CIS 520, Machine Learning, Fall 2018: Assignment 7

Shubhankar Patankar

November 28, 2018

Collaborators:

Simran Arora

1 Hidden Markov Models

For the following four probabilities, we want to find $P(X_{1:2} = (sing, TV)|z_{1:2} = (a, b))$ for states $a, b \in \{Happy,\ Sad\}$:

1. $(a,\ b) = (Happy,\ Happy)$

   The probability of the first state being Happy state is $P\{Z_1 = Happy\} = (\frac{1}{2})$. Given this first state, the $Pr(Sing,\ Happy) = \frac{5}{10}$. Using the transition probabilities $Pr(Z_{t+1} = Happy|Z_t = Happy) = \frac{4}{5}$ and the $Pr(TV|Happy) = \frac{2}{10}$, by the multiplication rule, the total probability of the situation $(a,\ b)$ is $(\frac{1}{2})(\frac{5}{10})(\frac{4}{5})(\frac{2}{10}) = \frac{40}{1000} = 0.04$

2. $(a,\ b) = (Happy,\ Sad)$

   The probability of the first state being a Happy state is $P\{Z_1 = Happy\} = (\frac{1}{2})$. Given this first state, the $Pr(Sing,\ Happy) = \frac{5}{10}$. Using the transition probabilities $Pr(Z_{t+1} = Sad|Z_t = Happy) = \frac{1}{5})$ and the $Pr(TV|Happy) = \frac{7}{10}$, by the multiplication rule, the total probability of the situation $(a,\ b)$ is $(\frac{1}{2})(\frac{5}{10})(\frac{1}{5})(\frac{7}{10}) = \frac{35}{1000} = 0.035$

3. $(a,\ b) = (Sad,\ Happy)$

   The probability of the first state being a Sad state is $P\{Z_1 = Sad\} = (\frac{1}{2})$. Given this first state, the $Pr(Sing,\ Sad) = \frac{1}{10}$. Using the transition probabilities $Pr(Z_{t+1} = Happy|Z_t = Sad) = \frac{1}{2})$ and the $Pr(TV|Happy) = \frac{2}{10}$, by the multiplication rule, the total probability of the situation $(a,\ b)$ is $(\frac{1}{2})(\frac{1}{10})(\frac{1}{2})(\frac{2}{10}) = \frac{2}{400} = 0.005$

4. $(a,\ b) = (Sad,\ Sad)$

   The probability of the first state being a sad state is $P\{Z_1 = Sad\} = (\frac{1}{2})$. Given this first state, the $Pr(Sing,\ Sad) = \frac{1}{10}$. Using the transition probabilities $Pr(Z_{t+1} = Sad|Z_t = Sad) = \frac{1}{2})$ and the $Pr(TV|Sad) = \frac{7}{10}$, by the multiplication rule, the total probability of the situation $(a,\ b)$ is $(\frac{1}{2})(\frac{1}{10})(\frac{1}{2})(\frac{7}{10}) = \frac{7}{400} = 0.0175$

Based on these probabilities, the most likely hidden state sequence $Z_{1:2} = (Happy,\ Happy)$, with probability 0.04. Individually, if $(sing, TV)$ are observed then the most likely state for day 2 is Sad because the $Pr(Sad, Sad) + Pr(Happy, Sad) > Pr(Sad, Happy) + Pr(Happy, Happy)$. The probability of day 2 being Sad is $0.035 + 0.0175 = 0.0525$, while the probability of day 2 being Happy is $0.04 + 0.005 = 0.045$.

Without any observed actions on the days (i.e. ignoring the fact that (Sing and TV) are observed), from the onset, the most likely hidden state on day 2 is computed by the total probability formula as follows. $P(Z_2 = Happy) = Pr(Z_2 = Happy|Z_1 = Happy)Pr(Z_1 = Happy) + Pr(Z_2 = Happy|Z_1 = Sad)Pr(Z_1 = Sad) = (\frac{4}{5})(\frac{1}{2}) + (\frac{1}{2})(\frac{1}{2}) = \frac{4}{10} + \frac{1}{4} = 0.65$. Meanwhile $P(Z_2 = Sad) = Pr(Z_2 = Sad|Z_1 = Happy)Pr(Z_1 = Happy) + Pr(Z_2 = Sad|Z_1 = Sad)Pr(Z_1 = Sad) = (\frac{1}{5})(\frac{1}{2}) + (\frac{1}{2})(\frac{1}{2}) = \frac{1}{10} + \frac{1}{4} = 0.35$. Thus we can see that the individually most likely hidden state on day 2 is the Happy state.

