

1 Estimation with hidden variables and the EM algorithm

Thus far we focused on distribution families $p(x|\theta)$ of a relatively simple form where estimators (e.g., maximum likelihood) could be found in closed form. In the other extreme we discussed non-parametric methods that made no assumptions but required more data. In the middle there are parametric families that are more descriptive than e.g. Gaussians.

Uncertainty is generated by variables we can't observe (latent/hidden). It is generally impractical to consider a model that explicitly models those. However, in many cases, we have an understanding of what the important variables are and how the observations are dependent on those.

Examples:

- The distribution of height within men and women of a given age group. If we know the gender, it is reasonable to model height as a normal variable. Otherwise the height distribution will have two peaks, one for each gender.
- Toss a coin. If it's the same coin, the distribution is Bernouli. What if we know one out of two coins is chosen each time. Conditioned on the coin chosen, the distribution will be Bernouli.
- We record the activity of a neuron exposed to three stimuli. For each stimulus it might be Poisson but otherwise a mixture.
- We record from an electrode that at different times might reflect the activity of one of two neurons.

Sometimes we might even observe a hidden variable in some cases and in some cases not. For example, we record the response of k neurons, but in some trials some of the neurons cannot be reliably measured.

In all these cases there is a variable we cannot observe. Denote it by z . And we have a reasonable parametric model for $p(x, z|\theta)$ where θ is a set of parameters. For example in the height example (single observation), assume the observation is generated by choosing a gender z with probability c_z and then getting a height according to $\mathcal{N}(\mu_z, \sigma^2)$.

$$p(X = x, Z = z|\theta) = c_z \mathcal{N}(\mu_z, \sigma^2) \quad (1)$$

Thus:

$$p(x|\theta) = \sum_z p(x, z|\theta) = \sum_z c_z \mathcal{N}(\mu_z, \sigma^2) \quad (2)$$

This is referred to as a mixture distribution (this is a legal model if $c_z \geq 0$ and $\sum_z c_z = 1$).

If we have multiple observations x_1, \dots, x_n they will correspond to z_1, \dots, z_n (which we assume are independent) and

$$p(X = x^n, Z = z^n | \theta) = \prod_{i=1}^n p(x_i, z_i | \theta) \quad (3)$$

so

$$\begin{aligned} p(x^n | \theta) &= \sum_{z^n} \prod_i p(x_i, z_i | \theta) \\ &= \sum_{z_1, \dots, z_{n-1}} \sum_{z_n} \prod_i p(x_i, z_i | \theta) = \sum_{z_1, \dots, z_{n-1}} \prod_{i=1}^{n-1} p(x_i, z_i | \theta) \sum_{z_n} p(x_n, z_n | \theta) \\ \dots &= \prod_i \sum_{z_i} p(x_i, z_i | \theta) = \prod_i p(x_i | \theta) \end{aligned}$$

Thus IID samples from a mixture distribution correspond to marginalization over n latent variables.

2 Maximum likelihood with hidden data - EM

Say we have a model $p(x | \theta)$ that results from marginalization over some $p(x, z | \theta)$. We would like to find the maximum likelihood parameters θ (assume x may also stand for a size n sample, but our derivations will not assume that explicitly).

One thing to do is to simply write the likelihood $p(x | \theta)$ and maximize it as a function of θ via an optimization method of your choice (e.g., gradient ascent). There is however, a very elegant algorithm that often provides closed form updates for the parameters. It is called the Expectation Maximization algorithm and was invented by Dempster, Laird and Rubin (77).

Note: The likelihood function in the mixture case is typically non-convex and there is no procedure for finding a global optimum for it. Algorithms typically seek a local minimum in this case, and some can be shown to find it.

The key idea is this: assume that if we had complete information, i.e., someone gave us the value of z that we did not observe. In many cases the maximization of the *full likelihood* $p(x, z | \theta)$ can be found in closed form. For example, in the Gaussian mixture case, the maximum likelihood estimate for μ_z would be the empirical mean of the observations with $z^{(i)} = z$. The estimator for c_z would be the relative frequency of z in the sample.

