208891 Probabilistic Graphical Models **Spring 2019**

Lecture 12: February 10

Lecturer: Donlapark Pornnopparath

12.1 Parameter estimation

As mentioned in the previous sections, when data are missing, the parameters in the likelihood function cannot be decomposed in general, causing the function to be highly complex and non-convex. There are two optimization methods what we could use to approximate the MLE estimator.

12.1.1 Gradient ascent

This is a standard convex optimization method that can be applied to non-convex function. However, the solution that we obtain might only be a local maxima and not the global one. The main idea behind the algorithm is to always climb the surface of the log-likelihood function $LL(\boldsymbol{\theta}|D)$ uphill, meaning that, in each step, the parameters $\boldsymbol{\theta}$ are moved in the "same" direction as that of $\nabla_{\boldsymbol{\theta}} LL(\boldsymbol{\theta}|D)$. More precisely, we perform the following algorithm:

1. Initialize parameters $\boldsymbol{\theta}_0$.
2. At t–th step, we move the parameters using $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \eta\nabla LL(\boldsymbol{\theta}_{t-1}, D)$.

Here, $\eta \in \mathbb{R}^+$ is the *step size* which should be small enough that the parameters eventually stabilize at a local maxima.

Example 12.1. Suppose that we have a simple graph as in Figure 12.1 with three variables: $A \in \{0,1\}$, $B \in \{0,1\}$, $C \in \{0,1\}$. Then the parameters we have to learn are $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_{C|0,0}, \theta_{C|0,1}, \theta_{C|1,0}, \theta_{C|1,1})$. Suppose that we observe $D_1 = \{(0,?,1)\}$ and $D_2 = \{(1,?,0)\}$. Then the log-likelihood function of $D = D_1 \cup D_2$ is

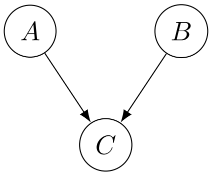

Figure 12.1: A little v

$$
\begin{aligned}
LL(\boldsymbol{\theta}|D) = &\log((1-\theta_A)\theta_B\theta_{C|0,1} + (1-\theta_A)(1-\theta_B)\theta_{C|0,0}). \\
&+ \log(\theta_A\theta_B(1-\theta_{C|1,1}) + \theta_A(1-\theta_B)(1-\theta_{C|1,0})).
\end{aligned}
$$

For example, in each step, the value of $\theta_{C|1,1}$ is added by

$$
\eta\frac{\partial LL(\boldsymbol{\theta}|D)}{\partial\theta_{C|1,1}} = -\frac{\eta\theta_A\theta_B}{\theta_A\theta_B(1-\theta_{C|1,1}) + \theta_A(1-\theta_B)(1-\theta_{C|1,0})} \tag{12.1}
$$

Suppose that we set $\eta = 0.1$ and the current values of parameters is $\theta_A = 0.1, \theta_B = 0.3$ and $\theta_{C|1,0} = \theta_{C|1,1} = 0.5$. Plugging these in (12.1), we obtain

$$\eta \frac{\partial LL(\boldsymbol{\theta}|D)}{\partial \theta_{C|1,1}} = 0.06.$$

$\diamond$

Thus, the value of $\theta_{C|1,1}$ is updated as follows:

$$\theta_{C|1,1} \leftarrow \theta_{C|1,1} + \eta \frac{\partial LL(\boldsymbol{\theta}|D)}{\partial \theta_{C|1,1}} = 0.5 + 0.06 = 0.56.$$

12.1.2 Expectation-maximization (EM)

In contrast to gradient ascent, the *expectation-maximization* algorithm is designed to optimize likelihood functions. Suppose that we have data $D = \{(x^{(n)}, z^{(n)})\}_{n=1}^N$ where $x^{(n)}$ is the observed variables and $z^{(n)}$ is the missing variables. Recall that our goal is to maximize the log-likelihood function

$$LL(\boldsymbol{\theta}|D) = \sum_n \log P(\boldsymbol{x}^{(n)}, \boldsymbol{z}^{(n)}|\boldsymbol{\theta}) = \sum_n \log \left(\sum_{\boldsymbol{z}^{(n)}} P(\boldsymbol{x}^{(n)}|\boldsymbol{z}^{(n)}, \boldsymbol{\theta}) P(\boldsymbol{z}^{(n)}|\boldsymbol{\theta}) \right).$$

This can be done with various optimization methods, which might involve heavy computations due to the logarithm being outside of the summation. This is not a problem if we use the EM algorithm.

EM algorithm: A parameter $\boldsymbol{\theta}_0$ is initialized. Then following two steps are performed to obtain $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ until convergence.

- *E-Step*: For each n, compute the posterior $p(\boldsymbol{z}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{\theta}_t)$ and the expected log-likelihood.

  $$Q_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z} \sim p(z^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{\theta}_t)} \log p(\boldsymbol{x}^{(n)}, \boldsymbol{z}|\boldsymbol{\theta}).$$

  Let's focus on the posterior term for now. Intuitively, we are trying to "fill" the missing variables with the underlying probability distributions.

