CS 475 Machine Learning: Lecture 12
EM Algorithm

Prof. Mark Dredze

1 EM Algorithm

In unsupervised learning, we have only the data $\mathbf{X}$. We want to write a function of this data (likelihood):

$$p(\mathbf{X}|\theta)$$

where θ are the model parameters.

As we saw with GMMs, maximizing this likelihood is very complicated. Instead, its simpler to explicitly write a joint probability of the latent variables:

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

We can write the joint by summing over all possible $\mathbf{Z}$. In many cases, it is much easier to maximize the joint likelihood. In fact, we've maximized joint likelihood several times in class.

We call this the complete-data likelihood function. We pretend we have all of the data ($\mathbf{Z}$), which of course makes learning much easier.

Allow me to introduce a new distribution $q(\mathbf{Z})$ over the latent variables. This is just some distribution over the values we expect the latent variables to take. The form of the distribution does not matter.

Let's write the log-likelihood. We will take it on faith for a moment that the log-likelihood can be broken down into:

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \mathrm{KL}(q||p)$$

where we define

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\mathrm{KL}(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

While this is the KL divergence, notice that the form is slightly changed. Normally, it would be:

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right\}$$

flipping the fraction. Since the fraction is inside a log, when we flip it we get the $-$ sign.

What have we written? Notice the differences between the two equations.

- They differ in terms of a sign. As the first one gets larger, the second one should get smaller.

- $\mathcal{L}$ contains the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$, while KL contains the conditional. The second term forces our distribution $q(\mathbf{Z})$ to be similar to the posterior of $\mathbf{Z}$. Recall that we said the joint is easier to use, which means we've made progress.

This will be an important decomposition for our understanding of EM. But first, let's understand where it comes from. Let's work backwards to prove it.

First, note that we can use the product rule to produce:

$$p(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta)$$

We can substitute this in for the numerator of $\mathcal{L}$.

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta)}{q(\mathbf{Z})} \right\}$$

Let's substitute both $\mathcal{L}$ and KL into the decomposition of $\log p(\mathbf{X}|\theta)$:

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta)}{q(\mathbf{Z})} \right\} - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

Move the sum of $\mathbf{Z}$ and $q(\mathbf{Z})$ out:

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \left[\log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta)}{q(\mathbf{Z})} \right\} - \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\} \right]$$

Distribute the logs:

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \left[\log \left\{ p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta) \right\} - \log q(\mathbf{Z}) - \log p(\mathbf{Z}|\mathbf{X}, \theta) + \log q(\mathbf{Z}) \right]$$

Distribute the remaining log and cancel the $\log q(\mathbf{Z})$.

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \left[\log p(\mathbf{Z}|\mathbf{X}, \theta) + \log p(\mathbf{X}|\theta) - \log p(\mathbf{Z}|\mathbf{X}, \theta) \right]$$

Cancel the remaining log and move terms outside of the $\sum_{\mathbf{Z}}$:

$$\log p(\mathbf{X}|\theta) \sum_{\mathbf{Z}} q(\mathbf{Z})$$

Remember that $q(\mathbf{Z})$ is a probability distribution which sums to 1, so we have just $\log p(\mathbf{X}|\theta)$, which proves the decomposition.

1.1 Understanding EM

Now that we know this decomposition of the log-likelihood holds, what does it mean?

Recall that $\text{KL}(q||p) \geq 0$, and is only 0 if $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$. Therefore, it must be that $\mathcal{L}(q, \theta) \leq \log p(\mathbf{X}|\theta)$ and is only equal when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$. Therefore, we say $\mathcal{L}$ is a (sort of) lower bound on the log likelihood. This means if we maximize $\mathcal{L}$ we maximize the log-likelihood.

Recall that EM is a two step iterative process for finding maximum likelihood. Let's use this decomposition to understand EM.

Suppose we have the current value of the parameters as θ^{old}. In the E step, we maximize $\mathcal{L}(q, \theta^{old})$ with respect to $q(\mathbf{Z})$ by keeping θ^{old} constant. What is the solution to this maximization? Notice that $\log p(\mathbf{X}|\theta^{old})$ does not depend on $q(Z)$ so we can only maximize $\mathcal{L}$ by driving the KL to 0. When is the KL = 0? When $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$. Therefore, the goal of the E step is to make $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{old})$, or the distribution of $\mathbf{Z}$ should be equal to the expected $\mathbf{Z}$ given the data and my parameters θ^{old}. Thinking back to GMMs, this means we want our hidden variables (the clusters) to equal exactly what we'd expect them to be given $\mathbf{X}$ and the current parameters of our model θ^{old} (where the clusters are located.) If KL is 0, then the lower bound becomes equal to the log-likelihood.

