CS456: Machine Learning

Mixture model

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

February 28, 2020

Objectives

- To learn how to perform clustering using mixture model

Objectives

- To learn how to perform clustering using mixture model
- To relate mixture model to k-means clustering

Outlines

- Mixture model
- Gaussian Mixture Model (GMM) for clustering
- GMM learning algorithm

Overview

- Sometimes basic probability distribution cannot explain complex data

Overview

- Sometimes basic probability distribution cannot explain complex data
- Mixture model constructs complex probability distribution by linear combination of basis distributions

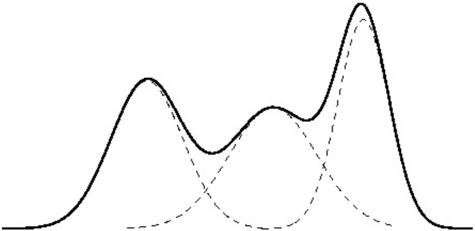


Figure: There's no distribution corresponds to black line but it can be explained by 3 Gaussians

Mixture model

- A probability distribution f is a mixture of K component distributions $f_1, f_2, \ldots, f_K$ if

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

where π_k is the mixing weights and $\pi_k > 0, \sum_k \pi_k = 1$

Mixture model

- A probability distribution f is a mixture of K component distributions $f_1, f_2, \ldots, f_K$ if

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

where π_k is the mixing weights and $\pi_k > 0, \sum_k \pi_k = 1$

- The distribution f can be any probabilistic distribution

Mixture model

- A probability distribution f is a mixture of K component distributions $f_1, f_2, \ldots, f_K$ if

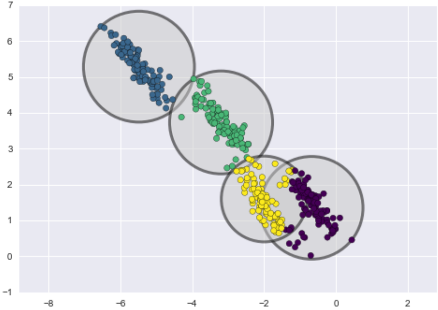
$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

where π_k is the mixing weights and $\pi_k > 0, \sum_k \pi_k = 1$

- The distribution f can be any probabilistic distribution

- Usually, people employ Gaussians (Normal distribution)

Mixture model for clustering



- Given a set of data, we would like to find basis distributions that when combined can explain the data (fit data) as much as possible

- Note: wo do not have access to y (correct cluster assignment)

K-means' hard assignment

- Recall how k-means assign data point to a cluster based on some distance measure e.g., Euclidean distance

$$z_i = \arg\min_{k=1:K} d(\mathbf{x}_i, \mu_k)$$

K-means' hard assignment

- Recall how k-means assign data point to a cluster based on some distance measure e.g., Euclidean distance

$$z_i = \arg\min_{k=1:K} d(\mathbf{x}_i, \mu_k)$$

- For example if $K = 3$ we need to decide

$$\begin{aligned} z_i &= \arg\min_{k=1:K} [0.2, 0.8, 1.3] \\ &= 1 \end{aligned}$$

K-means' hard assignment

- Recall how k-means assign data point to a cluster based on some distance measure e.g., Euclidean distance

$$z_i = \arg\min_{k=1:K} d(\mathbf{x}_i, \mu_k)$$

- For example if $K = 3$ we need to decide

$$\begin{aligned} z_i &= \arg\min_{k=1:K} [0.2, 0.8, 1.3] \\ &= 1 \end{aligned}$$

- Function arg min hard assigns data point to cluster

Soft assignment

- Instead of hard assignment, we can keep z_i as a vector representing the probabilities that $\mathbf{x}_i$ belongs to each of the clusters

Soft assignment

- Instead of hard assignment, we can keep z_i as a vector representing the probabilities that $\mathbf{x}_i$ belongs to each of the clusters
- For example, if $\mathbf{x}_i$ is more likely to come from cluster 2 out of 3 possible clusters we have $z_i = [0.01, 0.95, 0.04]$

Soft assignment

- Instead of hard assignment, we can keep z_i as a vector representing the probabilities that $\mathbf{x}_i$ belongs to each of the clusters
- For example, if $\mathbf{x}_i$ is more likely to come from cluster 2 out of 3 possible clusters we have $z_i = [0.01, 0.95, 0.04]$
- Note that z_i sums to 1 (probability that $\mathbf{x}_i$ belongs any of the 3 clusters must be 1)

The degree of membership

- What are the probabilities in z_i, and how are they calculated ?

