Clustering
Lecture 5: Mixture Model

Jing Gao
SUNY Buffalo
Outline

• **Basics**
  – Motivation, definition, evaluation

• **Methods**
  – Partitional
  – Hierarchical
  – Density-based
  – Mixture model
  – Spectral methods

• **Advanced topics**
  – Clustering ensemble
  – Clustering in MapReduce
  – Semi-supervised clustering, subspace clustering, co-clustering, etc.
Using Probabilistic Models for Clustering

• **Hard vs. soft clustering**
  – Hard clustering: Every point belongs to exactly one cluster
  – Soft clustering: Every point 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
Gaussian Distribution

Changing $\mu$ shifts the distribution left or right

Changing $\sigma$ increases or decreases the spread

Probability density function $f(x)$ is a function of $x$ given $\mu$ and $\sigma$

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right)$$
Define likelihood as a function of $\mu$ and $\sigma$ given $x_1, x_2, \ldots, x_n$

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

Which Gaussian distribution is more likely to generate the data?
Gaussian Distribution

- Multivariate Gaussian

\[
\mathcal{N}(x | \mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}
\]

mean \hspace{1cm} covariance

- Log likelihood

\[
L(\mu, \Sigma) = \sum_{i=1}^{n} \ln \mathcal{N}(x_i | \mu, \Sigma) = \sum_{i=1}^{n} \left( -\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) - \pi \ln |\Sigma| \right)
\]
Maximum Likelihood Estimate

• MLE
  – Find model parameters $\mu, \Sigma$ that maximize log likelihood

\[ L(\mu, \Sigma) \]

• MLE for Gaussian

\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \]
Gaussian Mixture

- Linear combination of Gaussians

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leq \pi_k \leq 1 \]

parameters to be estimated
Gaussian Mixture

- To generate a data point:
  - first pick one of the components with probability $\pi_k$
  - then draw a sample $\mathcal{X}_i$ from that component distribution
- Each data point is generated by one of $K$ components, a latent variable $\tilde{z}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{iK})$ is associated with each $\mathcal{X}_i$
  \[ \sum_{k=1}^{K} \tilde{z}_{ik} = 1 \text{ and } p(\tilde{z}_{ik} = 1) = \pi_k \]
Gaussian Mixture

• Maximize log likelihood

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

• Without knowing values of latent variables, we have to maximize the incomplete log likelihood
Expectation-Maximization (EM) Algorithm

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

\[
r_{ik} \equiv E(z_{ik}) = \frac{p(z_{ik} = 1 \mid x_i, \pi, \mu, \Sigma) \cdot p(x_i \mid z_{ik} = 1, \pi, \mu, \Sigma)}{\sum_{k=1}^{K} p(z_{ik} = 1) \cdot p(x_i \mid z_{ik} = 1, \pi, \mu, \Sigma)} = \frac{\pi_k \mathcal{N}(x_i \mid u_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid u_k, \Sigma_k)}
\]

- Note that \( r_{ik} \in [0, 1] \) instead of \( \{0, 1\} \) but we still have \( \sum_{k=1}^{K} r_{ik} = 1 \) for all \( i \)
**Expectation-Maximization (EM) Algorithm**

- **M-step:** maximize the expected complete log likelihood

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

- **Parameter update:**

\[
\pi_k = \frac{\sum_i r_{ik}}{n} \quad \mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}
\]

\[
\Sigma_k = \frac{\sum_i r_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i r_{ik}}
\]
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
    $\sum_k = \delta^2 I$
  – As $\delta^2 \to 0$, $r_{ik} \to 0$ or 1, two methods coincide
L = 20
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**
Mixture Model

• **Strengths**
  – Give probabilistic cluster assignments
  – Have probabilistic interpretation
  – Can handle clusters with varying sizes, variance etc.

• **Weakness**
  – Initialization matters
  – Choose appropriate distributions
  – Overfitting issues
Take-away Message

- Probabilistic clustering
- Maximum likelihood estimate
- Gaussian mixture model for clustering
- EM algorithm that assigns points to clusters and estimates model parameters alternatively
- Strengths and weakness