In the M step we do the opposite: hold $q(\mathbf{Z})$ fixed and maximize $\mathcal{L}$ with respect to θ to give θ^{new}. This will cause $\mathcal{L}$ to increase (if possible). Since KL can only increase (it cannot be negative), this increases the log-likelihood function. This means we will have a new posterior $p(\mathbf{Z}|\mathbf{X}, \theta^{new})$, and since $q(\mathbf{Z})$ is fixed, the KL will increase. Therefore, the increase in the log-likelihood is greater than the increase in $\mathcal{L}$.

Therefore, in both the E step and the M step we are increasing the likelihood. We cannot increase the likelihood further when the maximum value for θ is θ^{old}. Therefore, EM is guaranteed to maximize the likelihood function until the model converges. This is a very powerful result: we have a general procedure for maximizing likelihood for any hidden variable model.

Notice that $\log p(\mathbf{X}|\theta^{old})$ *does not depend on* $q(Z)$

This is a point worth repeating. We can evaluate the likelihood of $\mathbf{X}$ without knowing anything about $q(\mathbf{Z})$. First, notice that $q(\mathbf{Z})$ doesn't show up in the definition of $p(\mathbf{X}|\theta)$. Instead, if we need the probability of $\mathbf{Z}$ we use the model parameters θ. In that case, why do we bother having a $q(\mathbf{Z})$ at all? Think of $q(\mathbf{Z})$ as the expectation of the $\mathbf{Z}$ variables given the current model parameters. If we stopped the model right now and asked, "what do you think $\mathbf{Z}$ should be?" it would answer $q(\mathbf{Z})$. In that sense, $q(\mathbf{Z})$ is just $P(\mathbf{Z}|\mathbf{X}, \theta)$. However, one of the tricks of EM is to write this variable out explicitly and then *not* update it at the same time as the model. It would be the equivalent of asking the model for $\mathbf{Z}$, writing that down, updating the model, and then asking again "what do you think $\mathbf{Z}$ should be?" We'd get a different answer. That difference is what increases the KL. We can think of the E-step, minimizing KL, as fixing this discrepancy; throwing away our old $q(\mathbf{Z})$ for a new one. When we compute the data likelihood $\log p(\mathbf{X}|\theta)$, we don't care what the model had said previously, which we stored in $q(\mathbf{Z})$. In other words, we don't use $q(\mathbf{Z})$ to compute $\log p(\mathbf{X}|\theta)$. If that is true, then by fixing θ we aren't changing $\log p(\mathbf{X}|\theta)$. If that remains the same, then as the KL decreases, $\mathcal{L}(q, \theta)$ must increase by the same amount since their sum is fixed.

1.2 Cross Entropy

We have previously learned about entropy, which is defined as

$$H(x) = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \ .$$

Entropy indicates the uncertainty in a random variable. A related measure is cross entropy, which measures the average number of bits (how much information) is needed to identify an event from a set of possible events if the coding scheme is based on distribution q instead of the true distribution p. Cross entropy is defined as:

$$H(p,q) = -\sum_i p(x_i) \log q(x_i) \ .$$

When $p = q$, the this reduces to the entropy of p. Disagreements between p and q only add to the randomness.

2 Examining EM

Let's examine what happens to $\mathcal{L}$ after an E step. We know that the E step makes $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{old})$, so if we substitute this into the equation for $\mathcal{L}$ we get:

$$\mathcal{L}(q,\theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \log p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \log p(\mathbf{Z}|\mathbf{X},\theta^{old})$$

We've just distributed the log and $q(\mathbf{Z})$, then substituted $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ for $q(\mathbf{Z})$.

What have we written?
The first term: this is the negative cross entropy between $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ and $p(\mathbf{X}, \mathbf{Z}|\theta)$. p is $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ and q is $p(\mathbf{X}, \mathbf{Z}|\theta)$. This means that maximizing this term makes p and q similar and will reduce to the entropy of p.

The second term: The entropy of $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$. This is fixed since it doesn't depend on q or θ^{old}.

