Problem 1. (SUM–PRODUCT Algorithm)

(a) Fig. 1(a) contains a drawing of this graphical model, known as the “Hidden Markov Model.” The $X_i$ variables form a Markov chain backbone and are usually not directly observed. Instead, we indirect measurements $Y_i$, from which we wish to infer the states of $X_i$. The probability distribution is

$$P(X_1\ldots X_T, Y_1\ldots Y_T) = P(X_1)P(Y_1|X_1) \prod_{i=2}^{T} P(X_i|X_{i-1})P(Y_i|X_i).$$

(b) Define

$$A = \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 1/3 & 1/3 & 1/3 \\ 1/6 & 1/3 & 1/2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad p_1 = \begin{bmatrix} 1/2 \\ 1/4 \end{bmatrix}.$$

Also, let $e_i$ denote the indicator vector having 1 at the $i$-th position and zero everywhere else. The potential functions can be written as

$$\Psi(X_i) = \begin{cases} p_1, & \text{if } i = 1 \\ 1, & \text{o.w.} \end{cases}$$

$$\Psi(Y_i) = 1;$$

$$\Psi^b(Y_i) = e_{y_{i+1}};$$

$$\Psi(X_{i-1}, X_i) = P(X_i|X_{i-1}) = A;$$

$$\Psi(X_i, Y_i) = P(Y_i|X_i) = B.$$ 

Recall the formula for the message from node $i$ to node $j$.

$$m_{ji}(x_i) = \sum_{x_j} \Psi^b(x_j)\Psi(x_i, x_j) \prod_{k \in \mathcal{N}(j) \setminus i} m_{kj}(x_j).$$

Using Matlab notation, we'll use `.*` to denote element-wise array multiplication, that is to say, $[a \ b] \ast [c \ d] = [a \ast c \ b \ast d]$. WLOG, let's take $X_1$
to be the root. We can write the messages in matrix notation.

\[ m_{y_i|x_i} = B * e_{y_i+1}; \]
\[ m_{x_i|x_{i-1}} = A * (m_{x_{i+1}|x_i} * m_{y_{i+1}|x_i}); \]
\[ m_{x_1|x_2} = A^T * (p_1 * m_{y_1|x_1}); \]
\[ m_{x_2|x_3} = A^T * (m_{x_1|x_2} * m_{y_2|x_2}); \]
\[ m_{x_1|y_1} = B^T * (p_1 * m_{x_2|x_1}); \]
\[ m_{x_2|y_2} = B^T * (m_{x_1|x_2} * m_{x_3|x_2}); \]
\[ m_{x_3|y_3} = B^T * m_{x_2|x_3}. \]

Technically there are also messages from \( X_i, Y_i, i = 4 \ldots T \). But since they are all unobserved leaf nodes, their messages are just the all-one vectors.

\[
\begin{align*}
    m_{y_1|x_1} &= \begin{bmatrix} 1 \\ 0 \\
    \end{bmatrix}, &
    m_{y_2|x_2} &= \begin{bmatrix} 1 \\ 0 \\
    \end{bmatrix}, &
    m_{y_3|x_3} &= \begin{bmatrix} 0 \\ 1 \\
    \end{bmatrix}, \\

    m_{x_3|x_2} &= \begin{bmatrix} 0.1667 \\ 0.3333 \\ 0.5 \\
    \end{bmatrix}, &
    m_{x_2|x_1} &= \begin{bmatrix} 0.1944 \\ 0.1667 \\ 0.1389 \\
    \end{bmatrix}, \\

    m_{x_1|x_2} &= \begin{bmatrix} 0.3333 \\ 0.2500 \\ 0.1667 \\
    \end{bmatrix}, &
    m_{x_2|x_3} &= \begin{bmatrix} 0.2500 \\ 0.1944 \\ 0.1389 \\
    \end{bmatrix}, \\

    m_{x_1|y_1} &= \begin{bmatrix} 0.1389 \\ 0.0347 \\
    \end{bmatrix}, &
    m_{x_2|y_2} &= \begin{bmatrix} 0.1389 \\ 0.0833 \\
    \end{bmatrix}, &
    m_{x_3|y_3} &= \begin{bmatrix} 0.4444 \\ 0.1389 \\
    \end{bmatrix},
\end{align*}
\]

