Mixture Models

1 Introduction

[TBW]

2 EM for a mixture model

We want to train a Gaussian mixture model using EM. Let assume we have this mixture model:

$$g(x) = \sum_{k=1}^{K} \pi_k g_k(\mathbf{x}),\ g_k = \mathcal{N}\left(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}\right)$$

such that $\sum_{i=1}^{K} \pi_k = 1$ and $\Theta = \left\{ \{\pi_k, \boldsymbol{\mu}_k\}_{k=1}^{K}, \sigma^2 \right\}$ are unknown variables. Assuming that we have the training data $\{\mathbf{x}_n, g_n\}_{i=1}^{n}$, we can write the likelihood as following:

$$\mathcal{L} = \log \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_k, \sigma^2 \mathbf{I}\right) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_k, \sigma^2 \mathbf{I}\right). \quad (1)$$

Now we could find the MLE estimation of the parameters using gradient with respect to parameters of the model:

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0,\ \frac{\partial \mathcal{L}}{\partial \mu_k} = 0,\ \frac{\partial \mathcal{L}}{\partial \sigma^2} = 0.$$

Because taking the derivatives of the likelihood in Eq. 1 is hard, we could change its form by adding a categorical latent variables, $\{z_k\}_{k=1}^{n}\ s.t.\ z_k = 0, \ldots, K$, that determine each of the samples come from which component:

$$z_i \sim \text{Cat}(K, p)$$

$$\mathbf{x}_i | z_i = k\ \sim \mathcal{N}\left(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}\right)$$

$$\mathcal{L} = \sum_{k=1}^{K} \sum_{i:z_i=k}^{n} \log \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_k, \sigma^2 \mathbf{I}\right) + \log \pi_k.$$

If we know z_i, the MLE estimation for each of the parameters could be found by

$$\hat{\pi}_k = \frac{n_k}{n},\ \hat{\mu}_k = \frac{1}{n_k} \sum_{i:z_i=k} \mathbf{x}_i,\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i\right)^T \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i\right).$$

Now based the values of the model parameters we can calculate the probability of $\mathbf{x}_i$ belonging to one the component k by

$$\gamma_{ik} = \Pr\left(z_i = k | \mathbf{x}_i, \theta\right) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_k, \sigma^2 \mathbf{I}\right)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_{k'}, \sigma^2 \mathbf{I}\right)}. \tag{2}$$

Using the above criterion we can modify the likelihood updates as

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \gamma_{ik}}{n},\ \hat{\mu}_k = \frac{\sum_{i=1}^{n} \gamma_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ik}},\ \hat{\sigma}^2 = \frac{\sum_i \gamma_{ik} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i\right)^T \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i\right)}{\sum_{i=1}^{n} \gamma_{ik}} \tag{3}$$

One important question is that how to initialize the model parameters? A good way to construct initial guesses for μ_1 and μ_2 is simply to choose K of the training data at random. For σ^2 we can set it equal to the sample variance $\sum_{i=1}^{N} \left(\mathbf{x}_i - \hat{\mathbf{x}}\right)^2 / N$, and the mixing proportions, $\pi_k = 0.5$.

2.1 Gaussian Mixture Model as special case of k-means clustering!

First we have a short review on k-means clustering. In k-means clustering algorithm we aim to partition $X = \{x_1, \ldots, x_n\}$ into k clusters. Thus we define $\{\mu_i\}_{i=1}^{K}$ as centre of clusters, and $\left\{\{r_{i,k}\}_{i=1}^{n}\right\}_{k=1}^{K}$ as indicator variables. Each $r_{i,k}$ is 1 if and only if, x_i belongs to cluster k. The goal of the clustering is to minimize the sum of distances of points in the same cluster from the mean of the cluster:

$$J(\mathbf{r}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{i,k} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

It can be shown that iterative repetition of the following two steps will result in the above objective function:
Step1: Assuming $\boldsymbol{\mu}$ is determined, we can find $\mathbf{r}$ by

$$r_{i,k} = 1 \text{ if } k = \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

Step2: Assuming $\mathbf{r}$ is fixed, we can find each centre of cluster by

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_i \mathbf{x}_i}{\sum_{i=1}^{n} r_i} \tag{4}$$

If we assume that $\sigma \to 0$, the Gaussian distribution becomes one infinite mass at mean. So $\gamma_{i,k}$ in Eq. 2 becomes 1 only for $\mathbf{x}_i$ which is closest to $\boldsymbol{\mu}_k$, which is like $r_{i,k}$ in k-means clustering. Consequently Eq. 3 reduces to Eq. 4, and in overall results in k-means clustering algorithm.