2 Missing Data

1. $$Test\ \ Accuracy\ \ Full = 0.9600$$

2. **Random NaNs**

   Method 1: $Test\ \ Accuracy = 0.9200$
   Method 2: $Test\ \ Accuracy = 0.8667$

3. **Non-Random NaNs**

   Method 1: $Test\ \ Accuracy = 0.9333$
   Method 2: $Test\ \ Accuracy = 0.9467$

4. Method 1 (mean imputation) is more accurate than Method 2 (indicator variables) for randomly missing data. Method 2 is more accurate for data that is not missing at random. Both methods are less accurate overall than the case when no data is missing. Imputing the mean when the data is not missing at random adds noise to the data making Method 1 less desirable to Method 2, which makes no assumptions about the nature of the missing values. On the other hand, for randomly missing data, adding indicators is less useful than imputing the means because the means of a feature resemble the missing values more than they do when the data is not missing at random. The lack of order in randomly missing data is conducive to mean imputation. Both methods outperform the case where neither Method 1 nor Method 2 is employed, because both methods provide some additional information that is unavailable in the original data. The additional information can only help the ML method being used.

3 Bayesian Networks

1. $P(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_2)p(x_6|x_3, x_4)p(x_5|x_3)$

2. Yes. The joint probability distribution P' in 3.2 can be represented in the class of joint probability distributions P given by the Bayesian network. In P' we have a subset of P because P' takes P without the dependency of X_6 on X_4; without the dependency of X_3 on X_1 or X_2; and without the dependency of X_4 on X_2. Since P' has more independent random variable relationships (more conditional independencies), it is a subset of P. Some of the conditional dependencies in P are redundant in the case of P'. P is general and is able to represent the given joint probability distribution in 3.2.

3. Smaller. Without the edge from X_3 to X_6 we have $X_6 \perp\!\!\!\perp X_2|X_4$ instead of $X_6 \perp\!\!\!\perp X_2|X_4, X_3, X_1$. The resulting network will represent a narrower class of distributions than the original network because of the fewer dependencies. There are more conditional independencies associated with the resulting network so the resulting network will be smaller than the original network.

4. (a) This is **FALSE** because there is an active trail for $X_3 \leftarrow X_2 \rightarrow X_4$ where X_2 is not observed. In $D - Separation$ we know that variable X_i and X_j are independent if and only if there is no active trail between X_i and X_j. Thus we have that X_3 and X_4 are not independent unless we have observed X_2 to remove the active trail.

   (b) This is **TRUE** because we know by the Local Markov Assumption that a variable X is independent of its non-descendants given its parents. In this case, X_1 has no parents and X_4 is a non-descendant. Thus, we have that X_1 and X_4 are independent and by definition of independence, this means that $p(x_1, x_4) = p(x_1)p(x_4)$

   (c) This is **FALSE** for $p(x_1, x_2|x_6)$. There is an active trail $X_1 \rightarrow X_3 \rightarrow X_6$ and there is an active trail $X_2 \rightarrow X_3 \rightarrow X_6$. We cannot say that X_1 and X_2 are conditionally independent given X_6.

   (d) This is **FALSE** for $p(x_1, x_6|x_3)$. There is an active path from X_1 to X_6. Since X_3 is observed, X_1 and X_2 are dependent. Since X_3 is observed, X_3 and X_4 have an active trail through X_2. Since X_4 is not observed, X_2 and X_6 are dependent. Since both X_1 and X_6 have active trails from X_2 and all consecutive triples in X_1 to X_3 to X_2 to X_4 to X_6 have active trails, X_1 and X_6 have an active trail between each other.

4 Belief Net Construction

For this problem, two probabilities are assumed to be equal if they are within 0.05 of each other.

