# Lecture-14: Mixture of Gaussians and EM Algorithm

Sameera Bharadwaja H

Sept 24, 2019

## 1 Recap

In the previous lecture, we discussed algorithms for parametric estimation and generative modelling. In generative modelling, we have $(x_1,...,x_n)$ independent and identically distributed (i.i.d.) samples drawn from $p_\theta(x), \theta \subseteq \mathbb{R}^d$. The estimation of $\theta$ can be done using two methods: Maximum Likelihood Estimation (MLE) and Bayesian/ Maximum Aposteriori probability (MAP) method. In MLE, we have seen that the estimator is:

$$\hat{\theta} = \arg\max_\theta \mathscr{L}(\theta) \triangleq \arg\max_\theta p_\theta(x_1,...,x_n) = \arg\max_\theta \prod_{i=1}^{n} p_\theta(x_i) \tag{1}$$

Here, $\mathscr{L}(\theta)$ is called the likelihood function and the last equality follows from the fact that the samples $(x_1,...,x_n)$ are i.i.d. We can also maximize the log-likelihood instead: $\log \mathscr{L}(\theta) = \sum_{i=1}^{n} \log[p_\theta(x_i)]$.

If we take, $p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2}\right)$ where $\theta = \mu$. Then we get:

$$\log \mathscr{L}(\theta) = \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(x_i-\mu)^2}{2}\right] \tag{2}$$

Take derivative w.r.t $\theta$ and equate to zero to obtain: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

## 2 Mixture of Gaussians

A sample likelihood function for a mixture of two Gaussians is given as:

$$\mathscr{L}(\theta) = p_1 \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_1)^2}{2}\right) + (1-p_1)\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_2)^2}{2}\right) \tag{3}$$

Here, $\theta = (p_1, \mu_1, \mu_2)$ and $K$, the number of component distributions in the mixture is 2. A 'brute-force' way to solve the above equation would be to take partial derivatives with each of $p_1$, $\mu_1$ and $\mu_2$ and equate to zero to obtain the required solution that optimizes the likelihood. This is a complicated non-linear equation and further, a generic Gaussian mixture model (GMM) might have a large number of parameters and not just three as shown in the above example ($K = 2$).

Thus, solving the MLE equation in general is very hard. Also, in general it is non-concave. Thus, even if we solve, we don't know if it is a maximum or a minimum; or a global or a local optimum or a saddle point. It is important to find an efficient algorithm to solve such an equation. We shall discuss one such algorithm in the next section.

# 3 Expectation Maximization (EM) algorithm

## 3.1 Background and Motivation

EM algorithm is an efficient way to solve problems with complex likelihood function as seen in Gaussians mixture models. Further note that the EM algorithm is generic and that the component distributions need not be Gaussian. Consider:

$$p_\theta(x) = \sum_y p_\theta(x,y) \tag{4}$$

Taking log, we get:

$$\log\ p_\theta(x) = \log\left(\sum_y p_\theta(x,y)\right) \tag{5}$$

Note that $Y$ in the previous example is a random variable s.t. $P(Y=1) = p_1$ and $P(Y=0) = (1-p_1)$.
Consider $n$ i.i.d samples $x_1, ..., x_n$. We have:

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} \log\ p_\theta(x_i) = \sum_{i=1}^{n} \log\left(\sum_y p_\theta(x_i,y)\right) \tag{6}$$

Due to the presence of summation over y inside the log term, taking partial derivatives and equating to zero, gives a very complicated non-linear system of non-convex functions. Thus, in the following we modify the function such that the new function is easier to optimize. Define:

$$F(Q,\theta) = \sum_{i=1}^{n} \sum_y Q_{iy}\ \log(p_\theta(X = x_i, Y = y)) \tag{7}$$

where, $Q_{iy} = p_\theta(Y = y|X = x_i)$. Observe that $y$'s are not the labels of data samples whereas $x_i$'s are data samples from the actual distribution.

If we had the sample set: $(x_1, y_1), ...(x_n, y_n)$, we can get the log-likelihood as:

$$\log\ p_\theta(X = x_i, Y = y_i) = \log(p_\theta(Y = y_i)p_\theta(X = x_i|Y = y_i)) \tag{8}$$

Summing over $i$ on both sides, we get:

$$\sum_i \log p_\theta(X = x_i, Y = y_i) = \sum_i \log(p_\theta(Y = y_i)p_\theta(X = x_i|Y = y_i)) \tag{9}$$

which is an easier equation to solve than the equation (6) where there is a summation over $y$ inside $\log(.)$. But, we only have $x_i$'s and the $y_i$'s are missing.

To summarize the motivational points for considering EM algorithm:
1. Solve for likelihood equation for more complicated distribution
2. Solve for likelihood when some data is missing (e.g. $y_i$ is missing)

## 3.2 EM Algorithm

EM algorithm is an iterative algorithm. We compute, $(Q^{(1)}, \theta^{(1)}), (Q^{(2)}, \theta^{(2)}), ....$
It has two steps:
1. Expectation step: Get better Q using $Q_{iy}^{(t+1)} = p_{\theta^{(t)}}(Y = y|X = x_i)$
2. Maximization step: Maximize the likelihood given by $\theta^{(t+1)} = \arg\max_\theta \ \ F(Q^{(t+1)}, \theta)$

Due to form of equation (7), maximization step is easy and expectation step is an approximation to MLE.