Since θ^{old} is not changing, this is just a constant. Only the first term is changing with θ: $p(\mathbf{X}, \mathbf{Z}|\theta)$. We call this the $\mathcal{Q}$ function.

$$\mathcal{Q}(\theta, \theta^{old}) + \text{constant}$$

where

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \log p(\mathbf{X},\mathbf{Z}|\theta)$$

You can think of $\mathcal{Q}$ as the expectation of the complete-data log likelihood evaluated for some parameter θ based on our current belief (θ^{old}) about how $\mathbf{Z}$ is distributed.

Therefore, we can see clearly that as we maximize $\mathcal{L}$ we are actually maximizing the joint likelihood $p(\mathbf{X}, \mathbf{Z}|\theta)$ (likelihood of the complete data).

Recall in the beginning we said that maximizing the joint likelihood would be easier then the complex partial likelihood why? The log-likelihood of the data is:

$$\log p(\mathbf{X}|\theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)$$

The log cannot move past the sum. This seems very tricky.

However, by pursuing EM, we see that the complete data likelihood $p(\mathbf{X}, \mathbf{Z}|\theta)$ appears only as $\log p(\mathbf{X}, \mathbf{Z}|\theta)$. Why is this good? We often write $p(\mathbf{X}, \mathbf{Z}|\theta)$ using an exponential model. Taking the log of such a model is pretty easy.

To summarize, we have shown that if we rewrite the likelihood in terms of the complete data likelihood, we can maximize the lower bound $\mathcal{L}$ using a two step procedure.
The E step causes the hidden variables $\mathbf{Z}$ to match the *expectation* of our model according to θ^{old}.
The M step maximizes $\mathcal{L}$ with respect to θ, hence the *maximization*.
Finally, we see that this approach is advantageous because we only have to write $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ inside a log term, which is easy to do.

3 GMMs with EM

Let's see how EM is applied in practice, by again looking at GMMs.

Quickly review GMM model.

We begin by writing the complete data log likelihood:

$$\log p(\mathbf{X}|\theta) = \log\{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)\} \ .$$

In the case of GMMs, we know that:

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \mathbf{\Sigma}_k)^{z_k} \ .$$

Therefore, we can write the likelihood function as:

$$p(\mathbf{X}, \mathbf{Z}|\theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \mathbf{\Sigma}_k)^{z_{nk}} \ .$$

where $\theta = \{\mu_k, \mathbf{\Sigma}_k, \pi_k\}_{k=1}^{K}$.

Taking the log:

$$\log p(\mathbf{X}, \mathbf{Z}|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk}\{\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \mathbf{\Sigma}_k)\} \ .$$

Note that the log appears inside the summation over K, i.e., insider the summation over the hidden variables $\mathbf{Z}$.

Maximizing this complete data log likelihood, we obtain can obtain solutions for the mean and covariance of each Gaussian. However, in this case, because we observe z_{nk} this either places or removes an example from the cluster. Therefore, the solution for the mean and variance is identical to that for a standard Gaussian.

We get a similar result for π:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}$$

Since we "observe" z, this summation is simple: just count!

So far we've just shown how to maximize the complete data likelihood function, which is the M-step of EM. The E-step requires us to compute the posterior distribution for $\mathbf{Z}$. Using Bayes theorem and equations for $p(z)$ and $p(x|z)$, which we had last time, we can get:

$$p(\mathbf{Z}|\mathbf{X}, \theta) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k)]^{z_{nk}} \ .$$

Note that since this factors over n, each z_n is independent. According to this posterior, what is the expected value for z_{nk}? We can show that this is the responsibility of k for the example $\mathbf{x}_n$: $\gamma(z_{nk})$. Therefore, we can write the expected value of the complete data log likelihood function as:

$$\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\theta)] = \gamma(z_{nk})\{\log \pi_k + \log \mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k)\}.$$

We can maximize this using our iterative procedure. We first fix the parameters (θ) and compute the responsibilities γ. The responsibilities are the hidden variables, so this is the E step. Next, in the M step we fix the responsibilities and compute the new parameters θ, which will maximize the log likelihood. Note that the E step is minimizing the KL (finding the best responsibilities given the parameters) and the M step is maximizing the complete data likelihood, which lower bounds the data likelihood. We've shown that will maximize the data likelihood, so we have a procedure for finding the max solution (not optimal) for GMM.