- Unobserved Data
- Expectation Maximization: the Algorithm
- Expectation Maximization: an Example
- Expectation Maximization: the Derivation
  - Treatment 1
  - Treatment 2
  - Justification for E Step
  - Justification for M Step
- Expectation Maximization: an Illustration
  - E step round 1
  - M step round 1
  - E step round 2

Unobserved Data

On a sunny Silicon-Valley style afternoon during your internship at a fast-growing data analytics company, your boss told you to train a Bayesian Network over some dataset. You inspected the dataset and found some great news — the dataset only has four data records. So you quickly replied your boss:"No problem, I can do this!". The dataset comprises three features and you therefore decided to construct a simple three-node Bayesian network A\rightarrow B\rightarrow C. The model would learn its parameters by maximum likelihood estimation, which boils down to a grade-school counting problem on this lovely tiny dataset. Sweet!

Example	A	B	C
1	1	1	0
2	1	?	0
3	0	0	1
4	0	1	1

Wait a minute, something's wrong in the second data record — there's no value for feature B. Houston, we have a problem.

Expectation Maximization: the Algorithm

This internship anecdote in fact can be formulated into a general problem. Given a dataset containing observed (explicit) variable set \mathbf{X} and unobserved (latent) variable set \mathbf{Z}, how do we

model the joint distribution $P(\mathbf{X}, \mathbf{Z}|\mathbf{\theta})$ using a set of parameters $\mathbf{\theta}$?

The idea is simple, we (1) estimate the latent variable and (2) hopefully achieve a better estimation of parameters thereafter. This idea is captured by the Expectation Maximization algorithm — for each data point $\langle \mathbf{Z}_k, \mathbf{X}_k\rangle$ containing both observed and latent variable, in each iteration of the EM algorithm,

E step: compute

$$\forall k, P(\mathbf{Z}_k|\mathbf{X}_k, \mathbf{\theta}) = \frac{P(\mathbf{Z}_k,\mathbf{X}_k|\mathbf{\theta})}{P(\mathbf{X}_k|\mathbf{\theta})} = \frac{P(\mathbf{Z}_k,\mathbf{X}_k | \mathbf{\theta})}{\sum_zP(\mathbf{Z}_k=z,\mathbf{X}_k | \mathbf{\theta})}$$

In our context of boolean Bayesian network, observe that
$\mathbb{E}[\mathbf{Z}_k] = 1\cdot P(\mathbf{Z}_k = 1|\mathbf{X}_k, \mathbf{\theta}) + 0\cdot P(\mathbf{Z}_k = 0|\mathbf{X}_k, \mathbf{\theta}) =P(\mathbf{Z}_k = 1|\mathbf{X}_k, \mathbf{\theta})$

M step:

$$\mathbf{\theta} \leftarrow \arg\max_\mathbf{\theta'}\mathbb{E}_{P(\mathbf{Z}|\mathbf{X}, \mathbf{\theta})}\log P(\mathbf{Z},\mathbf{X}| \mathbf{\theta'})$$

The expressions look pretty horrifying, so let me elaborate on each step. The E step is nothing but calculating the posterior conditional probability using standard definition. To actually compute the value, however, in the context of Bayesian network, involves decompose the joint distribution based on the network topology. Therefore, we need to initialize a set of network parameters prior to the first iteration of our EM algorithm.

The second step turns out to resemble our grade-school counting problem to solve the MLE estimation for a fully observed Bayesian network. The only difference is that when the counting process reaches any data record containing latent variable, we count the "expected" appearance of that record instead of a hard 1 or 0. I know it's pretty abstract, so let me provide an example.

Expectation Maximization: an Example

After looking up the EM algorithm, you decided to handle your task described in the introductory section. To begin with, you initialized all of your network parameters to 0.6. Why 0.6? Well, to be honest, I don't know, maybe because six is a lucky number in the Chinese culture. In any case, you started your

adventure by computing \mathbb{E}[B_2], the expected value of feature B in the second record. According to the EM algorithm,

$$\mathbb{E}[B_2]= P(B_2 = 1|A_2 = 1, C_2 = 0, \theta) = \frac{P(B_2=1, A_2 = 1, C_2 = 0|\theta)}{P(B_2=1, A_2 = 1, C_2 = 0|\theta) + P(B_2=0, A_2 = 1, C_2 = 0|\theta)}$$

