Missing Data and the EM Algorithm

Guy Lebanon

In many cases the observed data contains missing values i.e. $X^{(1)},\ldots,X^{(n)} \overset{\text{iid}}{\sim} p$ where $X^{(i)}$ can be partitioned to two vectors $X^{(i)} = (Y^{(i)}, Z^{(i)})$ where $Y^{(i)}$ is observed and $Z^{(i)}$ is missing. Note that the dimensionality of the vectors $Y^{(i)}$ and $Z^{(i)}$ may depend on i but their sum is always d. For example we may have $Y^{(1)} = X_2^{(1)}, Z^{(1)} = (X_1^{(1)}, X_3^{(1)})$ but $Y^{(2)} = (X_1^{(2)}, X_2^{(2)}), Z^{(1)} = X_3^{(1)}$. In this case the likelihood $\sum \log p_\theta(X^{(i)})$ cannot be computed or maximized. A common alternative is to maximize the likelihood of the observed data

$$\hat{\theta} = \arg\max_\theta p_\theta(\text{observed data}) = \arg\max_\theta \sum_{i=1}^{d} \log p_\theta(Y^{(i)}) = \arg\max_\theta \sum_{i=1}^{d} \log \sum_{Z^{(i)}} p_\theta(Y^{(i)}, Z^{(i)}) \tag{1}$$

where the sum over $Z^{(i)}$ is potentially multidimensional (over all possible values of the missing entries) and is an integral when $Z^{(i)}$ is continuous. In many cases the summation over $Z^{(i)}$ above is intractable. This is especially true when the amount of missing data grows since the number of terms in the sum grows exponentially with the dimensionality of $Z^{(i)}$.

The expectation maximization (EM) algorithm maximizes instead a lower bound on the likelihood above, constructed to be tight at the current guess $\theta^{(t)}$. Repeatedly constructing such bounds and maximizing them converges to a local maximum, often at a much lower computational cost than gradient descent for (1). The EM algorithm is based on maximizing the following bound on the likelihood of the observed data

$$\ell(\theta) = \sum_{i=1}^{n} \log \sum_{Z^{(i)}} p_\theta(Y^{(i)}, Z^{(i)}) = \sum_{i=1}^{n} \log \sum_{Z^{(i)}} q_i(Z^{(i)}) \frac{p_\theta(Y^{(i)}, Z^{(i)})}{q_i(Z^{(i)})} = \sum_{i=1}^{n} \log \mathsf{E}_{q_i} \left(\frac{p_\theta(Y^{(i)}, Z^{(i)})}{q_i(Z^{(i)})} \right) \tag{2}$$

$$\geq \sum_{i=1}^{n} \mathsf{E} \left(\log \frac{p_\theta(Y^{(i)}, Z^{(i)})}{q_i(Z^{(i)})} \right) = \sum_{i=1}^{n} \sum_{Z^{(i)}} q_i(Z^{(i)}) \log \frac{p_\theta(Y^{(i)}, Z^{(i)})}{q_i(Z^{(i)})} \tag{3}$$

(q_i are nonzero distributions) where we used Jensen's inequality applied to the convex $f(x) = -\log x$

Proposition 1. *For a RV X and a convex function f we have $\mathsf{E} f(X) \geq f(\mathsf{E} X)$. Moreover, if f is strictly convex, equality holds iff X is degenerate i.e. i.e. $P(X = \mathsf{E} X) = 1$.*

Note that the denominator does not depend on θ and therefore can be removed in maximization over θ. Above, we actually have a parameterized family of bounds - one bound for each selection of the distributions $q_1,\ldots,q_n$. Recall that Jensen's inequality is equality for deterministic RV and therefore the selection

$$q_i(Z^{(i)}) \propto p_{\theta'}(Y^{(i)}, Z^{(i)}) \quad \Rightarrow \quad q_i(Z^{(i)}) = \frac{p_{\theta'}(Y^{(i)}, Z^{(i)})}{\sum_{Z^{(i)}} p_{\theta'}(Y^{(i)}, Z^{(i)})} = p_{\theta'}(Z^{(i)}|Y^{(i)})$$

would yield a bound with equality at θ'. The algorithm iterates between the following steps to convergence.

E step: compute the bound on the observed likelihood

$$Q(\theta, \theta^{(t)}) \overset{\text{def}}{=} \sum_{i=1}^{n} \sum_{Z^{(i)}} p_{\theta^{(t)}}(Z^{(i)}|Y^{(i)}) \log p_\theta(Y^{(i)}, Z^{(i)}) = \sum_{i=1}^{n} \mathsf{E}_{p_{\theta^{(t)}}} \left(\log p_\theta(Y^{(i)}, Z^{(i)})|Y^{(i)} \right)$$

M step: maximize the bound to update new value $\theta^{(t+1)}$

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

The fact that each iteration in the EM algorithm increases the likelihood may be seen by

$$\ell(\theta^{(t+1)}) \geq \sum_{i=1}^{n} \sum_{Z^{(i)}} p_{\theta^{(t)}}(Z^{(i)}|Y^{(i)}) \log p_{\theta^{(t+1)}}(Y^{(i)}, Z^{(i)}) \geq \sum_{i=1}^{n} \sum_{Z^{(i)}} p_{\theta^{(t)}}(Z^{(i)}|Y^{(i)}) \log p_{\theta^{(t)}}(Y^{(i)}, Z^{(i)}) = \ell(\theta^{(t)})$$

where the first inequality follows from Jensen's inequality (for the specified $q_i = p_{\theta^{(t)}}(Z^{(i)}|Y^{(i)})$), the second from the maximization step in EM, and the equality follows from the tightness of the bound at $\theta^{(t)}$.