**Lemma 3.1.** *Define,*

$$G(Q, \theta) = F(Q, \theta) - \sum_{i=1}^{n} \sum_y Q_{iy}\ \log\ Q_{iy} \tag{10}$$

*Then, EM algorithm can be re-written as,*

$$Q^{(t+1)} = \arg\max_Q \ \ G(Q, \theta^{(t)}) \tag{11}$$

$$\theta^{(t+1)} = \arg\max_\theta \ \ G(Q^{(t+1)}, \theta) \tag{12}$$

*Also,*

$$G(Q^{(t+1)}, \theta^{(t)}) = \mathscr{L}(\theta^{(t)}) \tag{13}$$

*where, $\mathscr{L}(\theta)$ is as defined in equation (6)*

*Proof.* Start from equation (10). Fix $Q$ and optimize over $\theta$. The second term of equation (10) is only dependent on $Q$ and independent of $\theta$. So, we can operate only on the first term and doing so, we get:

$$\arg\max_{\theta} \quad G(Q^{(t+1)}, \theta) = \arg\max_{\theta} \quad F(Q^{(t+1)}, \theta) = \theta^{(t+1)} \tag{14}$$

Thus, equation (12) is same as the maximization step.

Observe that $Q$ is a matrix whose elements are $Q_{iy}$. Fix $i$ and observe that rows are probabability mass over $y$ (i.e., PMF of $y$). So the constraint to be applied is: $\sum_y Q_{iy} = 1$ for each $i$ (each row of $Q$ sums to 1). Consider,

$$G(Q, \theta) = F(Q, \theta) - \sum_i \sum_y Q_{iy} \log Q_{iy} \tag{15}$$

From equation (7),

$$G(Q, \theta) = \sum_{i=1}^{n} \sum_y Q_{iy} \ \log(p_\theta(X = x_i, Y = y)) - \sum_i \sum_y Q_{iy} \ \log \ Q_{iy} \tag{16}$$

Simplifying, we get:

$$\begin{aligned} G(Q, \theta) &= \sum_i \sum_y Q_{iy} \ \log \left( \frac{p_\theta(X = x_i, Y = y)}{Q_{iy}} \right) \\ &\leq \sum_i \log \left( \sum_y Q_{iy} \ \frac{p_\theta(X = x_i, Y = y)}{Q_{iy}} \right) \\ &= \sum_i \log \left( p_\theta(X = x_i) \right) \end{aligned}$$

The penultimate step is obtained from the fact that log is a concave function and using Jensen's inequality after noting that convex function is negative of concave function, the inequality reverses and we get $\mathbb{E}[\phi(X)] \leq \phi(\mathbb{E}[X])$ with $\phi(.) = \log(.)$.
The R.H.S. is the log-likelihood function $\mathscr{L}(\theta)$. This means that for any $Q_{iy}$:

$$G(Q, \theta) \leq \mathscr{L}(\theta) \tag{17}$$

Now substitute, $Q_{iy} = p_\theta(Y = y | X = x_i)$ in equation (16). We get,

$$\begin{aligned} G(Q, \theta) = &\sum_{i=1}^{n} \sum_y p_\theta(Y = y | X = x_i) \ \log(p_\theta(Y = y, X = x_i)) \\ &- \sum_{i=1}^{n} \sum_y p_\theta(Y = y | X = x_i) \ \log(p_\theta(Y = y | X = x_i)) \end{aligned} \tag{18}$$

$$
\begin{aligned}
G(Q,\theta) &= \sum_i \sum_y p_\theta(Y=y|X=x_i)\ \log\left(\frac{p_\theta(Y=y,X=x_i)}{p_\theta(Y=y|X=x_i)}\right) \\
&= \sum_i \sum_y p_\theta(Y=y|X=x_i)\ \log\left(p_\theta(X=x_i)\right) \\
&= \sum_i \log\ p_\theta(X=x_i) \sum_y p_\theta(Y=y|X=x_i) \\
&= \sum_i \log\ p_\theta(X=x_i) \\
&= \mathscr{L}(\theta)
\end{aligned}
$$

Bayes theorem is applied to get the second step in the above set of equations and the fact that probabilities sum to 1 over its domain to get step 4 from step 3.
Thus, we get: $G(Q,\theta) = \mathscr{L}(\theta)$ when $Q_{iy} = p_\theta(Y=y|X=x_i)$. Combining the above results, we can convey the following:

$$
\arg\max_Q \quad G(Q,\theta^{(t)}) = Q^{(t+1)} = P_{\theta^{(t)}}(Y=y|X=x_i) \tag{19}
$$

which is same as the Expectation step. This proves equation (11). □

**Theorem 3.2.**

$$
\mathscr{L}(\theta^{(t+1)}) \geq \mathscr{L}(\theta^{(t)}),\ \ \forall\, t = 1,2,... \tag{20}
$$