Finally,

\[
\begin{align*}
    p(x_1|\bar{y}_1, \bar{y}_2, \bar{y}_3) &= \begin{bmatrix} 0.7 \\ 0.3 \\ 0 \\
    \end{bmatrix}, &
    p(x_2|\bar{y}_1, \bar{y}_2, \bar{y}_3) &= \begin{bmatrix} 0.4 \\ 0.6 \\ 0 \\
    \end{bmatrix}, &
    p(x_3|\bar{y}_1, \bar{y}_2, \bar{y}_3) &= \begin{bmatrix} 0 \\ 0 \\ 1 \\
    \end{bmatrix}.
\end{align*}
\]

(c) See Fig. 1(b) for the graph. The probability distribution is

\[
P(X_{1...T}, Y_{1...T}, Z_{1...T}) = P(X_1) \prod_{i=2}^T P(X_i|X_{i-1}) \prod_{i=1}^T P(Y_i|X_i)P(Z_i|Y_i).
\]
Figure 1: (a) The HMM model. (b) The HMM model with an additional hidden layer. (c) An equivalent model to (a) when the transition matrix is the identity matrix. (d) An equivalent model to (a) when the $X_i$'s are independent.
(d) For continuous variables, the summation turns into integration when computing the message.

\[ m_{z_i|y_i} = \int_{z_i} \delta(z_i = \bar{z}_i) P(z_i|y_i) dz_i = P(\bar{z}_i|y_i) = \frac{1}{2\pi} \exp^{-\frac{1}{2}(\bar{z}_i - y_i)^2}. \]  

(9)

The other messages are as in Eqns. (2-8), but with the \( m_{z_i|y_i} \) messages replacing the evidence potentials on \( Y_i \).

\[
m_{x_1^{} y_1^{} } = \begin{bmatrix} 0.1405 \\ 0.1405 \\ 0.1405 \end{bmatrix}, \quad m_{x_2^{} y_2^{} } = \begin{bmatrix} 0.1405 \\ 0.0517 \\ 0.0965 \end{bmatrix}, \quad m_{x_3^{} y_3^{} } = \begin{bmatrix} 0.0215 \\ 0.0215 \\ 0.0215 \end{bmatrix},
\]

\[
m_{y_1^{} x_1^{} } = \begin{bmatrix} 0.1405 \\ 0.1405 \\ 0.1405 \end{bmatrix}, \quad m_{y_2^{} x_2^{} } = \begin{bmatrix} 0.1405 \\ 0.0517 \\ 0.0965 \end{bmatrix}, \quad m_{y_3^{} x_3^{} } = \begin{bmatrix} 0.0215 \\ 0.0215 \\ 0.0215 \end{bmatrix},
\]

\[
m_{x_3^{} x_2^{} } = \begin{bmatrix} 0.0340 \\ 0.0465 \\ 0.0590 \end{bmatrix}, \quad m_{x_2^{} x_1^{} } = \begin{bmatrix} 0.0051 \\ 0.0048 \\ 0.0045 \end{bmatrix},
\]

\[
m_{x_1^{} x_2^{} } = \begin{bmatrix} 0.0527 \\ 0.0468 \\ 0.0410 \end{bmatrix}, \quad m_{x_2^{} x_3^{} } = \begin{bmatrix} 0.0062 \\ 0.0054 \\ 0.0045 \end{bmatrix},
\]

\[
m_{x_1^{} y_1^{} } = \begin{bmatrix} 0.0037 \\ 0.0011 \end{bmatrix}, \quad m_{x_2^{} y_2^{} } = \begin{bmatrix} 0.0040 \\ 0.0024 \end{bmatrix}, \quad m_{x_3^{} y_3^{} } = \begin{bmatrix} 0.0116 \\ 0.0045 \end{bmatrix},
\]

The conditional marginals are

\[
p(x_1|\bar{y}_{1,2,3}) = \begin{bmatrix} 0.5222 \\ 0.2463 \\ 0.2315 \end{bmatrix}, \quad p(x_2|\bar{y}_{1,2,3}) = \begin{bmatrix} 0.3688 \\ 0.4482 \\ 0.1830 \end{bmatrix}, \quad p(x_3|\bar{y}_{1,2,3}) = \begin{bmatrix} 0.1970 \\ 0.1692 \\ 0.6338 \end{bmatrix}.
\]

(e) If the transition matrix is the identity matrix, then the \( X_i \) variables are all equal to each other with probability 1. Hence a simpler model would just have one \( X \) node and conditionally independent \( Y_i \) nodes. See Fig. 1(c). Since the \( X_i \)'s cannot be unequal, and \( Y_i \) is deterministic once we know \( X_i \), the \( Y \) values are also equal with probability 1. Hence the current configuration of \( Y_{1,2,3} = [0 \ 0 \ 1] \) has probability zero, which manifests itself in the all-zero
messages from $X$ to $Y_1$ and $Y_2$.

