Mixture Models and EM - Tutorial 14

Aytunc Sahin
aytunc.sahin@inf.ethz.ch

May 2019

Introduction to EM

We can use a mixture of Gaussians to model complex distributions and capture multimodality. A Gaussian Mixture Model (GMM) can be written as a convex combination of different Gaussians as follows:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

We also introduce a latent random variable $\mathbf{z}$ where $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$. We define the prior distribution over $\mathbf{z}$ using mixing coefficients:

$$p(z_k = 1) = \pi_k \quad \text{with} \quad 0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

Since only one z_k is 1, we can write the prior and the conditional distribution as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \quad \text{and} \quad p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Note that for every *observed* $\mathbf{x}$ there is a corresponding *latent* (*unobserved*) $\mathbf{z}$. Now we look at the posterior probability of $\mathbf{z}$ using Bayes' theorem:

$$\begin{aligned} p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned}$$

We can also view $p(z_k = 1|\mathbf{x})$ as the responsibility r_k that component k takes for explaining the observation $\mathbf{x}$. Now we suppose that we have N observed data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and the log-likelihood using a GMM can be written as

$$\log p(\mathbf{x}_{1:N}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$$

Maximizing this log-likelihood is much more difficult than maximizing a single Gaussian because we have the sum inside of the logarithm, thus the logarithm cannot directly act on a Gaussian. If we take the derivative of the log-likelihood with respect to model parameters and set it to zero, we get

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} r_{nk}} \\
\pi_k &= \frac{\sum_{n=1}^{N} r_{nk}}{\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}}
\end{aligned}$$

Moreover, we have $N_k = \sum_{n=1}^{N} r_{nk}$ and $N = \sum_{k=1}^{K} N_k$. All the parameters have an intuitive meaning, for example $\boldsymbol{\mu}_k$ is calculated by the weighted average of all points $\mathbf{x}_n$ according to the posterior probability r_{nk} that component k was responsible for generating $\mathbf{x}_n$ and π_k is calculated by the average responsibility which that component takes for explaining the data points.

As a side note, these maximum likelihood estimators do not admit a closed form solution. The responsibilities r_{nk} depend on the model parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ by

$$r_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

and the model parameters depend on the r_{nk}. This suggests that an iterative solution can be used. First we choose some initial values for the model parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$. In the *expectation* step, we use those values to calculate the posterior probability r_{nk}. Then, in the *maximization* step, we use these probabilities to get a better estimate of the model parameters. After each iteration, we calculate the log-likelihood value and we stop when the change in log-likelihood is below a threshold.

Some Useful Concepts for EM

In this section, we will review some concepts which will be useful in the analysis of EM algorithm.

Entropy

For a discrete probability distribution p, the entropy is defined as

$$\mathcal{H}(p) = \sum_x -p(x) \log p(x)$$

Distributions which are spread more evenly across many values will have a *relatively* higher entropy than the distributions which are concentrated around a few values. Entropy is a measure of the unpredictability of the state, or equivalently, of its average information content. For example, the entropy of a Bernoulli random variable with parameter μ is $-(\mu \log \mu + (1 - \mu) \log(1 - \mu))$ and it is maximized when $\mu = \frac{1}{2}$.

Jensen's Inequality

If f is a convex function, we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Note that if X is constant we get an equality. Suppose we have $f(x) = x^2$, which is a convex function. Then, using Jensen's Inequality, we have $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, which you may recall from the definition of $\mathrm{Var}(X)$. Moreover, if f is a concave function (e.g. $f(x) = \log x$), we reverse the inequality sign.

KL Divergence

KL divergence measures how one probability distribution is different than the other. For discrete probability distributions p and q, it is defined as

$$KL(p \parallel q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

KL divergence is only defined for every x where $q(x) = 0$ implies $p(x) = 0$. When $p(x)$ is zero, KL divergence is still defined since its limit is still zero. Moreover, KL divergence is not symmetric, i.e. $KL(p \parallel q) \neq KL(q \parallel p)$ in general. Also it becomes zero when $p = q$. Now we will prove that KL divergence is always positive using Jensen's Inequality.

$$KL(p \parallel q) = -\sum_x p(x) \log \left(\frac{q(x)}{p(x)} \right) \geq -\log \sum_x p(x) \frac{q(x)}{p(x)} = 0$$

