Introduction to Machine Learning **Fall Semester**

Recitation 2

Lecturer: Regev Schweiger *Scribe: Regev Schweiger*

2.1 Maximum Likelihood

2.1.1 Poisson

Consider a Poisson distribution. A Poisson distribution is defined by a parameter $\lambda > 0$ and the probability is defined over the integers and denoted by $Pois(\lambda)$. The motivation is that it models an arrival rates of individuals with an average arrival rate of λ. The probability of having k individual arrive when $X \sim Pois(\lambda)$ is,

$$\Pr[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Assume we have a sample of n points $S = \{z_i, \ldots, z_n\}$ where each z_i is drawn independently from a distribution $Pois(\lambda)$. The likelihood function would be,

$$L_S(\lambda) = \Pr[S|\lambda] = \prod_{i=1}^{n} \Pr[z_i|\lambda] = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda_i^z}{z_i!}.$$

Many times, it is more convenient to work with the *log-likelihood*, simply taking the logarithm of the likelihood, and the product becomes a sum. Note that maximizing the likelihood is equivalent to maximizing the log-likelihood. In our case, the log-likelihood is:

$$\ell_S(\lambda) = \log L_S(\lambda) = \sum_{i=1}^{n} (-\lambda + z_i \log \lambda - \log(z_i!))$$

We would like to find the λ that maximizes the likelihood, denoted by λ_{ML}. Since the terms $\log(z_i!)$ do not depend on λ we can ignore them in the maximization. We have,

$$\hat{\lambda} = \arg\max_{\lambda} \left(-n\lambda + \left(\sum_{i=1}^{n} z_i \right) \log \lambda \right)$$

Taking the derivative and equating with zero we have,

$$0 = -n + \left(\sum_{i=1}^{n} z_i \right) \frac{1}{\lambda}$$

and the solution is,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} z_i}{n}.$$

We need to verify that this is indeed a maximum. The second derivative is

$$\left(\sum_{i=1}^{n} z_i\right) \frac{-1}{\lambda^2} < 0$$

and therefore we found a maximum.

2.1.2 Two Bits

Now, imagine we have two bits (e.g., two coins flips). However, the probability of the second coin flip depends on the result of the first coin flip. Formally, we will have three parameters: $p_0, p_{0|0}, p_{0|1}$, and we define $p_1 = 1 - p_0$, $p_{1|0} = 1 - p_{0|0}$, $p_{0|1} = 1 - p_{1|1}$. The probability to see a result $(x_1, x_2) \in \{0,1\}^2$ is

$$\Pr(X_1 = x_1, X_2 = x_2) = p_{x_1} \cdot p_{x_2|x_1}$$

We will to find the maximum likelihood estimates of the parameters. Given a sample $(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)$, the log likelihood is:

$$\ell(p_0, p_{0|0}, p_{0|1}; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p_{x_1^i} + \log p_{x_2^i|x_1^i}$$

Let us do $p_{0|0}$ as an example:

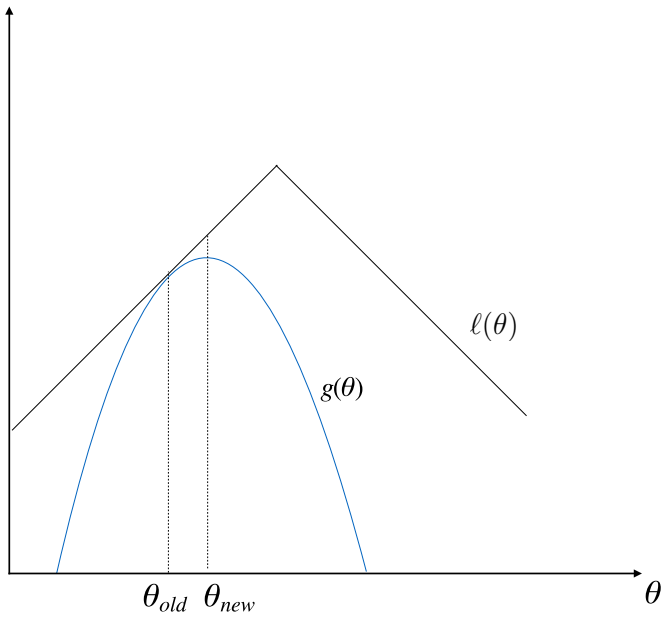
$$\frac{\partial \ell}{\partial p_{0|0}} = \sum_{i=1}^{n} \frac{\delta_{x_1^i=0, x_2^i=0}}{p_{0|0}} - \frac{\delta_{x_1^i=0, x_2^i=1}}{1 - p_{0|0}}$$