1. **Determining Dependencies**

   First add A to the belief net. Then add B. Check if B has a trail from A:

$$P(B \mid A) = \frac{1600}{3000} = 0.5333$$

$$P(B \mid \neg A) = \frac{800}{1400} = 0.5714$$

$$P(B) = \frac{2400}{4400} = 0.5455$$

$$|P(B \mid A) - P(B)| = 0.0122 < 0.05$$

$$|P(B \mid \neg A) - P(B)| = 0.0259 < 0.05$$

   Therefore, B does not depend on A. Next, test C.

$$P(C \mid A, B) = \frac{800}{1600} = 0.5$$

$$P(C \mid \neg A, B) = \frac{600}{800} = 0.75$$

$$P(C \mid A, \neg B) = \frac{1200}{1400} = 0.8571$$

$$P(C \mid \neg A, \neg B) = \frac{400}{600} = 0.6667$$

$$P(C) = \frac{3000}{4400} = 0.6818$$

$$|P(C \mid A, B) - P(C)| = 0.1818 > 0.05$$

$$|P(C \mid \neg A, B) - P(C)| = 0.0682 > 0.05$$

$$|P(C \mid A, \neg B) - P(C)| = 0.1753 > 0.05$$

$$|P(C \mid \neg A, \neg B) - P(C)| = 0.0151 < 0.05$$

   Since $P(C \mid A, B) \neq P(C \mid \neg A, B) \neq P(C \mid A, \neg B) \neq P(C)$, C is linked to the preceding variables.
   Primary test for C's link with A:

$$P(C \mid A) = \frac{2000}{3000} = 0.6667$$

$$P(C \mid \neg A) = \frac{1000}{1400} = 0.7143$$

$$P(C) = 0.6818$$

$$|P(C \mid A) - P(C)| = 0.0151 < 0.05$$

$$|P(C \mid \neg A) - P(C)| = 0.0325 < 0.05$$

   Therefore, seemingly C does not appear to depend on A.
   Primary test for C's link with B:

$$P(C \mid B) = \frac{1400}{2400} = 0.5833$$

$$P(C \mid \neg B) = \frac{1600}{2000} = 0.8$$

$$P(C) = 0.6818$$

$$|P(C \mid B) - P(C)| = 0.0985 > 0.05$$
$$|P(C \mid \neg B) - P(C)| = 0.1182 > 0.05$$

Therefore, C depends on B. Secondary test for joint dependence of C on B with A.

$$|P(C \mid B, A) - P(C \mid B)| = 0.0833 > 0.05$$
$$|P(C \mid B, \neg A) - P(C \mid B)| = 0.1667 > 0.05$$

Since $P(C \mid B, A) \neq P(C \mid \neg A, B) \neq P(C \mid B)$, knowing something about A tells us something about C conditioned on B. Therefore a link is present between A and C. The same inference can be drawn from the inequalities below.

$$|P(C \mid A, \neg B) - P(C \mid \neg B)| = 0.0571 > 0.05$$
$$|P(C \mid \neg A, \neg B) - P(C \mid \neg B)| = 0.1333 > 0.05$$

Next, test D.

$$P(D \mid A, B, C) = \frac{600}{800} = 0.75$$
$$P(D \mid \neg A, B, C) = \frac{200}{600} = 0.3333$$
$$P(D \mid A, \neg B, C) = \frac{400}{1200} = 0.3333$$
$$P(D \mid A, B, \neg C) = \frac{0}{800} = 0$$
$$P(D \mid \neg A, \neg B, C) = \frac{200}{500} = 0.5$$
$$P(D \mid A, \neg B, \neg C) = \frac{0}{200} = 0$$
$$P(D \mid \neg A, B, \neg C) = \frac{0}{200} = 0$$
$$P(D \mid \neg A, \neg B, \neg C) = \frac{0}{200} = 0$$
$$P(D) = \frac{1400}{4400} = 0.3182$$

$$|P(D \mid A, B, C) - P(D)| = 0.4318 > 0.05$$
$$|P(D \mid \neg A, B, C) - P(D)| = 0.0148 < 0.05$$
$$|P(D \mid A, \neg B, C) - P(D)| = 0.0151 < 0.05$$
$$|P(D \mid A, B, \neg C) - P(D)| = 0.3182 > 0.05$$
$$|P(D \mid \neg A, \neg B, C) - P(D)| = 0.1818 > 0.05$$
$$|P(D \mid A, \neg B, \neg C) - P(D)| = 0.3182 > 0.05$$
$$|P(D \mid \neg A, B, \neg C) - P(D)| = 0.3182 > 0.05$$
$$|P(D \mid \neg A, \neg B, \neg C) - P(D)| = 0.3182 > 0.05$$

Since all terms in the differences above are not equal to $P(D)$, D depends on the previously added variables.