The degree of membership

- What are the probabilities in z_i, and how are they calculated ?
- Essentially, each probabilities in z_i is the probability of $z_i = k$ after we see $\mathbf{x}_i$, which is $\underline{p(z_i = k|\mathbf{x}_i)} \leftarrow$ soft label

The degree of membership

- What are the probabilities in z_i, and how are they calculated ?
- Essentially, each probabilities in z_i is the probability of $z_i = k$ after we see $\mathbf{x}_i$, which is $\underline{p(z_i = k|\mathbf{x}_i)} \leftarrow$ soft label
- By Bayes' rule we know that $p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\sum_z p(x|z)p(z)}$
- It turns out to calculate cluster membership we first calculate

The degree of membership

- What are the probabilities in z_i, and how are they calculated ?
- Essentially, each probabilities in z_i is the probability of $z_i = k$ after we see $\mathbf{x}_i$, which is $\underline{p(z_i = k|\mathbf{x}_i)} \leftarrow$ soft label
- By Bayes' rule we know that $p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\sum_z p(x|z)p(z)}$
- It turns out to calculate cluster membership we first calculate
  - $p(x|z)$
  - $p(z)$

Cluster representation

- $p(x|z)$ can be modelled by a probability distribution

Cluster representation

- $p(x|z)$ can be modelled by a probability distribution
  - If we choose Normal distribution, $p(x|z)$ is calculated by

$$p(x|z) = \mathcal{N}(x; \mu_z, \Sigma_z)$$

Cluster representation

- $p(x|z)$ can be modelled by a probability distribution
  - If we choose Normal distribution, $p(x|z)$ is calculated by

$$p(x|z) = \mathcal{N}(x; \mu_z, \Sigma_z)$$

- $p(z) := \pi$ is cluster prior probability

Cluster representation

- $p(x|z)$ can be modelled by a probability distribution
  - If we choose Normal distribution, $p(x|z)$ is calculated by

$$p(x|z) = \mathcal{N}(x; \mu_z, \Sigma_z)$$

- $p(z) := \pi$ is cluster prior probability
  - It represents the ratio of the points we think they are in this cluster according to $p(z|x)$ divided by total number of data points
  - $p(z = 1) = \frac{\sum_{i=1}^{N} p(z_i=1|x)}{N}$

What we have so far ?

- We assume each clusters is modelled by some probability distribution e.g., Normal distribution,

$$p(x|z) = \mathcal{N}(x; \mu_z, \Sigma_z)$$

What we have so far ?

- We assume each clusters is modelled by some probability distribution e.g., Normal distribution,

$$p(x|z) = \mathcal{N}(x; \mu_z, \Sigma_z)$$

- Further we also assume probability of observing cluster $p(z) := \pi_z$

What we have so far ?

- We assume each clusters is modelled by some probability distribution e.g., Normal distribution,

$$p(x|z) = \mathcal{N}(x; \mu_z, \Sigma_z)$$

- Further we also assume probability of observing cluster $p(z) := \pi_z$
- Knowing the the two probabilities allows us to compute cluster membership probabilities $p(z = k|x)$

What do we want ?

- We want to find μ, Σ, π which best describes the data

What do we want ?

- We want to find μ, Σ, π which best describes the data
  - If it was a supervised learning we would have information of class label y, and estimating μ, Σ, π would be trivial
- We also want to find cluster assignment $\mathbf{z}$ (just like in k-means)

Estimating **z**

- If we fix μ, Σ, π, the values of **z** can be easily calculate using Bayes' rule

Estimating $\mathbf{z}$

- If we fix μ, Σ, π, the values of $\mathbf{z}$ can be easily calculate using Bayes' rule