- *M-Step*: Compute the new parameters using

  $$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \arg\max_{\boldsymbol{\theta}} \sum_n Q_n \\ &= \arg\max_{\boldsymbol{\theta}} \sum_n \mathbb{E}_{\boldsymbol{z} \sim p(z^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{\theta}_t)} \log p(\boldsymbol{x}^{(n)}, \boldsymbol{z}|\boldsymbol{\theta}) \end{aligned}$$

  In contrast to the original optimization problem, the logarithm is now inside the summation, which makes it easier to find a solution. Even though the solution that we obtain in this step does not exactly solve the original problem, it can be shown that $LL(\boldsymbol{\theta}_t|D)$ is increasing and converges to a maxima as $t \to \infty$.

Example 12.2. We use the same setup as in Example 12.1 but with $D_1 = \{(1, ?, 1)\}$ and $D_2 = \{(?, 1, 0)\}$. Suppose that we have obtained $\boldsymbol{\theta}_t = (0.2, 0.3, 0.5, 0.8, 0.1, 0.4)$. The posterior distributions in the E-Step are

$$\begin{aligned} p(b = 1|1, 1, \boldsymbol{\theta}_t) &= \frac{0.2 \cdot 0.3 \cdot 0.4}{0.2 \cdot 0.3 \cdot 0.4 + 0.2 \cdot 0.7 \cdot 0.1} = 0.63 \\ p(a = 1|1, 0, \boldsymbol{\theta}_t) &= \frac{0.2 \cdot 0.3 \cdot 0.6}{0.2 \cdot 0.3 \cdot 0.6 + 0.8 \cdot 0.3 \cdot 0.9} = 0.14. \end{aligned}$$

Let $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_{C|0,0}, \theta_{C|0,1}, \theta_{C|1,0}, \theta_{C|1,1})$ be the parameters that we want to optimize in the M-Step. It follows that

$$\begin{aligned}
Q_1(\boldsymbol{\theta}) &= p(b=0|1,1,\boldsymbol{\theta}_t)\log\left(\theta_A(1-\theta_B)\theta_{C|1,0}\right) \\
&\quad + p(b=1|1,1,\boldsymbol{\theta}_t)\log\left(\theta_A\theta_B\theta_{C|1,1}\right) \\
&= 0.37\log\left(\theta_A(1-\theta_B)\theta_{C|1,0}\right) + 0.63\log\left(\theta_A\theta_B\theta_{C|1,1}\right) \\
Q_2(\boldsymbol{\theta}) &= p(a=0|1,0,\boldsymbol{\theta}_t)\log\left((1-\theta_A)\theta_B(1-\theta_{C|0,1})\right) \\
&\quad + p(a=1|1,0,\boldsymbol{\theta}_t)\log\left(\theta_A\theta_B(1-\theta_{C|1,1})\right) \\
&= 0.86\log\left((1-\theta_A)\theta_B(1-\theta_{C|0,1})\right) + 0.14\log\left(\theta_A\theta_B(1-\theta_{C|1,1})\right).
\end{aligned}$$

Notice that we can optimize $Q = Q_1 + Q_2$ by considering each of the parameters separately. For example, the optimal value of $\theta_{C|1,1}$ can be found by solving

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_{C|1,1}} = \frac{0.63}{\theta_{C|1,1}} - \frac{0.14}{1-\theta_{C|1,1}} = 0,$$

which yields $\hat{\theta}_{c|1,1} = 0.63/(0.63+0.14) = 0.82$. ◇

There are many statistical models that involve *latent variables* i.e. variables that are not observed from the data. By treating the latent variables as the missing variables, we can utilize the EM algorithm to learn these models from data.

Example 12.3 (Gaussian Mixture Models (GMM))**.** In this model, we try to group data points $\{x_1, x_2, \ldots\}$ into K clusters, each of which is distributed as Gaussian. In other words, we are trying to create a probabilistic model of (x_n, z_n) where $z_n \in \{1, 2, \ldots, K\}$ is a latent variable. The graphical model is $Z \to X$, meaning that

$$p(x,z) = p(x|z)p(z)$$

where

$$p(x|z=k) = N(x|\mu_k, \Sigma_k)$$

is a multivariate Gaussian with mean u_k and covariance matrix Σ_k. The approach to compute the posterior is a little bit different than the previous example. Let $\boldsymbol{\theta}_t$ consist of all parameters of $p(x|z)$ and $p(z)$ after t steps. Then the posterior of z can be computed as follows:

$$p(z|x,\boldsymbol{\theta}_t) = \frac{p(x|z,\boldsymbol{\theta}_t)p(z|\boldsymbol{\theta}_t)}{\sum_{k=1}^{K} p(x|k,\boldsymbol{\theta}_t)p(k|\boldsymbol{\theta}_t)}.$$

Then in the M-Step, we update the parameters using

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \arg\max_{\boldsymbol{\theta}} \sum_{x\in D} \mathbb{E}_{z\sim p(z|x,\boldsymbol{\theta}_t)} \log p(x,z|\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{K}\sum_{x\in D} p(k|x,\boldsymbol{\theta}_t)\log p(x|k,\boldsymbol{\theta})p(k|\boldsymbol{\theta}),
\end{aligned}$$

which can be solved using standard calculus. Alternatively, we could use tools from information theory which we will discuss in the next section. ◇