But, clearly we do not have the unobserved variables... The idea in EM is as follows. Given some parameter vector $\hat{\theta}^t$ we can calculate the probability $p(z | x; \hat{\theta}^t)$ that z takes a given value. If this probability were one for a particular value we would be in the fully observed scenario. Even

if not, we can treat the probability as the probability that the unobserved variable will take this value. EM then suggests we consider the following likelihood:

$$Q(\theta; \hat{\theta}^t) = \sum_z p(z|x; \hat{\theta}^t) \log p(x, z|\theta) \quad (4)$$

This averages the log-likelihood according to the posterior probability obtained from $\hat{\theta}^t$. It is then natural to maximize this with respect to θ .

The EM algorithm generates a sequence of parameter estimates $\hat{\theta}^1, \dots, \hat{\theta}^t, \hat{\theta}^{t+1}, \dots$ such that $p(x|\hat{\theta}^t) \leq p(x|\hat{\theta}^{t+1})$, and the algorithm converges to a stationary point of the likelihood function.

Here is the procedure. Repeat for $t = 1, \dots, T$:

Expectation step: Construct the function $Q(\theta; \hat{\theta}^t)$. This requires calculating $p(z|x; \hat{\theta}^t)$ for the current parameter values.

Maximization step: Set $\hat{\theta}^{t+1} = \arg \max Q(\theta; \hat{\theta}^t)$.

Claim: The EM algorithm converges to a stationary point of the likelihood function $p(x|\theta)$.

We will first show that the likelihood increases at every step (note this does not imply convergence to a stationary point yet). Denote $\ell(\theta) = \log p(x|\theta)$. Then $\ell(\hat{\theta}^t) \leq \ell(\hat{\theta}^{t+1})$.

Start with rewriting $Q(\theta, \hat{\theta}^t)$:

$$\begin{aligned} Q(\theta, \hat{\theta}^t) &= \sum_z p(z|x, \hat{\theta}^t) \log p(x, z|\theta) = \sum_z p(z|x, \hat{\theta}^t) \log [p(z|x, \theta)p(x|\theta)] \\ &= \log p(x|\theta) + \sum_z p(z|x, \hat{\theta}^t) \log p(z|x, \theta) \\ &= \log p(x|\theta) - D_{KL}[p(z|x, \hat{\theta}^t)|p(z|x, \theta)] - H[p(z|x, \hat{\theta}^t)] \end{aligned}$$

Say θ^{t+1} maximizes the above. Then (discarding the constant)

$$\log p(x|\hat{\theta}^{t+1}) - D_{KL}[p(z|x, \hat{\theta}^t)|p(z|x, \hat{\theta}^{t+1})] \geq \log p(x|\hat{\theta}^t) - D_{KL}[p(z|x, \hat{\theta}^t)|p(z|x, \theta^t)]$$

So

$$\log p(x|\hat{\theta}^{t+1}) - \log p(x|\hat{\theta}^t) \geq D_{KL}[p(z|x, \hat{\theta}^t)|p(z|x, \hat{\theta}^{t+1})] \quad (5)$$

This shows improvement at each step. Another way of understanding this is by defining the function $F(\theta, \hat{\theta}^t) = Q(\theta, \hat{\theta}^t) + H[p(z|x, \hat{\theta}^t)]$. This is a lower bound on the true likelihood and equals the likelihood for $\theta = \hat{\theta}^t$. It's easy to see why this implies maximizing F (equivalent to maximizing Q yields improvement).

Note: It's easy to show that fixed point of EM correspond to stationary points of the likelihood. Note that these might actually be local minima but they would be very unstable. It is possible to show under mild conditions that EM converges to a fixed point and that if we don't start out at a local maximum we will converge to a stationary point (can still be a saddle point unless more conditions are satisfied).

