Introduction to Machine Learning **Fall Semester**

Lecture 12

Lecturer: Amir Globerson *Scribe: Amir Globerson*

12.1 Latent variable models 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. Today we will consider parametric families that are more descriptive than e.g. Gaussians, and make use of latent variables.

Uncertainty is generated by variables we can't observe (latent/hidden). It is generally impractical to consider a model that explicitly models all the sources of uncertainty (e.g., modelling the exact physical properties of a tossed coin such that these properties fully determine the outcome of the toss). 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.
- Consider a Naive Bayes model for spam detection. But now assume we do not see the true class (i.e., the Y variable is hidden), only the bag of words representation. The features we will see are a *mixture* of two types of features (i.e., spam and non-spam).

In all these cases there is a variable we cannot observe. Denote it by Z. And we have a reasonable parametric model for $p(X=x,Z=z;\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ 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 sampling a height from the Gaussian $\mathcal{N}(x,\mu_z,\sigma^2)$ (we use $\mathcal{N}(x,\mu,\sigma^2)$ to denote the density of the Gaussian at point x). So:

$$p(x,z;\boldsymbol{\theta}) = c_z\mathcal{N}(x,\mu_z,\sigma_z^2) \tag{12.1}$$

From this it follows that the distribution over X (which is the variable we actually observe) is given by

$$p(x;\boldsymbol{\theta}) = \sum_z p(x,z;\boldsymbol{\theta}) = \sum_z c_z\mathcal{N}(x,\mu_z,\sigma_z^2) \tag{12.2}$$

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

$$p(x^n, z^n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i, z_i; \boldsymbol{\theta}) \tag{12.3}$$

so

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

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

12.2 Maximum likelihood with hidden data - EM

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

One thing to do is to simply write the likelihood $p(x; \boldsymbol{\theta})$ and maximize it as a function of $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}^t$ we can calculate the probability $p(z|x; \boldsymbol{\theta}^t)$ 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:

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^t) = \sum_z p(z|x;\boldsymbol{\theta}^t)\log p(x,z;\boldsymbol{\theta}) \tag{12.4}$$

This averages the log-likelihood according to the posterior probability obtained from $\boldsymbol{\theta}^t$. It is then natural to maximize this with respect to $\boldsymbol{\theta}$. The EM algorithm generates a sequence of parameter esimates $\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^t,\boldsymbol{\theta}^{t+1},\ldots$ such that $p(x;\boldsymbol{\theta}^t)\le p(x;\boldsymbol{\theta}^{t+1})$, and the algorithm converges to a stationary point of the likelihood function.

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

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

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

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 $\ell(\theta) = \log p(x;\boldsymbol{\theta})$. Then $\ell(\boldsymbol{\theta}^t)\le \ell(\boldsymbol{\theta}^{t+1})$.

Start with rewriting $Q(\boldsymbol{\theta};\boldsymbol{\theta}^t)$:

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

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

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

So

$$\log p(x;\boldsymbol{\theta}^{t+1}) - \log p(x;\boldsymbol{\theta}^t) \ge D_{KL}[p(z|x,\boldsymbol{\theta}^t)|p(z|x,\boldsymbol{\theta}^{t+1})] \tag{12.5}$$

This shows improvement at each step. Another way of understanding this is by defining the function $\bar{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^t) = Q(\boldsymbol{\theta};\boldsymbol{\theta}^t) + H[p(z|x,\boldsymbol{\theta}^t)]$. This is a lower bound on the true likelihood and equals the likelihood for $\boldsymbol{\theta} = \boldsymbol{\theta}^t$. It's easy to see why this implies maximizing $\bar{Q}$ (equivalent to maximizing Q yields improvement). This is therefore an instance of the minorization maximization approach you saw in the recitation.

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).

12.2.1 EM for a Gaussian mixture

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

$$p(x^n;\boldsymbol{\theta}) = \prod_i \sum_z c_z \mathcal{N}(x_i;\mu_z,\sigma_z^2) \tag{12.6}$$

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 $p(x^n, z^n; \boldsymbol{\theta})$ where we marginalize over the unobserved z. It thus perfectly fits the EM framework.

