Machine Learning (CS 5350/CS 6350) *29 Oct 2009*

Expectation Maximization

We want to do maximum likelihood in a model $p(x \mid \theta)$ that contains *hidden variables* z, so:

$$p(x \mid \theta) = \sum_z p(x, z \mid \theta)$$

In GMM, x is the data, θ is the means and variances, and z is the cluster assignments.

Our problem is that if we get a data set $x_{1:N}$ and take a product, we cannot move the log inside the sum:

$$\log p(x_{1:N} \mid \theta) = \sum_n \log \sum_z p(x, z \mid \theta)$$

What we're going to do is to find a *lower bound* to the $\log \sum$ and then maximize *that* instead. If the lower bound is tight we can keep pushing it up, then the true log likelihood will also go up.

Quick foray into convexity...

Recall from HW1 that a function $f : \mathbb{R}^D \to \mathbb{R}$ is *convex* if for all $x_1, x_2 \in \mathbb{R}^D$ and all $\lambda \in [0, 1]$:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$$

Pictorially, this means that all chords lie above the function (i.e., the function looks like a "bowl." As you proved in HW1, a function is convex if its second derivative is non-negative. We can use this to show that the function $f(x) = -\log x$ is convex (on $[0, \infty)$) because the second derivative is x^{-2}.

The second thing we need is *Jensen's Inequality:*

Theorem: Let f be convex. If $x_1, x_2, \ldots, x_N$ are in the domain of f and $\lambda_1, \lambda_2, \ldots, \lambda_N \geq 0$ with $\sum_n \lambda_n = 1$, then: $f\left(\sum_n \lambda_n x_n\right) \leq \sum_n \lambda_n f(x_n)$.

Proof is by induction on N. (Sketch:) For $N = 2$ this is just the definition of convexity. Otherwise, we do an inductive step by grouping the $N + 1$ data points into one group of N and one group of 1. We apply the inductive hypothesis to the N and then use the definition of convexity to merge the two.

(Jensen's inequality also applies in most cases for integrals, but we'll not prove this here.)

Putting these two facts together, we get:

$$\log \sum_n \lambda_n x_n \geq \sum_n \lambda_n \log x_n$$

So what does this have to do with probabilistic models with hidden variables? Well, we had:

$$\log \sum_z p(x, z \mid \theta) = \log \sum_z p(z \mid \theta) p(x \mid z, \theta)$$

Here, $p(z \mid \theta)$ plays the role of λ in Jensen's inquality. We need to be a bit more formal, but this is the basic idea.

Formally, we're going to do our standard trick of multiplying by 1 and rearranging terms. In this case, we're multiplying by 1 by first multiplying, then dividing, by some quantity $q(z)$ for some distribution q. For now,

just let q be *any* distribution; we'll figure out what it should be optimally next.

$$\begin{aligned}
\log p(x \mid \theta) &= \log \sum_z p(x, z \mid \theta) \\
&= \log \sum_z q(z) \frac{1}{q(z)} p(x, z \mid \theta) \\
&\geq \sum_z q(z) \log \left[\frac{1}{q(z)} p(x, z \mid \theta) \right] \\
&= \sum_z q(z) \log p(x, z \mid \theta) - \sum_z q(z) \log q(z)
\end{aligned}$$

This gives us the following lower-bound on the data likelihood:

$$\log p(x \mid \theta) \geq \underbrace{\mathbb{E}_{z \sim q}\Big[\log p(x, z \mid \theta) \Big]}_{\text{expected complete log-lik}} + \underbrace{\mathbb{E}_{z \sim q}\Big[\log q(z \mid \theta) \Big]}_{\text{entropy of } q}$$

The entropy term is independenty of θ so we can ignore it when trying to maximize. Note that the above holds for *any* q (well, it has to have the same support as $p \ldots$); the remaining question is: which q should we use?

We'll do this by "guessing and getting lucky." Let $q(z) = p(z \mid x, \theta)$. Now, we get:

$$\begin{aligned}
\sum_z q(z) \log \left[\frac{1}{q(z)} p(x, z \mid \theta) \right] &= \sum_z p(z \mid x, \theta) \log \frac{p(x, z \mid \theta)}{p(z \mid x, \theta)} \\
&= \sum_z p(z \mid x, \theta) \log \frac{p(z \mid x, \theta) p(x \mid \theta)}{p(z \mid x, \theta)} \\
&= \sum_z p(z \mid x, \theta) \log p(x \mid \theta) \\
&= \Big[\log p(x \mid \theta) \Big] \sum_z p(z \mid x, \theta) \\
&= \log p(x \mid \theta)
\end{aligned}$$

So with this choice of z, the inequality from above becomes an equality!

This gives us:

$$\log p(x \mid \theta) \geq \underbrace{\mathbb{E}_{z \sim p(\cdot \mid x, \theta)}\Big[\log p(x, z \mid \theta) \Big]}_{\text{expected complete log-lik}} + \underbrace{\mathbb{E}_{z \sim p(\cdot \mid x, \theta)}\Big[\log p(z \mid x, \theta) \Big]}_{\text{entropy of } p(z \mid x, \theta)}$$

This leads to the expectation maximization algorithm:

- Initialize $t = 0$
- Initialize θ^t somehow
- Repeat until convergence of θ:
  - E-step: Compute $z = \sum_z z \times p(z \mid x, \theta^t)$
  - M-step: Maximize $\theta^{t+1} = \max_{\hat{\theta}} \mathbb{E}_z p(x, z \mid \hat{\theta})$
  - Set $t \leftarrow t + 1$