$$m_{y_1x_1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad m_{y_2x_2} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad m_{y_3x_3} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

$$m_{xy_1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, m_{xy_2} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, m_{xy_3} = \begin{bmatrix} 0.75 \\ 0.75 \\ 0.75 \end{bmatrix}.$$

These messages are exactly equivalent to what would have been sent between $X_i, Y_i$ in the full model using the new transition matrix. The messages between $X_i, X_{i+1}$ in the full model would have been the element-wise product of its received messages and the node potentials. So, for example, the message from $X_2$ to $X_1$ would have been the all zero-vector, because it has received conflicting information from $Y_2$ and $X_3$ about what state $X_2$ should be in.

(f) The new transition matrix has identical rows, so the transition probabilities are the same no matter what the current state is. Hence the $X_i$ variables are no longer dependent on the particular value of $X_{i-1}$, i.e. $X_i$’s are now independent. The new model is drawn in Fig. 1(d). $X_2$ and $X_3$ are now also root nodes, and need to be endowed with their own marginals. Let $p_t = P(X_t)$. We need to compute $p_2$ and $p_3$ from $p_1$ and $A$, using the formula for computing marginals in a Markov chain.

$$P(x_{t+1} = j) = \sum_i P(x_t = i)P(x_{t+1} = j|x_t = i) = \sum_i P(x_t = i) * a_j = a_j,$$

where $a_j$ is the $j$-th element of any row of $A$ (since all rows are the same). Hence the marginals $p_t$ are equal to each other and the rows of $A$.

$$p_2 = p_3 = \begin{bmatrix} 1/2 \\ 1/4 \\ 1/4 \end{bmatrix}.$$

The messages are

$$m_{y_1x_1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad m_{y_2x_2} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad m_{y_3x_3} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

$$m_{xy_1} = \begin{bmatrix} 0.75 \\ 0.25 \\ 0.25 \end{bmatrix}, m_{xy_2} = \begin{bmatrix} 0.75 \\ 0.25 \\ 0.25 \end{bmatrix}, m_{xy_3} = \begin{bmatrix} 0.75 \\ 0.25 \\ 0.25 \end{bmatrix}.$$
The marginals are
\[ p(x_1|\bar{y}_{1,2,3}) = \begin{bmatrix} 0.6667 \\ 0.3333 \\ 0 \end{bmatrix}, \quad p(x_2|\bar{y}_{1,2,3}) = \begin{bmatrix} 0.6667 \\ 3333 \\ 0 \end{bmatrix}, \quad p(x_3|\bar{y}_{1,2,3}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \]

If we were to do message passing on the full model using the new transition matrix, we would discover that the messages passed from \( X_{i+1} \) to \( X_i \) are all scalars (or constant vectors), and the messages from \( X_i \) to \( X_{i+1} \) are multiples of \( \begin{bmatrix} 2 & 1 & 1 \end{bmatrix} \). This makes sense because the \( X_{i+1} \)'s provide no additional information to \( X_i \), but \( X_{i+1} \) has to somehow figure out its correct marginal distribution from \( X_i \).

**Problem 3. (Naive Bayes)**

(a) See Fig. 2 for a drawing of the graphical model. Letting \( X \) denote the vector \( X_1, X_2, \ldots, X_d \), the probability distribution associated with the graph is
\[ P(X, Y) = P(Y) \prod_{i=1}^{d} P(X_i|Y), \]
where \( P(Y) \) is called the prior class probability, and \( P(X_i|Y) \) are the class conditional probabilities.

(b) Using the definition of conditional probability and the joint probability defined above, we get
\begin{align*}
P(Y = 1|x) &= \frac{P(x, Y = 1)}{P(x)} \quad (12) \\
&= \frac{P(Y = 1) \prod_{i=1}^{d} P(x_i|Y = 1)}{P(Y = 1) \prod_{i=1}^{d} P(x_i|Y = 1) + (1 - P(Y = 1)) \prod_{i=1}^{d} P(x_i|Y = 0)} \quad (13)
\end{align*}

(c) When classifying webpages using word occurrences on the page, the naive Bayes model may not be an accurate reflection of reality. Under the naive
Bayes model, words occur on the page independently of each other once we know the class label. However, in real life, knowing what type of webpage we are looking at, and seeing some words on the page, would give us ideas about what other words should also occur. For example, suppose we are looking at new paper articles, then when we see the word "United," then one might also expect to see the word "States." If we are looking at university department webpages, then one would expect the word "Computer" to often co-occur with "Science" or "Engineering," as opposed to "repair." Naive Bayes is "naive" because it makes the simplifying (and obviously wrong) assumption of conditional independence. However, in practice, one finds that in many cases its performance often rivals that of more complicated methods, despite its naive assumptions.