Assume that at time t we have the estimates c^t, μ^t, σ^t. Lets look at the Q function:

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

Note that $p(z_i|x_i,\boldsymbol{\theta}^t)$ can be calculated given the current value of the parameters as:

$$p(z_i=m|x_i;\boldsymbol{\theta}^t) = \frac{p(z_i=m,x_i;\boldsymbol{\theta}^t)}{\sum_k p(z_i=k,x_i;\boldsymbol{\theta}^t)} = \frac{p(x_i|z_i=m;\boldsymbol{\theta}^t)p(z_i=m)}{\sum_k p(x_i|z_i=k;\boldsymbol{\theta}^t)p(z_i=k)}$$

which results in:

$$p(z_i=m|x_i;\boldsymbol{\theta}^t) = \frac{\mathcal{N}(x_i,\mu_m^t,\sigma_m^{2,t})c_m^t}{\sum_k \mathcal{N}(x_i,\mu_k^t,\sigma_k^{2,t})c_k^t} \tag{12.7}$$

We have now expressed $Q(\boldsymbol{\theta};\boldsymbol{\theta}^t)$, and can proceed to the M step of maximizing it. Taking the log and derivative w.r.t. μ_m yields

$$\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^t)}{\partial \mu_m} = \sum_i p(z_i=m|x_i,\boldsymbol{\theta}^t)\frac{(x_i-\mu_m)}{\sigma_m^2} \tag{12.8}$$

Hence the maximizer (and next estimate) is

$$\mu_m^{t+1} = \frac{1}{\sum_i p(z_i=m|x_i,\boldsymbol{\theta}^t)}\sum_{i=1}^n p(z_i=m|x_i,\boldsymbol{\theta}^t)x_i \tag{12.9}$$

Similarly we have expression for the variance:

$$\sigma_m^{2,t+1} = \frac{1}{\sum_i p(z_i = m|x_i, \boldsymbol{\theta}^t)} \sum_{i=1}^{n} p(z_i = m|x_i, \boldsymbol{\theta}^t)(x_i - \mu_m^{t+1})^2 \tag{12.10}$$

Below we show that the new c is given by:

$$c_m^{t+1} = \frac{1}{n} \sum_{i=1}^{n} p(z_i = m|x_i, \boldsymbol{\theta}^t) \tag{12.11}$$

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

$$\mathcal{L}(c_m, \lambda) = Q(\boldsymbol{\theta}; \boldsymbol{\theta}^t) - \lambda \left(\sum_m c_m - 1 \right) \tag{12.12}$$

The relevant part of Q is only the one where $z_i = m$.

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^t) = const + \sum_i p(z_i = m|x_i; \boldsymbol{\theta}^t) \log p(x_i, z_i = m) = const + \sum_i p(z_i = m|x_i; \boldsymbol{\theta}^t) \log c_m \tag{12.13}$$

$$\frac{\partial \mathcal{L}(c_m, \lambda)}{\partial c_m} = \frac{1}{c_m} \sum_i p(z_i = m|x_i, \boldsymbol{\theta}^t) - \lambda \tag{12.14}$$

Since $\sum_m c_m = 1$ we have:

$$1 = \frac{1}{\lambda} \sum_{m,i} p(z_i = m|x_i, \boldsymbol{\theta}^t) = \frac{n}{\lambda} \tag{12.15}$$

12.3 Bayesian Parameter Estimation

In standard ML, we treat $\boldsymbol{\theta}$ as an unknown parameter which we optimize over. In the Bayesian framework, one views $\boldsymbol{\theta}$ itself as a random variable Θ. We therefore also assume there is a prior distirbution on $\boldsymbol{\theta}$, denote by $q(\boldsymbol{\theta})$.

The generation process is then to sample a θ from $q(\boldsymbol{\theta})$ and then sample $x_1, \ldots, x_n$ independently from $p(x_i|\boldsymbol{\theta})$. The overall process is then:

$$p(x_1, \ldots, x_n, \boldsymbol{\theta}) = q(\boldsymbol{\theta}) \prod_i p(x_i|\boldsymbol{\theta}) \tag{12.16}$$

The nice thing is this now allows us to consider the posterior distirbution $p(\boldsymbol{\theta}|x_1,\ldots,x_n)$ which reflects our uncertainty about the parameter, after having viewed the data.

$$p(\boldsymbol{\theta}|x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n,\boldsymbol{\theta})}{p(x_1,\ldots,x_n)} \propto q(\boldsymbol{\theta}) \prod_i p(x_i|\boldsymbol{\theta}) \tag{12.17}$$

By $\propto$ we mean this is equality up to a muliplicative constant which normalizes the distribution.