Primary test for D's link with A:

$$P(D \mid A) = \frac{1000}{3000} = 0.3333$$

$$P(D \mid \neg A) = \frac{400}{1400} = 0.2857$$

$$|P(D \mid A) - P(D)| = 0.0151 < 0.05$$

$$|P(D \mid \neg A) - P(D)| = 0.0325 < 0.05$$

Therefore, seemingly D does not appear to depend on A.

Primary test for D's link with B:

$$P(D \mid B) = \frac{800}{2400} = 0.3333$$

$$P(D \mid \neg B) = \frac{600}{2000} = 0.3$$

$$|P(D \mid B) - P(D)| = 0.0151 < 0.05$$

$$|P(D \mid \neg B) - P(D)| = 0.0182 < 0.05$$

Therefore, seemingly D does not appear to depend on B.

Primary test for D's link with C:

$$P(D \mid C) = \frac{1400}{3000} = 0.4667$$

$$P(D \mid \neg C) = \frac{0}{1400} = 0.3182$$

$$|P(D \mid C) - P(C)| = 0.1485 > 0.05$$

$$|P(D \mid \neg C) - P(C)| = 0.3182 > 0.05$$

Therefore, D depends on C.

Secondary test for joint dependence of D on C with A and/or B.

$$P(D \mid A, C) = \frac{1000}{2000} = 0.5$$

$$P(D \mid B, C) = \frac{800}{1400} = 0.5714$$

$$|P(D \mid A, C) - P(D \mid C)| = 0.0333 < 0.05$$

$$|P(D \mid B, C) - P(D \mid C)| = 0.1047 > 0.05$$

Therefore, knowing B tells us more about D conditioned on C, implying that B and D are linked.

$$P(D \mid \neg A, C) = \frac{400}{1000} = 0.4$$

$$|P(D \mid \neg A, C) - P(D \mid C)| = 0.0667 > 0.05$$

$$|P(C \mid B, \neg A) - P(C \mid B)| = 0.1667 > 0.05$$

Therefore, knowing A tells us more about D conditioned on C, implying that A and D are also linked.

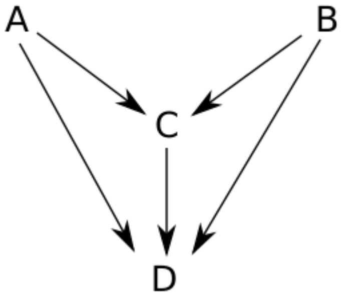
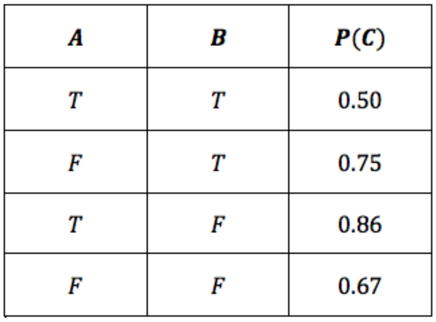


Figure 1: Constructed Belief Net

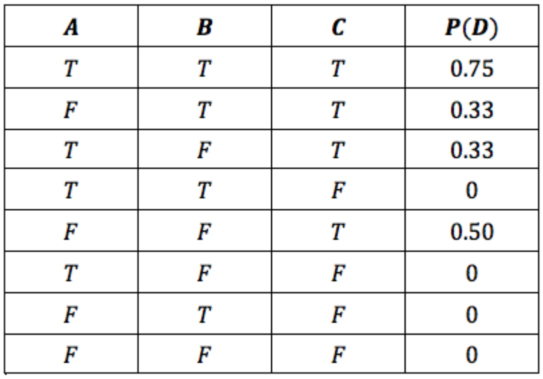
Conditional Probability Tables

The table below outlines the probabilities of C given values of A and B.



