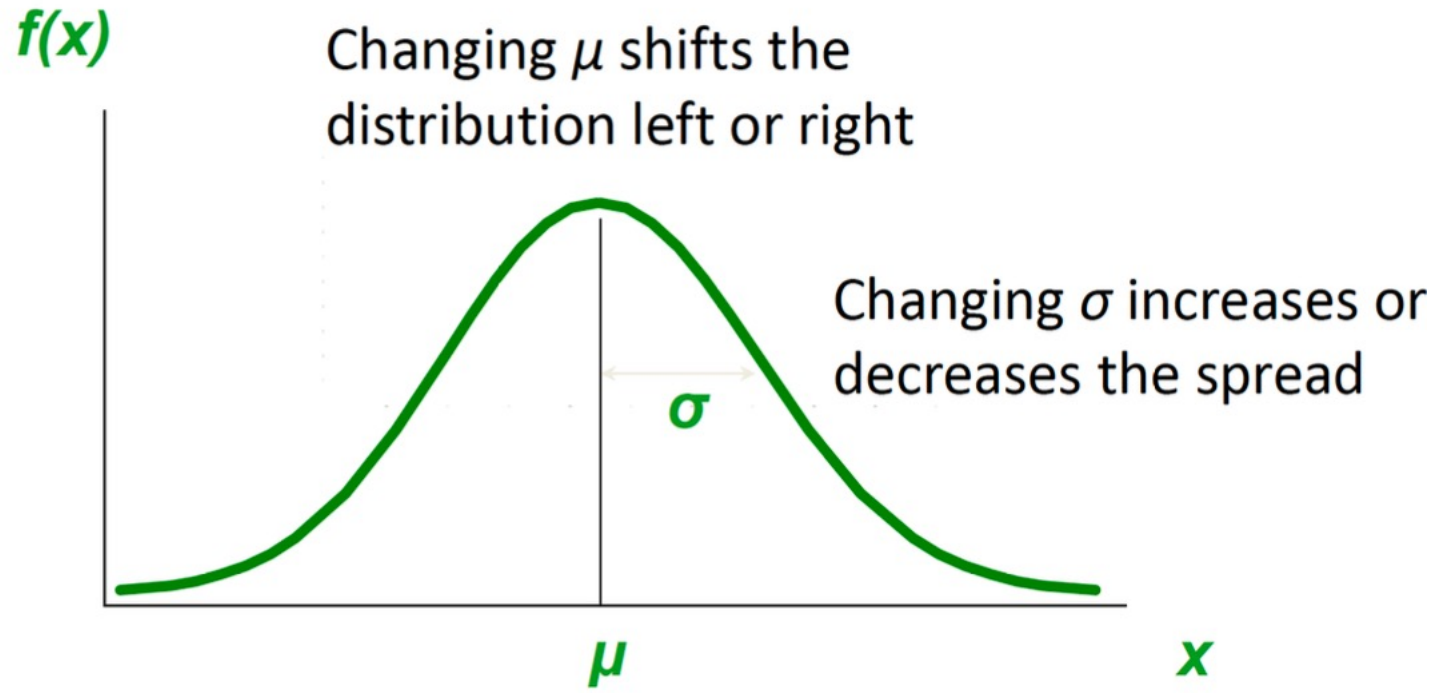


Mixture based Clustering

Mixture Based Clustering

- Using Probabilistic Models for Clustering
  - Hard vs. soft clustering:
    - Hard clustering: Every sample belongs to exactly one cluster
    - Soft clustering: Every sample belongs to several clusters with certain degrees
- Probabilistic clustering:
  - Each cluster is mathematically represented by a parametric distribution
  - The entire data set is modeled by a mixture of these distributions