The posterior is a distribution over $\boldsymbol{\theta}$ values rather than a single estimator as we had in ML. If we want to return a single value there are two popular choices:

- MAP estimator - Return the θ that maximizes the posterior.
- Expected Value - Return the expected value of Θ under the distribution $p(\boldsymbol{\theta}|x_1,\ldots,x_n)$.

Before giving examples, let us consider what happens as we get more data. Consider the log of the posterior, scaled by $\frac{1}{n}$:

$$\frac{1}{n}\log p(\boldsymbol{\theta}|x_1,\ldots,x_n) = \frac{1}{n}\log q(\boldsymbol{\theta}) + \frac{1}{n}\sum_i \log p(x_i|\boldsymbol{\theta}) \tag{12.18}$$

As $n \to \infty$ the first term will vanish, and the second term is the likelihood. So we conclude that as we have more data, the prior will have less and less effect on the maximizer of the posterior, and the estimator will get closer to the ML estimator.

12.3.1 Gaussian Example

Example (Gaussian family and prior): consider the case where θ is the mean of a Gaussian distribution:

$$p(x|\theta) = \mathcal{N}(x,\theta,1) \tag{12.19}$$

And the prior is $f(\theta) = \mathcal{N}(\theta,0,1)$. Given a set of IID observations, the posterior is then: Thus:

$$p(\theta|x^n) \propto \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\sum_i (x_i-\theta)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2} \tag{12.20}$$

Thus:

$$\begin{aligned} p(\theta|x^n) &\propto e^{-\frac{1}{2}\sum_i (x_i-\theta)^2} e^{-\frac{1}{2}\theta^2} \\ &\propto e^{-\frac{1}{2}((n+1)\theta^2 - 2\theta\sum_i x_i)} \\ &\propto e^{-\frac{n+1}{2}(\theta - \frac{1}{n+1}\sum_i x_i)^2} = \mathcal{N}(\theta, \frac{1}{n+1}\sum_i x_i, \frac{1}{n+1}) \end{aligned}$$

The Bayesian estimator is the mean of the posterior and thus we have:

$$\theta_{L2}(x^n) = \frac{1}{n+1}\sum_i x_i \tag{12.21}$$

Some observations about this:

- The estimator behaves as if we had another observation $x_{n+1} = 0$ (although we only have n observations). This *dummy* observation is the outcome of using the prior (which in this case is indeed has mean zero).
- The variance of the posterior shrinks with n. This implies that it will be more strongly peaked around its mean, and hence more *confident* of the estimate. It makes sense since we are getting more observations.
- The posterior is also a Gaussian (same as the prior and the parametric family). When a posterior has the same form as the prior for a given family, we say the prior is a conjugate prior for that family. Thus a Gaussian prior is a conjugate prior for the Gaussian family.

12.3.2 Beta Priors

The Beta distribution over variable $0 \le \theta \le 1$ is defined as:

$$p(\theta; \alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{12.22}$$

where $B(\alpha, \beta)$ is a normalization constant known as the Beta function. Its mean is $\frac{\alpha}{\alpha+\beta}$. We will next see that it is the conjugate prior of the Bernouli distribution, where the Bernouli family $p(x; \theta)$ is defined for $x \in \{0, 1\}$ and

$$p(x|\theta) = \left\{ \begin{array}{cc} \theta & x = 1 \\ 1-\theta & x = 0 \end{array} \right. \tag{12.23}$$

Say we want to estimate θ given x with a Bernouli family and Beta prior. Lets see what the posterior is:

$$\begin{aligned} p(\theta|x) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{x}(1-\theta)^{1-x} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+1-x-1} \end{aligned}$$

Thus we have that $f(\theta|x) = B(\alpha + x, \beta + 1 - x)$. The posterior mean is $\frac{\alpha+x}{\alpha+\beta+1}$. Repeating this n times we will have after n observations: $f(\theta|x^n) = B(\alpha + \sum_i x_i, \beta + 1 - \sum_i x_i)$. The posterior mean is thus: $\frac{\alpha+\sum_i x_i}{\alpha+\beta+n}$. This has the following nice interpretation: assume that originally we had α successes $(x = 1)$ and β failures $(x = 0)$. We now add n data points to this initial data, and the resulting estimate is exactly the proportion of successes in this new data.

Recall that the maximum likelihood in this case is simply $\frac{1}{n}\sum_i x_i$.