Denote by n_{00} the number of observed 00 pairs (similarly, n_{01}). Then, we have

$$\frac{\partial \ell}{\partial p_{0|0}} = \frac{n_{00}}{p_{0|0}} - \frac{n_{01}}{1 - p_{0|0}} = 0 \Rightarrow \hat{p}_{0|0} = \frac{n_{00}}{n_{00} + n_{01}}$$

2.2 Minorization Maximization (MM)

Suppose we have a function $\ell(\theta)$ that we want to maximize (i.e. find $\arg\max_\theta \ell(\theta)$). However, while we can calculate $\ell(\theta)$ for a given θ, it is hard to find its maximum. We want to develop an iterative procedure to find the maximum. In each iteration, we start with a given initial guess θ_{old}, and we want to find a new guess θ_{new}, so that:



$$\ell(\theta_{new}) \geq \ell(\theta_{old})$$

For example:

(In this case $\ell(\theta)$ may not really hard to maximize, but for the sake of this explanation, assume that it is.)

To do that, let's assume we have a family of simpler functions, $g(\theta; a_1, \ldots, a_k)$, (for some parameters $a_1, \ldots, a_k$), for which it is relatively easier for us to find the maximum. For example, we can limit ourselves to quadratic functions: $g(\theta; a_1, a_2, a_3) = a_1\theta^2 + a_2\theta + a_3$. We can indeed find the maximum of each such g easily, since we have a closed formula for it.

In addition, let's assume that we can easily find parameters $a_1, \ldots, a_k$ for which the corresponding function $g(\theta; a_1, \ldots, a_k)$ will be bounded by $\ell(\theta)$:

$$g(\theta; a_1, \ldots, a_k) \leq \ell(\theta)$$

And finally, let's assume that for our initial guess of θ_{old}, we can also find parameters such that:

$$g(\theta_{old}; a_1, \ldots, a_k) = \ell(\theta_{old})$$

This is clearly the case in our example (a high-school level exercise). Under the above conditions, we can perform the following steps:

1. Find parameters $a_1^*, \ldots, a_k^*$ for which:

$$g(\theta; a_1^*, \ldots, a_k^*) \leq \ell(\theta)$$

$$g(\theta_{old}; a_1^*, \ldots, a_k^*) = \ell(\theta_{old})$$

2. Find the maximum for the chosen g:

$$\theta_{new} = \arg\max_{\theta} g(\theta; a_1^*, \ldots, a_k^*)$$

For this value, from the properties of g, we get:

$$\ell(\theta_{new}) \geq g(\theta_{new}; a_1^*, \ldots, a_k^*) \geq g(\theta_{old}; a_1^*, \ldots, a_k^*) = \ell(\theta_{old})$$

The first inequality is because $\ell(\theta)$ is an upper bound for g; the second is because this is where the maximum is achieved; and the third is because of our choice of parameters $a_1, \ldots, a_k$.

This principle is called Minorization Maximization (MM), and this is the idea that stands behind Expectation Maximization (EM). EM is a special case of MM, for which $\ell(\theta)$ is the log-likelihood of the parameter(s) we want to maximize, given the data. The more complicated issues in EM are the choice of g-s, and the proof that they obey the required properties. But if you understand this principle, you understand why EM works.

2.3 Expectation Maximization (EM)

The EM principle is useful when we have a model with hidden (latent) variables. To understand how it works, we will go through the analysis with a simple example in mind.

Suppose we have the following process. We want to generate a series of n coin flips, $x_1, \ldots, x_n$. For each data sample we want to generate, we take two steps: On the first step, we first flip a coin with a probability θ which returns either 0 or 1. More precisely, define Z_i as the hidden variable that is the outcome of the first flip. Then $\Pr[Z_i = 1] = \theta$ and $\Pr[Z_i = 0] = 1 - \theta$. If $Z_i = 0$ then we flip a coin with bias $p_0 = 2/3$ to set the random variable X_i ($P(X_i = 1) = 2/3, P(X_i = 0) = 1/3$). If $Z_i = 1$ then we flip a coin with bias $p_1 = 1/4$ to set X_i ($P(X_i = 1) = 1/4, P(X_i = 0) = 3/4$). Schematically:

$$\xrightarrow{\theta} Z_i = z_i \xrightarrow{2/3 \text{ or } 1/4} X_i = x_i$$

We observe the sequence $\mathbf{x} = \{x_i\}_{i=1}^n$, for example $\mathbf{x} = (0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0)$ for $n = 13$. We would like to find the maximum likelihood estimator of θ, and to run an EM algorithm in order to do that.