Clustering

In clustering the task is to partition a dataset $Y^{(1)}, \ldots, Y^{(n)} \in \mathbb{R}^d$ into K disjoint sets so that each set has a spatially coherent set of points (we denote data here using Y rather than X for consistency with the rest of this note). Note that this is an unsupervised task i.e., labels are not available during the training phase.

The most well-known clustering technique is k-means: start by randomly initializing the cluster centroids $\mu_k^{(0)} \in \mathbb{R}^d, k = 1 \ldots, K$, and follow by iterating over the following two stages to convergence: (i) assign each $Y^{(i)}$ to a cluster corresponding to the nearest centroid among $\mu_1^{(t)}, \ldots, \mu_k^{(t)}$, (ii) update the cluster centroids based on the cluster membership obtained in (i) i.e.,

$$\mu_k^{(t+1)} = \text{average}(\{Y^{(i)} : \|Y^{(i)} - \mu_k^{(t)}\| = \min_{k'=1,\ldots,K} \|Y^{(i)} - \mu_{k'}^{(t)}\|\}).$$

A better performing clustering technique is EM for Gaussian mixture model. It is similar to k-means but differs in that $Y^{(i)}$ are assigned to each cluster with some probability (soft membership) rather than assigned with complete certainty to one cluster (hard membership) as k-means does. The Gaussian mixture model assumes the following generative model for our data

$$Y \sim p_\theta(Y) = \sum_{j=1}^{k} p_\theta(Z = j) p_\theta(Y|Z = j) = \sum_{j=1}^{K} \pi_j N(Y\,; \mu_j, \Sigma_j)$$

where Z is a hidden variable representing the Gaussian generating Y and $\pi_j = p(Z = j)$. In general, the unknown parameter θ contains μ_k, Σ_k, π_k for $k = 1, \ldots, K$ (but in some special case may contain only μ, Σ assuming π is known, or only μ assuming Σ, π are known). Once the parameter θ is estimated by maximizing the likelihood of the observed data we can cluster by assigning each $Y^{(i)}$ to the Gaussian most likely to have generated it. The likelihood is $\ell(\theta) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} \pi_j N(Y^{(i)}\,; \mu_j, \Sigma_j)$ and the corresponding EM is

$$\text{E Step:} \quad Q(\theta, \theta^{(t)}) = Q((\pi, \mu, \Sigma), (\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)})) = \sum_{i=1}^{n} \sum_{j=1}^{k} F_{ij}^{(t)} \log \pi_j N(Y^{(i)}\,; \mu_j, \Sigma_j)$$

$$F_{ij}^{(t)} = p_{\theta^{(t)}}(Z^{(i)} = j|Y^{(i)}) = \frac{N(Y^{(i)}\,; \mu_j^{(t)}, \Sigma_j^{(t)}) \pi_j^{(t)}}{\sum_{j'=1}^{K} N(Y^{(i)}\,; \mu_{j'}^{(t)}, \Sigma_{j'}^{(t)}) \pi_{j'}^{(t)}}$$

$$\text{M Step:} \quad \theta^{(t+1)} = (\pi^{(t+1)}, \mu^{(t+1)}, \Sigma^{(t+1)}) = \arg\max_{\pi,\mu,\Sigma} Q((\pi, \mu, \Sigma), (\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)})).$$

It is straightforward to show that the above maximization has the following closed form. Maximizing for π is similar to deriving the multinomial MLE and maximizing for μ, Σ is similar to the Gaussian MLE.

$$\pi^{(t+1)} = \sum_{i=1}^{n} F_{ij}^{(t)} / \sum_{i=1}^{n} \sum_{j'=1}^{K} F_{ij'} = \sum_{i=1}^{n} F_{ij}^{(t)} / n, \qquad \mu_j^{(t+1)} = \sum_{i=1}^{n} F_{ij}^{(t)} Y^{(i)} / \sum_{i=1}^{n} F_{ij}^{(t)}$$

$$\Sigma_j^{(t+1)} = \sum_{i=1}^{n} F_{ij}^{(t)} (Y^{(i)} - \mu_j^{(t+1)})(Y^{(i)} - \mu_j^{(t+1)})^\top / \sum_{i=1}^{n} F_{ij}^{(t)}.$$