EXPECTATION MAXIMIZATION

JASON PACHECO
4/13/11

EM IN GENERAL

Imagine a probabilistic model with

- X: observations
- θ: Parameters
- $p(X|\theta)$: Likelihood function.

We introduce "hidden" variables Z and by the Law of Total Probability can write

$$p(X|\theta) = \sum_Z p(X, Z|\theta)$$

"Complete-data Likelihood"

Problem: Learn parameters θ by ML

$$\theta_{ML} = \arg\max_{\theta} \sum_Z p(X, Z|\theta)$$

but likelihood function is intractable (e.g. sum $\sum_Z$ is over exponentially many terms).

Lower Bounding $p(x|\theta)$:

We use Jensen's Inequality

$$\log \mathbb{E}[f(x)] \geq \mathbb{E}[\log f(x)]$$

and derive a bound on the likelihood

$$\begin{aligned}
\log p(x|\theta) &= \log \textstyle\sum_z p(x, z|\theta) \\
&= \log \textstyle\sum_z q(z) \frac{p(x,z|\theta)}{q(z)} \\
&= \log \mathbb{E}_q\left[\frac{p(x,z|\theta)}{q(z)}\right] \\
&\geq \mathbb{E}_q\left[\log \frac{p(x,z|\theta)}{q(z)}\right] \\
&= \mathbb{E}_q[\log p(x,z|\theta)] + H(q)
\end{aligned}$$

The bound is "tight" iff $q(z) = p(z|x,\theta)$ (not proven here).

Key Idea:

Given a lower bound we can do ML as,

$$\max_\theta \log p(x|\theta) \geq \max_\theta \mathbb{E}_q[\log p(x,z|\theta)]$$

EM - STATEMENT:

Initialize parameter estimates $\theta^{(0)}$. Iterate the following until "Convergence"

E-STEP:

1) Improve bound by setting $q^{(i+1)}(z) = p(z \mid x, \theta^{(i)})$
2) Compute bound $\mathbb{E}_{q^{(i+1)}}[\log p(x, z \mid \theta)]$

↳ **NOTE:** Function of θ.

M-STEP:

Update parameter estimates

$$\theta^{(i+1)} = \arg\max_{\theta} \mathbb{E}_{q^{(i+1)}}[\log p(x, z \mid \theta)]$$

CONDITIONS FOR APPLYING EM:

1) Likelihood function $p(x \mid \theta)$ is intractable (otherwise we could just do exact ML)
2) Complete-data likelihood $p(x, z \mid \theta)$ is tractable.

e.g. (EM for Gaussian Mixture Models)

The GMM:

Data X_i drawn iid from K ~~clusters.~~ Gaussian distributions. Unknown assignments $z_i \in \{1, \dots, K\}$

Complete-Data Likelihood:

$$p(x_i, z_i \mid \mu, \Sigma) = \sum_{k=1}^{K} \mathbb{1}(z_i = k)\, \pi_k\, N(x_i \mid \mu_k, \Sigma_k)$$

where $p(z_i = k) \triangleq \pi_k$

Likelihood Function:

$$p(x \mid \mu, \Sigma) = \underbrace{\prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k\, N(x_i \mid \mu_k, \Sigma_k)}_{\substack{\text{INTRACTABLE!} \\ K^N \text{ terms}}}$$

EM for GMMs:

E-STEP:

STEP 1) Set $q(z_i = k) = p(z_i = k \mid x_i, \mu, \Sigma)$

$$p(z_i = k \mid x_i, \mu, \Sigma) = \frac{p(z_i = k, x_i \mid \mu, \Sigma)}{p(x_i \mid \mu, \Sigma)}$$

$$= \frac{\pi_k\, N(x_i \mid \mu_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l\, N(x_i \mid \mu_l, \Sigma_l)}$$

STEP 2) Compute updated bound

$$\mathbb{E}_q[\log p(X, Z|\theta)] = \cdots$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} q(z_i = k) \left\{ \log \pi_k + \log N(x_i | \mu_k, \Sigma_k) \right\}$$

M-STEP:

STEP 3) Maximize w.r.t μ, Σ

$$\mu_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} q(z_i = k)\, x_i$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} q(z_i = k)(x_i - \mu_k^{new})^2$$

STEP 4) Repeat until Convergence...

NOTE: The form of μ, Σ updates look like sample mean/variance weighted by probability of assignment to each cluster. Many times the term "soft" updates is used.