A	*B*	*P*(*C*)
T	*T*	0.50
F	*T*	0.75
T	*F*	0.86
F	*F*	0.67

The table below outlines the probabilities of C given values of A, B and C.



A	*B*	*C*	*P*(*D*)
T	*T*	*T*	0.75
F	*T*	*T*	0.33
T	*F*	*T*	0.33
T	*T*	*F*	0
F	*F*	*T*	0.50
T	*F*	*F*	0
F	*T*	*F*	0
F	*F*	*F*	0

$P(A) = 0.6818$, $P(B) = 0.5455$, $P(C) = 0.6818$

5 EM

1. **Model Parameters**

$$p(D_i, C_i, A_i, B_i) = p(D_i|C_i, A_i, B_i) \times p(C_i|A_i, B_i) \times p(A_i|B_i) \times p(B_i)$$

However, based on the belief net structure,

$$p(D_i, C_i, A_i, B_i) = p(D_i|C_i) \times p(C_i|A_i, B_i) \times p(A_i) \times p(B_i)$$

Therefore the model parameters $\boldsymbol{\theta}$ are:
$p(D_i|C_i)$, $p(D_i|\neg C_i)$, $p(C_i|A_i, B_i)$, $p(C_i|\neg A_i, B_i)$, $p(C_i|A_i, \neg B_i)$, $p(C_i|\neg A_i, \neg B_i)$, $p(A_i)$ and $p(B_i)$.

$$p(D_i \mid C_i = 1) = \begin{cases} p_1 & D_i = 1 \\ 1 - p_1 & D_i = 0 \end{cases}$$

$$\therefore p(D_i \mid C_i = 1) = {p_1}^{D_i}(1 - p_1)^{1-D_i}$$

$$p(D_i \mid C_i = 0) = \begin{cases} p_2 & D_i = 1 \\ 1 - p_2 & D_i = 0 \end{cases}$$

$$\therefore p(D_i \mid C_i = 0) = {p_2}^{D_i}(1 - p_2)^{1-D_i}$$

$$p(C_i \mid A_i = 1, B_i = 1) = \begin{cases} p_3 & C_i = 1 \\ 1 - p_3 & C_i = 0 \end{cases}$$

$$\therefore p(C_i \mid A_i = 1, B_i = 1) = {p_3}^{C_i}(1 - p_3)^{1-C_i}$$

$$p(C_i \mid A_i = 0, B_i = 1) = \begin{cases} p_4 & C_i = 1 \\ 1 - p_4 & C_i = 0 \end{cases}$$

$$\therefore p(C_i \mid A_i = 0, B_i = 1) = {p_4}^{C_i}(1 - p_4)^{1-C_i}$$

$$p(C \mid A_i = 1, B_i = 0) = \begin{cases} p_5 & C_i = 1 \\ 1 - p_5 & C_i = 0 \end{cases}$$

$$\therefore p(C_i \mid A_i = 1, B_i = 0) = {p_5}^{C_i}(1 - p_5)^{1-C_i}$$

$$p(C \mid A_i = 0, B_i = 0) = \begin{cases} p_6 & C_i = 1 \\ 1 - p_6 & C_i = 0 \end{cases}$$

$$\therefore p(C_i \mid A_i = 0, B_i = 0) = {p_6}^{C_i}(1 - p_6)^{1-C_i}$$

$$p(A_i) = \begin{cases} p_7 & A_i = 1 \\ 1 - p_7 & A_i = 0 \end{cases}$$

$$\therefore p(A_i) = {p_7}^{A_i}(1 - p_7)^{1-A_i}$$

$$p(B_i) = \begin{cases} p_8 & B_i = 1 \\ 1 - p_8 & B_i = 0 \end{cases}$$

$$\therefore p(B_i) = {p_8}^{B_i}(1 - p_8)^{1-B_i}$$

2. **E-Step**

For the given belief net, the observed states are $\mathbf{x_i} = \{A_i, B_i, D_i\}$.