2.3.1 Choice of $\ell(\theta)$

First, we define the target function: The log-likelihood of the parameter θ given the data (in general, this may be a vector of several parameters, $\boldsymbol{\theta}$). Namely:

$$\ell(\theta) = \ell(\theta; x_1, \ldots, x_n)$$

Example

In our example, the log-likelihood is:

$$\begin{aligned}
\ell(\theta; x_1, \ldots, x_n) &= \sum_{i=1}^{n} \log \Pr_{\theta}[X_i = x_i] \\
&= \sum_{i=1}^{n} \log \left((1-\theta) \cdot \left(\frac{2}{3}\right)^{x_i} \left(\frac{1}{3}\right)^{1-x_i} + \theta \cdot \left(\frac{1}{4}\right)^{x_i} \left(\frac{3}{4}\right)^{1-x_i} \right)
\end{aligned}$$

2.3.2 Choice of g-s

Now we turn to choose our family of distributions. Suppose $\mathbf{x} = x_1, \ldots, x_n$ is our observed data, and $\mathbf{z} = z_1, \ldots, z_n$ are the values of the hidden variables. For a specific instance of $\mathbf{z}$, as a function of θ, we can calculate the following quantity:

$$\log \Pr_{\theta}[\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}]$$

Suppose we don't know the precise values of $\mathbf{z}$, but instead of have a *distribution* $\mathbf{Z} \sim f(\mathbf{z})$ over all possible values of $\mathbf{Z}$. Now, $\log \Pr_{\theta}[X = \mathbf{x}, \mathbf{Z}]$ is no longer a number, but a random variable, for which we can now calculate its expectation, *as a function of* θ.

In addition, we will calculate another quantity derived from the distribution $f(\mathbf{z})$:

$$H(f) = -\sum_{\mathbf{z}} f(\mathbf{z}) \log f(\mathbf{z})$$

Usually we will not trouble ourselves with explicitly calculating $H(f)$, since it is independent of θ.

This will be our choice of function family g:

$$\begin{aligned}
g(\theta; f) &= E_{\mathbf{Z} \sim f(\mathbf{z})}[\log \Pr_{\theta}[\mathbf{X} = \mathbf{x}, \mathbf{Z}]] + H(f) \\
&= \sum_{\mathbf{z}} \Pr_{\mathbf{Z} \sim f(\mathbf{z})}[\mathbf{Z} = \mathbf{z}] \cdot \log \Pr_{\theta}[\mathbf{X} = \mathbf{x}, \mathbf{z}] + H(f)
\end{aligned}$$

Example

In our example, for a given $\mathbf{z}, \theta$:

$$\begin{aligned}\log \Pr_\theta[\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}] &= \sum_{i=1}^{n} \log \Pr_\theta[X_i=x_i, Z_i=z_i] \\ &= \sum_{i=1}^{n} \log \left(\left((1-\theta)\cdot \left(\frac{2}{3}\right)^{x_i} \left(\frac{1}{3}\right)^{1-x_i} \right)^{1-z_i} \cdot \left(\theta \cdot \left(\frac{1}{4}\right)^{x_i} \left(\frac{3}{4}\right)^{1-x_i} \right)^{z_i} \right)\end{aligned}$$

Now assume we are given the distribution of $\mathbf{z}$, and not concrete values. In our example, the space of all the possible distributions $f(\mathbf{z})$ is actually quite limited. First, we know that the Z_i are independent; secondly, $\Pr[Z_i=0] = 1-\Pr[Z_i=1]$ so we need only determine the parameters:

$$\mathbf{a} = (a_1,\ldots,a_n) = (\Pr[Z_1=1],\ldots,\Pr[Z_n=1])$$

This gives:

$$\begin{aligned} g(\theta;f) &= E_{\mathbf{Z}\sim f(\mathbf{z})}[\log \Pr_\theta[\mathbf{X}=\mathbf{x},\mathbf{Z}]] + H(f) \\ &= E_{\mathbf{Z}\sim f(\mathbf{z})}\left[\sum_{i=1}^{n} \log \Pr_\theta[X_i=x_i,Z_i]\right] + H(f) \\ &= \sum_{i=1}^{n} E_{\mathbf{Z}\sim f(\mathbf{z})}\left[\log \Pr_\theta[X_i=x_i,Z_i]\right] + H(f) \\ &= \sum_{i=1}^{n} \Pr_{\mathbf{Z}\sim f(\mathbf{z})}[Z_i=0]\cdot \log\left((1-\theta)\cdot\left(\frac{2}{3}\right)^{x_i}\left(\frac{1}{3}\right)^{1-x_i}\right) \\ &+ \Pr_{\mathbf{Z}\sim f(\mathbf{z})}[Z_i=1]\cdot \log\left(\theta\cdot\left(\frac{1}{4}\right)^{x_i}\left(\frac{3}{4}\right)^{1-x_i}\right) + H(f) \\ &= \sum_{i=1}^{n}(1-a_i)\cdot\left(\log(1-\theta) + x_i\cdot\log\left(\frac{2}{3}\right) + (1-x_i)\cdot\log\left(\frac{1}{3}\right)\right) \\ &+ a_i\cdot\left(\log\theta + x_i\cdot\log\left(\frac{1}{4}\right) + (1-x_i)\cdot\log\left(\frac{3}{4}\right)\right) + H(f)\end{aligned}$$

2.3.3 Why these g-s?

Recall that in the Minorization-Maximization principle, we required three properties from our functions g (note that the parameters of g are now f):

1. That we can choose a f for which $g(\theta; f) \leq \ell(\theta)$;
2. That we can choose a f for which $g(\theta_{old}; f) = \ell(\theta_{old})$;
3. That we can find the maximum of $g(\theta; f)$ relatively easily.