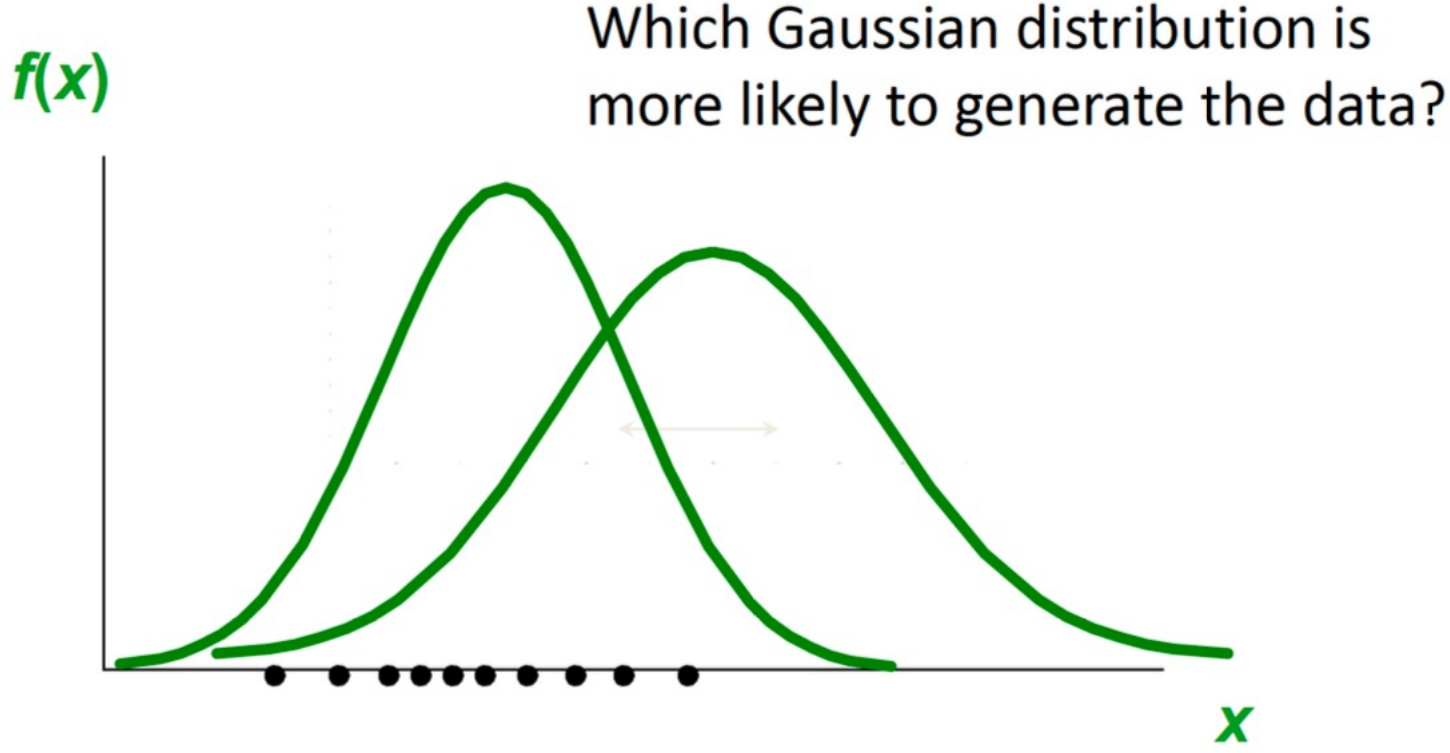
Mixture based Clustering - Gaussian Distribution



Probability density function $f(x)$ is a function of x given μ and σ

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Mixture based Clustering - Likelihood



Define likelihood as a function of μ and σ given $x_1, x_2, \ldots, x_n$

$$\prod_{i=1}^{n} N(x_i \mid \mu, \sigma^2)$$

Mixture based Clustering

- **Gaussian Distribution**

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

  - $\mu \in R^d$ is the mean and $\Sigma \in R^{d\times d}$ is the covariance matrix.

- **Likelihood**

$$L(\mu,\Sigma) = \sum_{i=1}^{n} \log N(\mathbf{x}_i|\mu,\Sigma)$$

$$= \sum_{i=1}^{n} \left(\frac{1}{2}(\mathbf{x}_i-\mu)^{\top}\Sigma^{-1}(\mathbf{x}_i-\mu) - \ln((2\pi)^{d/2}|\Sigma|^{1/2})\right)$$

Mixture based Clustering

- **Maximum Likelihood Estimate(MLE)**
- Find model parameters that maximize log likelihood L(μ, Σ)
- **MLE** for Gaussian

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

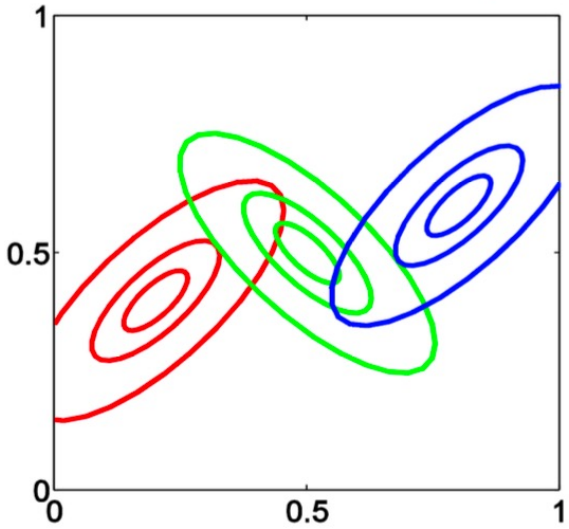
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^\top$$

Mixture based Clustering

- **Gaussian Mixture**
  - Linear combination of Gaussians

$$P(\mathbf{x}|\mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \Sigma_k)$$