- For example if we assume 3 clusters
  - $p(z_i = 1|\mathbf{x}_i) = p(z_i = 1)\mathcal{N}(\mathbf{x}_i; \mu_1, \Sigma_1)/p(\mathbf{x}_i)$
  - $p(z_i = 2|\mathbf{x}_i) = p(z_i = 2)\mathcal{N}(\mathbf{x}_i; \mu_2, \Sigma_2)/p(\mathbf{x}_i)$
  - $p(z_i = 3|\mathbf{x}_i) = p(z_i = 3)\mathcal{N}(\mathbf{x}_i; \mu_3, \Sigma_3)/p(\mathbf{x}_i)$

Estimating $\mathbf{z}$

- If we fix μ, Σ, π, the values of $\mathbf{z}$ can be easily calculate using Bayes' rule
- For example if we assume 3 clusters
  - $p(z_i = 1|\mathbf{x}_i) = p(z_i = 1)\mathcal{N}(\mathbf{x}_i; \mu_1, \Sigma_1)/p(\mathbf{x}_i)$
  - $p(z_i = 2|\mathbf{x}_i) = p(z_i = 2)\mathcal{N}(\mathbf{x}_i; \mu_2, \Sigma_2)/p(\mathbf{x}_i)$
  - $p(z_i = 3|\mathbf{x}_i) = p(z_i = 3)\mathcal{N}(\mathbf{x}_i; \mu_3, \Sigma_3)/p(\mathbf{x}_i)$
- This is equivalent to updating cluster assignment in k-means (but with soft label)

Estimating μ based on soft label

We will take $p(z|x)$ as an soft cluster assignment

Estimating μ based on soft label

We will take $p(z|x)$ as an soft cluster assignment

(NDA)

$$\mu_k = \frac{\sum_{i=1}^{N_k} \mathbf{x}_i}{N}$$

(GMM)

$$\mu_k = \frac{\sum_{i=1}^{N_k} p(z_i = k|\mathbf{x}_i)\mathbf{x}_i}{N}$$

Estimating Σ based on soft label

We will take $p(z|x)$ as an soft cluster assignment

Estimating Σ based on soft label

We will take $p(z|x)$ as an soft cluster assignment

(NDA)

$$\Sigma_k = \frac{\sum_{i=1}^{N_k} \mathbf{x}_i \mathbf{x}_i^T}{N}$$

(GMM)

$$\Sigma_k = \frac{\sum_{i=1}^{N_k} p(z_i = k|\mathbf{x}_i)(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{N}$$

Estimating π based on soft label

We will take $p(z|x)$ as an soft cluster assignment

Estimating π based on soft label

We will take $p(z|x)$ as an soft cluster assignment

(NDA)

$$\pi_k = \frac{N_k}{N}$$

(GMM)

$$\pi_k = \frac{\sum_{i=1}^{N} p(z_i = k|x)}{N}$$

Likelihood function

- To measure how well μ, Σ, π describe the data
- we can calculate the likelihood for the model

$$\mathcal{L}(\mu, \Sigma, \pi; \{\mathbf{x}_i\}_{i=1}^{N}) = \prod_{i=1}^{N} f(\mathbf{x}_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k) \tag{1}$$

Likelihood function

- To measure how well μ, Σ, π describe the data
- we can calculate the likelihood for the model

$$\mathcal{L}(\mu, \Sigma, \pi; \{\mathbf{x}_i\}_{i=1}^{N}) = \prod_{i=1}^{N} f(\mathbf{x}_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k) \qquad (1)$$

- Obtaining the log-likelihood

$$llh(\mu, \Sigma, \pi; \{\mathbf{x}_i\}_{i=1}^{N}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k) \right) \qquad (2)$$

GMM algorithm

1. Initialisation: choose k and initialise μ_k, Σ_k, π_k artibrarily for all k
2. Repeat until likelihood converges
   - estimate soft cluster assignment

$$p(z_i = k|\mathbf{x}_i) = \frac{p(z_i = k)\mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k)}{\sum_k p(z_i = k)\mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k)}$$

   - update μ, Σ, π

$$\mu_k = \frac{\sum_{i=1}^{N_k} p(z_i = k|\mathbf{x}_i)\mathbf{x}_i}{N}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N_k} p(z_i = k|\mathbf{x}_i)(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{N}$$

$$\pi_k = \frac{\sum_{i=1}^{N} p(z_i = k|x)}{N}$$

Objectives: revisited

- To learn how to perform clustering using mixture model
- To relate mixture model to k-means clustering