$$C_i = z \in \{0, 1\}$$

The full joint distribution including the hidden state C_i can be written as follows:

$$p(D_i, C_i, A_i, B_i) = \left[{p_1}^{D_i}(1-p_1)^{1-D_i}\right]^{C_i} \times \left[{p_2}^{D_i}(1-p_2)^{1-D_i}\right]^{1-C_i} \times$$

$$\left[{p_3}^{C_i}(1-p_3)^{1-C_i}\right]^{A_i B_i} \times \left[{p_4}^{C_i}(1-p_4)^{1-C_i}\right]^{(1-A_i)B_i} \times \left[{p_5}^{C_i}(1-p_5)^{1-C_i}\right]^{(1-B_i)A_i} \times$$

$$\left[{p_6}^{C_i}(1-p_6)^{1-C_i}\right]^{(1-A_i)(1-B_i)} \times \left[{p_7}^{A_i}(1-p_7)^{1-A_i}\right] \times \left[{p_8}^{B_i}(1-p_8)^{1-B_i}\right]$$

$$\therefore p(C_i = z \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1}) = \frac{p(\mathbf{x_i} \mid C_i = z)p(C_i = z)}{p(\mathbf{x_i})} = \frac{p(\mathbf{x_i}, C_i = z)}{\sum_{C_i} p(\mathbf{x_i}, C_i = z)}$$

Equivalently, this can be written as:

$$p(C_i = z \mid \mathbf{x_i}, \theta^{t-1}) = \frac{p(\mathbf{x_i}, C_i = z \mid \theta^{t-1})}{p(\mathbf{x_i}, C_i = 1 \mid \theta^{t-1}) + p(\mathbf{x_i}, C_i = 0 \mid \theta^{t-1})} = \frac{p(\mathbf{x_i}, C_i = z \mid \theta^{t-1})}{\sum_{C_i} p(\mathbf{x_i}, C_i = z)}$$

Or as:

$$p(C_i = z \mid \mathbf{x_i}, \theta^{t-1}) = \frac{p(D_i, C_i = z, A_i, B_i)}{\sum_{C_i} p(D_i, C_i = z, A_i, B_i)}$$

3. **M-Step**

$$\boldsymbol{\theta}^t = argmax_{\boldsymbol{\theta}}\ q(\boldsymbol{\theta}, \boldsymbol{\theta^{t-1}})$$

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \sum_i \sum_{C_i} p(C_i \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1}) \times ln\ p(\mathbf{x_i}, C_i \mid \boldsymbol{\theta})$$

$$= \sum_i \left[\left[p(C_i = 1 \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1})\ ln\ p(\mathbf{x_i}, C_i = 1 \mid \boldsymbol{\theta})\right] + \left[p(C_i = 0 \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1})\ ln\ p(\mathbf{x_i}, C_i = 0 \mid \boldsymbol{\theta})\right]\right]$$

$$\because p(D_i, C_i, A_i, B_i) = p(D_i|C_i) \times p(C_i|A_i, B_i) \times p(A_i) \times p(B_i)$$

$$\implies ln\ p(\mathbf{x_i}, C_i = z \mid \boldsymbol{\theta}) = ln\ P(D_i|C_i = z) + ln\ P(C_i = z|A_i, B_i) + ln\ P(A_i) + ln\ P(B_i)$$

Each of the ln terms in $q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$ can be substituted for with the above expression, appropriately replacing z.

4. **(C, D) or (C and D) Missing**

Since C is missing for the first half of the samples, the same updates as above can be used for $i = 1 : (1/2)n$, where n is the total number of samples.

$$i = 1 : (1/2)n$$

$$C_i = z \in \{0, 1\}$$

$$\mathbf{x_i} = \{A_i, B_i, D_i\}$$

E-Step:

$$p(C_i = z \mid \mathbf{x_i}, \theta^{t-1}) = \frac{p(\mathbf{x_i}, C_i = z \mid \theta^{t-1})}{p(\mathbf{x_i}, C_i = 1 \mid \theta^{t-1}) + p(\mathbf{x_i}, C_i = 0 \mid \theta^{t-1})} = \frac{p(\mathbf{x_i}, C_i = z \mid \theta^{t-1})}{\sum_{C_i} p(\mathbf{x_i}, C_i = z)}$$