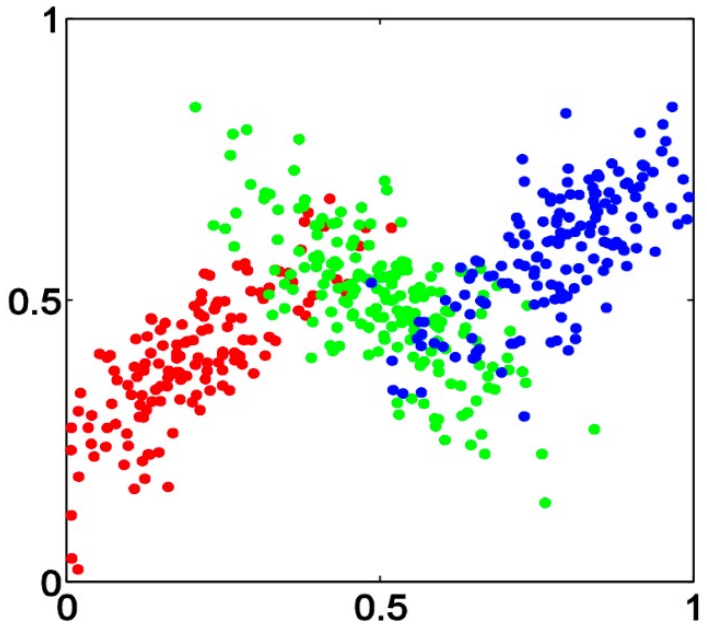
where $\sum_{k=1}^{K} \pi_k = 1, 0 \leq \pi_k \leq 1$



Mixture based Clustering - GMM

- To generate a data point:
  - first pick one of the components with probability π_k
  - then draw a sample x_i from that component distribution
- Each data point is generated by one of *K* components, a latent variable $z_i = (z_{i1}, \ldots, z_{iK})$ is associated with each x_i

$$\sum_{k=1}^{K} z_{ik} = 1 \text{ and } p(z_{ik} = 1) = \pi_k$$



Mixture based Clustering - GMM

- Maximize log likelihood

$$\ln p(x|\pi,\mu,\Sigma) = \sum_{i=1}^{n} \ln\{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k,\Sigma_k)\}$$

- Without knowing values of latent variables, we have to maximize the incomplete log likelihood

Mixture based Clustering - EM-Algorithm

- E-step: for given parameter values we can compute the expected values of the latent variables (responsibilities of data points)

$$
\begin{aligned}
r_{ik} \equiv E(z_{ik}) &= p(z_{ik}=1|x_i,\pi,\mu,\Sigma) \\
&= \frac{p(z_{ik}=1)p(x_i|z_{ik}=1,\pi,\mu,\Sigma)}{\sum_{k=1}^{K} p(z_{ik}=1)p(x_i|z_{ik}=1,\pi,\mu,\Sigma)} \\
&= \frac{\pi_k \mathcal{N}(x_i|u_k,\Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|u_k,\Sigma_k)}
\end{aligned}
$$

  - Note that $r_{ik} \in [0,1]$ instead of $\{0,1\}$ but we still have $\sum_{k=1}^{K} r_{ik} = 1$ for all i

Mixture based Clustering - EM-Algorithm

- M-step: maximize the expected complete log likelihood

$$E[\ln p(x, z|\pi, \mu, \Sigma)] = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \{\ln \pi_k + \ln \mathcal{N}(x_i|\mu_k, \Sigma_k)\}$$

- Parameter update:

$$\pi_k = \frac{\sum_i r_{ik}}{n} \qquad \mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}$$

$$\Sigma_k = \frac{\sum_i r_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i r_{ik}}$$

Mixture based Clustering - EM-Algorithm

EM-Algorithm

- Iterate E-step and M-step until the log likelihood of data does not increase any more
  - Converge to local optimal
  - Need to restart algorithm with different initial guess of parameters (as in K-means)
- Relation to K-means
  - Consider GMM with common covariance
  - $\Sigma_k = \sigma^2 I$
  - As $\sigma^2 \to 0$, $r_{ik} \to 0$ or 1, two methods coinside.

K-means vs GMM

- Objective function
  - Minimize sum of squared Euclidean distance
- Can be optimized by an EM algorithm
  - E-step: assign points to clusters
  - M-step: optimize clusters
  - Performs hard assignment during E-step
- Assumes spherical clusters with equal probability of a cluster

- Objective function
  - Maximize log-likelihood
- EM algorithm
  - E-step: Compute posterior probability of membership
  - M-step: Optimize parameters
  - Perform soft assignment during E-step
- Can be used for non-spherical clusters
- Can generate clusters with different probabilities