2.1 EM for a Gaussian mixture

We'd like to use this to estimate the following model:

$$p(x^n|\theta) = \prod_i \sum_z c_z \mathcal{N}(x_i; \mu_z, \sigma_z^2) \quad (6)$$

The free parameters are c_z (where these must be non-negative and sum to one) and μ_z, σ_z^2 (the latter should be non-negative).

We have seen that this model is equivalent to a model $p(x^n, z^n|\theta)$ where we marginalize over the unobserved z . It thus perfectly fits the EM framework.

Assume that at time t we have the estimates $\hat{c}^t, \hat{\mu}^t, \hat{\sigma}^t$. Lets look at the Q function:

$$\begin{aligned} Q(\theta, \hat{\theta}^t) &= \sum_{z^n} p(z^n|x^n, \hat{\theta}^t) \log p(x^n, z^n|\theta) \\ &= \sum_{z^n} p(z^n|x^n, \hat{\theta}^t) \sum_i \log p(x_i, z_i|\theta) \\ &= \sum_i \sum_{z^n} p(z^n|x^n, \hat{\theta}^t) \log p(x_i, z_i|\theta) \\ &= \sum_i \sum_{z^i} p(z_i|x_i, \hat{\theta}^t) \log p(x_i, z_i|\theta) \end{aligned}$$

Note that $p(z^i|x^i, \hat{\theta}^t)$ can be calculated in given the current value of the parameters as:

$$p(z^i = m|x^i, \hat{\theta}^t) \propto p(x_i|z_i = m, \hat{\theta}^t)p(z_i = m) = \mathcal{N}(x_i, \hat{\mu}_m^t, \hat{\sigma}_m^{2,t})\hat{c}_m^t \quad (7)$$

Note also that $p(x_i, z_i|\theta) = c_{z_i}\mathcal{N}(x_i, \mu_{z_i}, \sigma_{z_i}^2)$. Taking the log and derivative w.r.t. μ_m yields

$$\frac{\partial Q(\theta, \hat{\theta}^t)}{\partial \mu_m} = \sum_i p(z_i = m|x_i, \hat{\theta}^t) \frac{(x_i - \mu_m)^2}{\sigma_m^2} \quad (8)$$

Hence the maximizer (and next estimate) is

$$\hat{\mu}_m^{t+1} = \frac{1}{\sum_i p(z_i = m|x_i, \hat{\theta}^t)} \sum_{i=1}^n p(z_i = m|x_i, \hat{\theta}^t) x_i \quad (9)$$

Similarly we have expression for the other estimators:

$$\hat{c}_m^{t+1} = \frac{1}{n} \sum_{i=1}^n p(z_i = m|x_i, \hat{\theta}^t) \quad (10)$$

and

$$\hat{\sigma}_m^{2,t+1} = \frac{1}{\sum_i p(z_i = m|x_i, \hat{\theta}^t)} \sum_{i=1}^n p(z_i = m|x_i, \hat{\theta}^t) (x_i - \hat{\mu}_m^{t+1})^2 \quad (11)$$

To derive \hat{c}_m^{t+1} we use Lagrange multipliers (no need for non-negativity constraints since they come out automatically). The Lagrangian is:

$$\mathcal{L}(c_m, \lambda) = Q(\theta, \hat{\theta}^t) - \lambda \left(\sum_m c_m - 1 \right) \quad (12)$$

$$\frac{\partial \mathcal{L}(c_m, \lambda)}{\partial c_m} = \frac{1}{c_m} \sum_i p(z_i = m|x_i, \hat{\theta}^t) - \lambda \quad (13)$$

Since $\sum_m c_m = 1$ we have:

$$1 = \frac{1}{\lambda} \sum_{m,i} p(z_i = m|x_i, \hat{\theta}^t) = \frac{n}{\lambda} \quad (14)$$

2.2 The K-Means algorithm

It often happens that the posterior distributions are peaked at some value of the latent variable. It might make sense to just put all the mass in the maximum, thus simplifying the algorithm. This is what the K-means algorithm does. It has other justifications in terms of vector quantization, but can be interpreted as a variant of EM.