M-Step:

$$\boldsymbol{\theta}^t = argmax_{\boldsymbol{\theta}}\ q(\boldsymbol{\theta}, \boldsymbol{\theta^{t-1}})$$

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \sum_i \sum_{C_i} p(C_i = z \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1}) \times ln\ p(\mathbf{x_i}, C_i = z \mid \boldsymbol{\theta})$$

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \sum_i \Bigg[\big[p(C_i = 1 \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1})\ ln\ p(\mathbf{x_i}, C_i = 1 \mid \boldsymbol{\theta})\big] + \big[p(C_i = 0 \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1})\ ln\ p(\mathbf{x_i}, C_i = 0 \mid \boldsymbol{\theta})\big] \Bigg]$$

$$\because ln\ p(\mathbf{x_i}, C_i = z \mid \boldsymbol{\theta}) = ln\ P(D_i | C_i = z) + ln\ P(C_i = z | A_i, B_i) + ln\ P(A_i) + ln\ P(B_i)$$

Each of the ln terms in $q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$ can be substituted for with the above expression, appropriately replacing z.

D is missing for the next quarter of the samples.

$$i = (1/2)n : (3/4)n$$

$$D_i = w \in \{0, 1\}$$

$$\mathbf{x_i} = \{A_i, B_i, C_i\}$$

E-Step:

$$p(D_i = z \mid \mathbf{x_i}, \theta^{t-1}) = \frac{p(\mathbf{x_i}, D_i = w \mid \theta^{t-1})}{p(\mathbf{x_i}, D_i = 1 \mid \theta^{t-1}) + p(\mathbf{x_i}, D_i = 0 \mid \theta^{t-1})}$$

M-Step:

$$\boldsymbol{\theta}^t = argmax_{\boldsymbol{\theta}}\ q(\boldsymbol{\theta}, \boldsymbol{\theta^{t-1}})$$

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \sum_i \sum_{D_i} p(D_i = w \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1}) \times ln\ p(\mathbf{x_i}, D_i = w \mid \boldsymbol{\theta})$$

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \sum_i \Bigg[\big[p(D_i = 1 \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1})\ ln\ p(\mathbf{x_i}, D_i = 1 \mid \boldsymbol{\theta})\big] + \big[p(D_i = 0 \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1})\ ln\ p(\mathbf{x_i}, D_i = 0 \mid \boldsymbol{\theta})\big] \Bigg]$$

$$\because ln\ p(\mathbf{x_i}, D_i = w \mid \boldsymbol{\theta}) = ln\ P(D_i = w | C_i) + ln\ P(C_i | A_i, B_i) + ln\ P(A_i) + ln\ P(B_i)$$

Each of the ln terms in $q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$ can be substituted for with the above expression, appropriately replacing w.

Both C and D are missing for the final quarter of the samples.

$$i = (3/4)n : n$$

$$\mathbf{x_i} = \{A_i, B_i\}$$

$$C_i = z \in \{0, 1\}$$

$$D_i = w \in \{0, 1\}$$

E-Step:

$$p(C_i = z, D_i = w \mid \mathbf{x_i}, \theta^{t-1}) = \frac{p(\mathbf{x_i}, C_i = zD_i = w \mid \theta^{t-1})}{\sum_{C_i} \sum_{D_i} p(\mathbf{x_i}, C_i, D_i)}$$

M-Step:

$$\boldsymbol{\theta}^t = argmax_{\boldsymbol{\theta}}\ q(\boldsymbol{\theta}, \boldsymbol{\theta^{t-1}})$$

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \sum_i \sum_{C_i} \sum_{D_i} p(C_i = z, D_i = w \mid \mathbf{x_i}, \boldsymbol{\theta}^{t-1}) \times ln\ p(\mathbf{x_i}, C_i = z, D_i = w \mid \boldsymbol{\theta})$$

$$\because ln\ p(\mathbf{x_i}, C_i = z, D_i = w \mid \boldsymbol{\theta}) = ln\ P(D_i = w | C_i = z) + ln\ P(C_i = z | A_i, B_i) + ln\ P(A_i) + ln\ P(B_i)$$

Each of the ln terms in $q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$ can be substituted for with the above expression, appropriately replacing z and w for the possible values of C_i and D_i, respectively.