Since the network topology is A\rightarrow B\rightarrow C,

$$\begin{align*}\mathbb{E}[B_2]&=\frac{P(A_2 = 1)P(B_2 = 1 | A_2 = 1)P(C_2 = 0 | B_2 = 1)}{P(A_2 = 1)P(B_2 = 1 | A_2 = 1)P(C_2 = 0 | B_2 = 1) + P(A_2 = 1)P(B_2 = 0 | A_2 = 1)P(C_2 = 0 | B_2 = 0)}\\\\&=\frac{0.6 \cdot 0.6\cdot (1-0.6)}{0.6 \cdot 0.6\cdot (1-0.6) + 0.6 \cdot (1-0.6) \cdot (1-0.6)} = 0.6\end{align*}$$

Therefore, our "expected" dataset now becomes

Example	**A**	**B**	**C**
1	1	1	0
2	1	0.6	0
3	0	0	1
4	0	1	1

And we can proceed to the M step

$$\begin{align*}&P(A = 1) = \frac{\sum_{k = 1}^N\delta(A_k = 1)}{N} = 0.5 &\\\\&P(B = 1 | A = 1) = \frac{\sum_{k = 1}^N\delta(A_k = 1)\mathbb{E}[B_k]}{\sum_{k = 1}^N\delta(A_k = 1)} = 0.5&P(B = 1 | A = 1) = \frac{\sum_{k = 1}^N\delta(A_k = 1, B_k = 1)}{\sum_{k = 1}^N\delta(A_k = 1)}\\\\&P(B = 1 | A = 0) = \frac{\sum_{k = 1}^N\delta(A_k = 0)\mathbb{E}[B_k]}{\sum_{k = 1}^N\delta(A_k = 0)} = 0.8&P(B = 1 | A = 0) = \frac{\sum_{k = 1}^N\delta(A_k = 0, B_k = 1)}{\sum_{k = 1}^N\delta(A_k = 0)}\\\\&P(C = 1 | B = 1) = \frac{\sum_{k = 1}^N\delta(C_k = 1)\mathbb{E}[B_k]}{\sum_{k = 1}^N\mathbb{E}[B_k]} = 0.71&P(C = 1 | B = 1) = \frac{\sum_{k = 1}^N\delta(C_k = 1, B_k = 1)}{\sum_{k = 1}^N \delta(B_k = 1)} \\\\&P(C = 1 | B = 0) = \frac{\sum_{k = 1}^N\delta(C_k = 1)(1-\mathbb{E}[B_k])}{\sum_{k = 1}^N(1-\mathbb{E}[B_k])} = 0.38&P(C = 1 | B = 0) = \frac{\sum_{k = 1}^N\delta(C_k = 1, B_k = 1)}{\sum_{k = 1}^N\delta(B_k = 0)}\end{align*}$$

Note the difference between the M step (left column) and the naive MLE estimation as in the fully observed case (right column).

Expectation Maximization: the Derivation

Recall that in previous section, we say $\forall k,\mathbb{E}[\mathbf{Z}_k]= P(\mathbf{Z}_k|\mathbf{X}_k, \mathbf{\theta})$, but wait a second. Why is that? Here is the proof. Given complete data $\mathcal{D} = \Big(\langle\mathbf{Z}^{(1)}, \mathbf{X}^{(1)}\rangle, \cdots ,\langle\mathbf{Z}^{(N)}, \mathbf{X}^{(N)}\rangle\Big)$, we would like to find parameter θ such that the likelihood of observed data $\mathbf{X}$ is maximized. Formally,

$$\theta = \arg\max_\theta P(\mathbf{X} | \theta)$$

Following the ML convention, we will work with $\log P(\mathbf{X}|\theta)$. We will apply two algebraic transformations, respectively.

Treatment 1

Firstly, define arbitrary probability distribution over latent variable $\mathbf{Z}$ for every data record, $Q_i(\mathbf{Z}^{(i)})$, then

$$\begin{aligned}\log P(\mathbf{X|\theta}) &= \sum_{i = 1}^N\log P\big(\mathbf{X}^{(i)}|\theta\big) \\\log P\big(\mathbf{X}^{(i)}|\theta\big) &= \sum_{k = 1}^MQ_i\big(\mathbf{Z}^{(i)} = z_k\big)\log P\big(\mathbf{X}^{(i)}|\theta\big) = \mathbb{E}_{Q_i(\mathbf{Z}^{(i)})} \bigg[\log P\big(\mathbf{X}^{(i)}|\theta\big)\bigg]\\\Rightarrow \log P(\mathbf{X|\theta}) &= \sum_{i = 1}^N \mathbb{E}_{Q_i(\mathbf{Z}^{(i)})} \bigg[\log P\big(\mathbf{X}^{(i)}|\theta\big)\bigg]\end{aligned}$$

The crux is in the second equation, and it holds because $\log P\big(\mathbf{X}^{(i)}|\theta\big)$ is a constant — the expectation of a constant over any distribution is the constant itself.