(d) Let $p_0 = P(Y = 1)$, $\sigma_i = P(X_i = 1|Y = 1)$, and $\rho_i = P(X_i = 1|Y = 0)$. We rewrite the class conditionals using a notational shorthand.

$$P(X_i|Y = 1) = \sigma_i^{X_i}(1 - \sigma_i)^{1-X_i}$$
$$P(X_i|Y = 0) = \rho_i^{X_i}(1 - \rho_i)^{1-X_i}$$

Using the monotonicity of the log function and the fact that $\frac{a}{a+b} \geq 1/2$ iff $a \geq b$, we see that

$$P(Y = 1|x) \geq 1/2$$

$$\iff P(x, Y = 1) \geq P(x, Y = 0)$$

$$\iff p_0 \prod_{i=1}^{d} \sigma_i^{X_i}(1 - \sigma_i)^{1-X_i} \geq (1 - p_0) \prod_{i=1}^{d} \rho_i^{X_i}(1 - \rho_i)^{1-X_i}$$

$$\iff \log p_0 + \sum_i X_i \log \sigma_i + (1-X_i) \log (1-\sigma_i) \geq \log (1-p_0) + \sum_i X_i \log \rho_i + (1-X_i) \log (1-\rho_i)$$

$$\iff \sum_i X_i \left( \log \frac{\sigma_i}{1-\sigma_i} - \log \frac{\rho_i}{1-\rho_i} \right) \geq \sum_i \log \frac{1-\rho_i}{1-\sigma_i} - \log \frac{p_0}{1-p_0}. \quad (14)$$

The log $\frac{\sigma_i}{1-\sigma_i}$ terms in the equations above are called log-odds ratios and are important quantities in statistical decision theory.

Finally, we see that $b$ is just the constant terms at the right hand side of the last equation, and $a_i = \log \frac{\sigma_i}{1-\sigma_i} - \log \frac{\rho_i}{1-\rho_i}$.
Figure 2: The directed graphical model corresponding to the Naive Bayes algorithm.
Naive Bayes

(a)

\[ p(Y = 1|x) = \frac{p(Y = 1) \prod_{i=1}^{d} p(x_i|Y = 1)}{p(Y = 0) \prod_{i=1}^{d} p(x_i|Y = 0) + p(Y = 1) \prod_{i=1}^{d} p(x_i|Y = 1)} \]

(b)

Figure 5: Graphical model for Naive Bayes

(c)

This may not be an accurate model of the joint distribution since the naive Bayes model assumes independence between \(X_w\) and thus independence between words. It is a well established fact that written text typically exhibit some Markovian structure and thus non-independence between neighboring words. Additionally the words within the two classes we are trying to distinguish may inherently exhibit non-independence (for example if the web pages correspond to different topic areas).

(d)

\[ p(Y = 1|x) = \frac{p(Y = 1) \prod_{i=1}^{d} p(x_i|Y = 1)}{p(Y = 0) \prod_{i=1}^{d} p(x_i|Y = 0) + p(Y = 1) \prod_{i=1}^{d} p(x_i|Y = 1)} \]

\[ = \frac{1}{1 + \exp \left( \log \frac{p(Y = 0)}{p(Y = 1)} + \sum_{i=1}^{d} \log \frac{p(x_i|Y = 0)}{p(x_i|Y = 1)} \right)} \]

\[ = \frac{1}{1 + \exp \left( -A^T x + b \right)} \]

\[ p(Y = 1|x) \geq \frac{1}{2} \]

\[ \frac{1}{1 + \exp \left( -A^T x + b \right)} \geq \frac{1}{2} \]

\[ 2 \geq 1 + \exp \left( -A^T x + b \right) \]

\[ 1 \geq \exp \left( -A^T x + b \right) \]

\[ 0 \geq -A^T x + b \]

\[ A^T x \geq b \]

\[ \sum_{i=1}^{d} a_i x_i \geq b \]