A General View on EM

Now we will consider a more general interpretation of EM. All observed variables are denoted by $\mathbf{x}$, all latent variables are denoted by $\mathbf{z}$ and all model parameters are denoted by $\boldsymbol{\theta}$. Our aim is to maximize the log-likelihood function $\log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The optimization of $\log p(\mathbf{x}|\boldsymbol{\theta})$ difficult whereas the complete data log-likelihood $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ is easier to optimize. We also introduce a new distribution $q(\mathbf{z})$ over the latent variables. The following equation shows us that for every q, we get the following lower bound of the log-likelihood using Jensen's Inequality:

$$\begin{aligned}\log p(\mathbf{x}|\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} q(\mathbf{z}) \\ &\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} = \sum_{\mathbf{z}} q(\mathbf{z}) p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})\end{aligned}$$

One important question is when this lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ becomes tight. Now we will use another decomposition of the log-likelihood to answer that question.

$$\begin{aligned}\log p(\mathbf{x}|\boldsymbol{\theta}) &= \sum_{\mathbf{z}} q(\mathbf{z}) p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left(\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \frac{q(\mathbf{z})}{q(\mathbf{z})} \right) \\ &= \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} + \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \\ &= \mathcal{L}(q, \boldsymbol{\theta}) + KL(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}))\end{aligned}$$

We know that KL divergence is always positive and becomes zero when $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$. Therefore if we choose $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$, the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ becomes tight. Here is a summary of EM algorithm:

- **E-Step:** Set $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}_{old})$

  In this step, lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the distribution q while holding $\boldsymbol{\theta}$ fixed.

- **M-Step:** Set $\boldsymbol{\theta}_{new} = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}_{old}) \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$

  In this step, lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameters $\boldsymbol{\theta}$ while holding the distribution q fixed.

Mixture of Bernoullis

Now, we introduce the mixture of Bernoullis with EM, which is covered also in the homework. Suppose we have d independent Bernoulli random variables and the joint probability is defined as

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{d} \mu_i^{x_i} (1 - \mu_i)^{1 - x_i}$$

If we have K such distributions and we use a convex combination of them, the resulting mixture distribution is defined as

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)$$

If we look at the mean and covariance of this mixture distribution, we see that

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \quad \text{and} \quad \mathrm{Cov}(\mathbf{x}) = \sum_{k=1}^{K} \pi_k (\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T) - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T$$

and this mixture distribution can capture correlations between the variables unlike the independent Bernoulli random variables whose mean and covariance can be written as

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{and} \quad \mathrm{Cov}(\mathbf{x}) = \mathrm{diag}(\boldsymbol{\mu}(1-\boldsymbol{\mu})) = \boldsymbol{\Sigma}$$

The complete data log-likelihood given N observed points is

$$\log p(\mathbf{x}, \mathbf{z} | \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\log \pi_k + \sum_{i=1}^{d} x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki}) \right)$$

We need to calculate $\mathbb{E}[z_{nk}]$ for the E-Step and it can be calculated using:

$$\mathbb{E}[z_{nk}] = p(z_{nk} = 1 | \mathbf{x}_n) = \frac{\pi_k p(\mathbf{x}_n | \boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n | \boldsymbol{\mu}_j)}$$

Therefore, expected complete data log-likelihood can be written as

$$\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z} | \boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left(\log \pi_k + \sum_{i=1}^{d} x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki}) \right)$$

where we used $r_{nk} = \mathbb{E}[z_{nk}]$.

Now, let's see how to compute $\boldsymbol{\pi}^*$ for the M-Step. Since we are solving a constrained optimization problem, we need to construct the Lagrangian.

$$L(\lambda, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \log \pi_k - \lambda (\sum_{k=1}^{K} \pi_k - 1)$$

Setting derivatives to zero, we find that

$$\pi_k = \frac{\sum_{n=1}^{N} r_{nk}}{\lambda} = \frac{N_k}{N} \quad \text{where} \quad \lambda = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk}$$

The optimal $\boldsymbol{\mu}^*$ and more information can be found in the corresponding homework solution. For an excellent overview of EM algorithm, you can consult chapter 9 of [1] which is mainly used for this tutorial. Another good exposition is [3]. For a gentle introduction with a biomedical perspective, you can check [2]. For a more advanced treatment, you can check chapter 11 of [4].

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[2] Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897, 2008.

[3] Maya R. Gupta and Yihua Chen. Theory and use of the em algorithm. *Found. Trends Signal Process.*, 4(3):223–296, March 2011.

[4] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.