Treatment 2

On the other hand, by marginalizing out the latent variables $\mathbf{Z}$,

$$\begin{aligned}\log P(\mathbf{X|\theta}) &= \sum_{i = 1}^N\log P\big(\mathbf{X}^{(i)}|\theta\big) = \sum_{i = 1}^N \log \sum_{k = 1}^M P\big(\mathbf{Z}^{(i) } = z_k, \mathbf{X}^{(i)}|\theta \big) \\\&= \sum_{i = 1}^N \log \sum_{k = 1}^M Q_i\big(\mathbf{Z}^{(i)}= z_k\big) \frac{P\big(\mathbf{Z}^{(i) } = z_k, \mathbf{X}^{(i)}|\theta \big) }{Q_i(\mathbf{Z}^{(i)}= z_k)}\\\\& = \sum_{i = 1}^N \log \mathbb{E}_{ Q_i(\mathbf{Z}^{(i)})} \Bigg[\frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta \big) } {Q_i(\mathbf{Z}^{(i)})}\Bigg]\\\&\geq \sum_{i = 1}^N \mathbb{E}_{ Q_i(\mathbf{Z}^{(i)})} \Bigg[\log\frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta \big) }{Q_i(\mathbf{Z}^{(i)})}\Bigg] \qquad \text{(by Jensen's inequality)}\\\end{aligned}$$

Justification for E Step

The reason for us to utilize Jensen's inequality is that working with logarithm of expectation is hard. Moreover, we want to compare and contrast our results from both treatments. Now, both treatment turn $\log P(\mathbf{X}|\theta)$ into *a summation of* **expectation of some logarithm term over an arbitrary distribution Q_i** *over the entire dataset*. However, the inequality condition introduced by Jensen's inequality adds extra fun. Let's combine results from our two treamtments.

$$\log P(\mathbf{X}|\theta) = \sum_{i = 1}^N \mathbb{E}_{Q_i(\mathbf{Z}^{(i)})} \bigg[\log P\big(\mathbf{X}^{(i)}|\theta\big)\bigg]\geq \sum_{i = 1}^N \mathbb{E}_{ Q_i(\mathbf{Z}^{(i)})} \Bigg[\log\frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta \big) }{Q_i(\mathbf{Z}^{(i)})}\Bigg]$$

If we speculate some quantity q such that

$$\begin{aligin*}&\sum_{i = 1}^N \mathbb{E}_{Q_i(\mathbf{Z}^{(i)})} \bigg[\log P\big(\mathbf{X}^{(i)}|\theta\big)\bigg] = \sum_{i = 1}^N \mathbb{E}_{ Q_i(\mathbf{Z}^{(i)})} \Bigg[\log\frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta \big)}{Q_i\big(\mathbf{Z}^{(i)}\big)}\Bigg] + q \\\\q = &\sum_{i = 1}^N \mathbb{E}_{Q_i(\mathbf{Z}^{(i)})} \bigg[\log P\big(\mathbf{X}^{(i)}|\theta\big)\bigg] - \sum_{i = 1}^N \mathbb{E}_{ Q_i(\mathbf{Z}^{(i)})} \Bigg[\log\frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta \big)} {Q_i\big(\mathbf{Z}^{(i)}\big)}\Bigg]\\\\=&-\sum_{i = 1}^N \mathbb{E}_{Q_i(\mathbf{Z}^{(i)})} \bigg[\log \frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta \big)}{P\big(\mathbf{X}^{(i)}|\theta\big) Q_i\big(\mathbf{Z}^{(i)}\big)} \bigg]\\\\=&-\sum_{i = 1}^N \mathbb{E}_{Q_i(\mathbf{Z}^{(i)})} \bigg[\log \frac{P\big(\mathbf{Z}^{(i) }| \mathbf{X}^{(i)},\theta \big)}{Q_i\big(\mathbf{Z}^{(i)}\big)} \bigg]\\\\=&\sum_{i = 1}^N D_{KL}\Big(Q_i\big(\mathbf{Z}^{(i)}\big) || P\big(\mathbf{Z}^{(i) }| \mathbf{X}^{(i)},\theta \big)\Big)\end{aligin*}$$

Note that this is exactly the sum of Kullback-Liebler divergence between $Q_i\big(\mathbf{Z}^{(i)}\big)$ and the posterior distribution $P\big(\mathbf{Z}^{(i) }| \mathbf{X}^{(i)},\theta \big)$. Therefore, the result from treatment 2 acts as a lower bound on our objective $P(\mathbf{X} | \theta)$, and it will equal to our original objective function if and only if q = 0. By definition of KL divergence, this means that equlity holds when $Q_i\big(\mathbf{Z}^{(i)}\big) = P\big(\mathbf{Z}^{(i) }| \mathbf{X}^{(i)},\theta \big)$. And that's exactly what the E step is doing.