It turns out that for *every* distribution $f(\mathbf{z})$:

$$g(\theta; f) = E_{\mathbf{Z}\sim f(\mathbf{z})}[\log \Pr_{\theta}[\mathbf{X} = \mathbf{x}, \mathbf{Z}]] + H(f) \leq \ell(\theta; \mathbf{x}) = \ell(\theta)$$

This will be shown in class, so the proof will not be repeated here. This proves that property (1) holds for our choice of family of functions $g(\theta; f)$.

Secondly, it turns out that if we choose f to be the *posterior* distribution of $\mathbf{Z}$ conditioned on $\mathbf{x}$ and given our current estimate θ_{old}:

$$f^*(\mathbf{Z} = \mathbf{z}) = \Pr_{\theta_{old}}[\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x}]$$

then, for this choice of distribution, property (2) holds:

$$g(\theta_{old}; f^*(\mathbf{z})) = \ell(\theta_{old}; \mathbf{x})$$

The proof of which is also omitted here. Often, the function $E_{\mathbf{Z}\sim f^*(\mathbf{z})}[\log \Pr_{\theta}[\mathbf{X} = \mathbf{x}, \mathbf{Z}]]$ is also denoted $Q(\theta|\theta_{old})$, since its distribution f^* is determined by θ_{old}. Finally, we hope that property (3) also holds. Experience shows that for many classes of models, this is true.

Example

In our example, let's first calculate the posterior probability of $\mathbf{z}$. Again, due to independence, it's enough to do this for each z_i, x_i separately:

$$\begin{aligned} a_i^* = \Pr_{\theta_{old}}[Z_i = 1|X_i = x_i] &= \frac{\Pr_{\theta_{old}}[Z_i = 1, X_i = x_i]}{\Pr_{\theta_{old}}[X_i = x_i]} \\ &= \frac{\theta_{old} \cdot \left(\frac{2}{3}\right)^{x_i} \left(\frac{1}{3}\right)^{1-x_i}}{\theta_{old} \cdot \left(\frac{2}{3}\right)^{x_i} \left(\frac{1}{3}\right)^{1-x_i} + (1-\theta_{old}) \cdot \left(\frac{1}{4}\right)^{x_i} \left(\frac{3}{4}\right)^{1-x_i}} \end{aligned}$$

and

$$\Pr_{\theta_{old}}[Z_i = 0|X_i = x_i] = 1 - \Pr_{\theta_{old}}[Z_i = 1|X_i = x_i]$$

and then:

$$f^*(\mathbf{Z} = \mathbf{z}) = \Pr_{\theta_{old}}[\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x}] = \prod_{i=1}^{n} \Pr_{\theta_{old}}[Z_i = z_i|X_i = x_i]$$

We have therefore chosen our lower bound function to be $g(\theta; f^*)$. This is called the **E-Step**.

Now we try to maximize it - the **M-Step**. Recall that we have seen:

$$\begin{aligned} g(\theta; f) &= \sum_{i=1}^{n}(1-a_i^*)\cdot\left(\log(1-\theta)+x_i\cdot\log\left(\frac{2}{3}\right)+(1-x_i)\cdot\log\left(\frac{1}{3}\right)\right) \\ &+ a_i^*\cdot\left(\log\theta+x_i\cdot\log\left(\frac{1}{4}\right)+(1-x_i)\cdot\log\left(\frac{3}{4}\right)\right)+H(f) \end{aligned}$$

where a_i are now as defined above. Derive with respect to θ to find the maximum:

$$\begin{aligned} \frac{\partial g(\theta; f)}{\partial\theta} &= -\frac{1}{1-\theta}\cdot\sum_{i=1}^{n}(1-a_i^*)+\frac{1}{\theta}\cdot\sum_{i=1}^{n}a_i^*=0\Rightarrow \\ \theta_{new} &= \frac{\sum_{i=1}^{n}a_i^*}{n} \end{aligned}$$

We need to verify that this is a maximum, by checking the second derivative (not shown). This concludes one EM iteration. ge have therefore found a θ_{new} for which:

$$\ell(\theta_{new}; \mathbf{x}) \geq \ell(\theta_{old}; \mathbf{x})$$

as required.