*Proof.* We know that, $\arg\max_Q G(Q,\theta^{(t+1)}) = Q^{(t+2)}$.
Since, $G(Q,\theta) = \mathscr{L}(\theta)$,

$$
\begin{aligned}
\mathscr{L}(\theta^{(t+1)}) &= G(Q^{(t+2)},\theta^{(t+1)}) \\
&\geq G(Q^{(t+1)},\theta^{(t+1)}) \\
&\geq G(Q^{(t+1)},\theta^{(t)}) \\
&= \mathscr{L}(\theta^{(t)})
\end{aligned}
$$

where, the second step follows because $Q^{(t+2)}$ is optimal and further by using Lemma 3.1 in subsequent steps to arrive at the required result. □

Consider the following: $F(Q^{(t+1)},\theta^{(t+1)}) = \max_Q F(Q^{(t+1)},\theta) \triangleq h(\theta^{(t)},\theta^{(t+1)})$ because $Q^{(t+1)}$ depends on $\theta^{(t)}$.

**Theorem 3.3.** *If $h$ is continuous in $(\theta^{(t)},\theta^{(t+1)})$, then all the limit points of $\theta^{(t)}$ are stationary points of $\mathscr{L}(\theta)$ and $\mathscr{L}(\theta^{(t)})$ converges monotonically to $\mathscr{L}(\hat{\theta})$ where $\hat{\theta}$ is the stationary point of $\mathscr{L}(\theta)$.*

# 4 Non-Parametric Regression

Consider $n$ i.i.d samples $(x_1, y_1), ..., (x_n, y_n)$ drawn from an unknown distribution $D$ such that the relationship between $x_i$ and $y_i$ can be expressed as:

$$y_i = m_o(x_i) + \varepsilon_i \tag{21}$$

where, $m_o : \mathscr{X} \mapsto \mathbb{R}$ is the unknown function with input $x_i \in \mathscr{X}$ and $y_i \in \mathbb{R}$. Further, $\varepsilon_i$ is the i.i.d noise with unknown distribution and zero mean i.e., $\mathbb{E}[\varepsilon_i] = 0$. Also, $\varepsilon_i$ and $x_i$ are independent. We want to estimate function $m_o$ from these samples.

Unlike the classification and parametric estimation problems seen so far, this is an infinite dimensional estimation problem.

**Definition 4.1 (Mean Square Error (MSE)).** For input $X$, output $Y$ and estimate of $m_o(.)$ denoted by $\hat{m}(.)$, the mean square error is defined as $\mathbb{E}\left[(\hat{m}(X) - Y)^2\right]$.

*Remark* 4.2. The optimum estimator in MSE sense is $m^*(x) = \mathbb{E}[Y \mid X = x]$.

*Note* 4.3. This estimate however can not be directly computed as the distribution $P(Y \mid X)$ is unknown.

MSE can be written:

$$\mathbb{E}\left[(Y - \hat{m}(X))^2\right] = \mathbb{E}\left[[(Y - m_o(X)) + (m_o(X) - \mathbb{E}[\hat{m}(X)]) + (\mathbb{E}[\hat{m}(X)] - \hat{m}(X))]^2\right] \tag{22}$$

Expanding the squares yields,

$$\begin{aligned}\mathbb{E}\left[(Y - \hat{m}(X))^2\right] = \mathbb{E}\left[(Y - m_o(X))^2\right] + \mathbb{E}\left[(m_o(X) - \mathbb{E}[\hat{m}(X)])^2\right] + \\ \mathbb{E}\left[(\mathbb{E}[\hat{m}(X)] - \hat{m}(X))^2\right] + Cross - terms\end{aligned} \tag{23}$$

The cross terms can be shown to be zero. The first term on the RHS is $\mathbb{E}[{\varepsilon_i}^2]$, the minimum error achievable. The second term is the expectation over the square of the bias, $\mathbb{E}[bias^2]$ and the third term is the variance of estimated function, $var(\hat{m}(X))$. Bias of the algorithm is its affinity toward choosing a hypothesis from H or the approximation error. The variance can be reduced by increasing the number of samples. The bias can be reduced by choosing a richer hypothesis class. We can also notice that reducing the bias might lead to increase in the variance because to reduce the bias we need to enlarge the hypothesis class to estimate $\hat{m}_o$ which means estimating more parameters. This is termed as "Bias-Variance trade-off".

# 5 In the next class

We will see the following algorithms for non-parametric regression/ estimation
1. K-NN
2. Random Forest
3. Kernel Method (RKHS and others)

We will also obtain minimax bounds.

*Remark* 5.1. Another general paradigm for both classification and regression problems are neural networks which we will study later.

# 6 References

1 Shai Shalev-Shwartz and Shai Ben-David, "Understanding Machine Learning: From Theory to Algorithms", Chapter 24: Generative models.