Justification for M Step

By the end of the E step, the lower bound $\sum_{i = 1}^N \mathbb{E}_{ Q_i(\mathbf{Z}^{(i)})} \Bigg[\log\frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta \big) }{Q_i(\mathbf{Z}^{(i)})}\Bigg]$ coincides with our objective function by choosing Q_i wisely to be the posterior distribution. We need to show that the M step lifts this lower bound, thus effectively lifting the objective function as well.

By definition the M step seeks

$$\begin{aligned}\mathbf{\theta}' &=\arg\max_\mathbf{\theta'} \mathcal{L}(Q(\mathbf{Z}), \theta')\\\\\mathcal{L}(Q(\mathbf{Z}), \theta') &= \sum_{i = 1}^N \mathbb{E}_{ Q_i(\mathbf{Z}^{(i)})} \Bigg[\log\frac{P\big(\mathbf{Z}^{(i) }, \mathbf{X}^{(i)}|\theta' \big) }{Q_i(\mathbf{Z}^{(i)})}\Bigg]\geq \mathcal{L}(Q(\mathbf{Z}), \theta)\end{aligned}$$

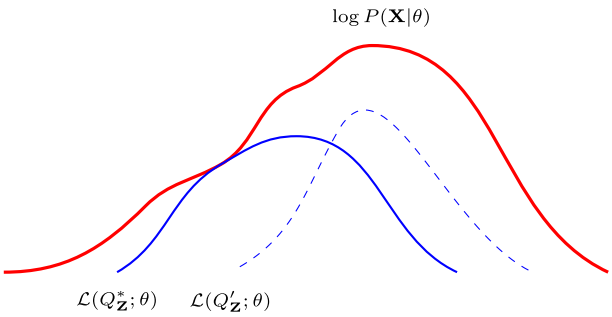
By choosing a different set of parameters, the posterior distribution changes, resulting in yet another non-negative KL divergence between $\mathcal{L}(Q(\mathbf{Z}), \theta')$ and $\mathcal{L}(Q(\mathbf{Z}), \theta)$. This discrepency will vanish during the E step of the next iteration. Therefore, by explicitly maximizing the lower bound, the M step will "implicitly" maximize $\log P(\mathbf{X|\theta})$ **at least as much as** it does with the lower bound.

Expectation Maximization: an Illustration

E step round 1

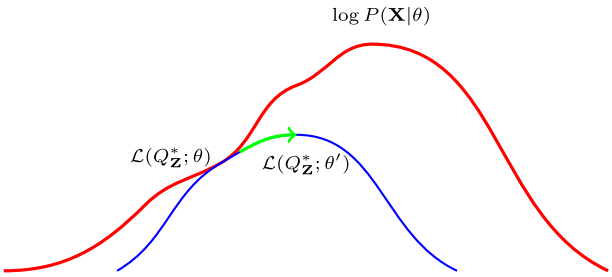
During the expectation step, the parameters θ are fixed. EM assigns the distribution $Q(\mathbf{Z})$ to be the posterior distribution as we discussed, such that the functional $\mathcal{L}(Q(\mathbf{Z}), \theta)$ is maximized. We denote this optimal distribution as $Q^*_{\mathbf{Z}}$.

To visualize this, imagine that there are many distribution $Q(\mathbf{Z})$ which are all lower bounds of our objective function $\log P(\mathbf{X}|\theta)$. The only one that satisfies $Q_i\big(\mathbf{Z}^{(i)}\big) = P\big(\mathbf{Z}^{(i) }| \mathbf{X}^{(i)},\theta \big)$ will coincide with the objective function given current parameter. This is our optimal distribution, shown in blue, while the remaining distributions are represented by the dashed curve.



M step round 1

During the maximization step, $Q^*_{\mathbf{Z}}$ is fixed and we optimize θ. This reflects in the visualization as the parameters "move" towards optimum θ', along the green arrow.



E step round 2

During the next expectation step, \theta' is fixed and we want to find a new optimal distribution now that the parameters have changed. The new optimum is denoted Q^{'*}_{\mathbf{Z}}. This time, Q^{'*}_{\mathbf{Z}} = P\big(\mathbf{Z}^{(i) }| \mathbf{X}^{(i)},\theta' \big), and we again increase the lower bound of the objective function, meantime leaving room for the following M step to further optimize since \theta' is not necessarily the optimal parameter under our new distribution.