6.867: Recitation Handout (Week 9)

November 11, 2016

Contents

1 3Flix

(a) 3Flix is a new movie website with only three movies - The Ring, R, Grudge, G, and The Shining, S. Alice, Bob and James are new users on 3Flix, and rate a few movies as shown below. We further know that Alice, Bob and James have similar tastes and the movies belong to the same genre, as a result we would like to estimate the rating matrix by a rank-1 approximation. How would the rating matrix approximation look like?

	R	G	S
A	3	2	*
B	*	2.5	1
J	1.5	*	*

Solution: By definition, $\|A\|_F^2 = \sum_{1\leqslant i,j\leqslant n} a_{ij}^2$. Now, since we use a rank-1 approximation, we have $u, v \in \mathbb{R}^{n\times 1}$. Let ranking matrix be approximated by uv^T, where $(i,j) \in M$ are the observed ratings. Since we want to minimize,

$$\min_{u,v} \|P_M(Y) - P_M(uv^T)\|_F^2 \tag{1}$$

Here $[P_M(X)]_{ij} = X_{ij}$ if $(i,j) \in M$ else $[P_M(X)]_{ij} = 0$, where M is the set of observed (i,j) pairs for user i and movie j. We observe that Eqn. (1) is $\min_{u,v} \sum_{ij\in M}(Y_{ij} - uv_{ij}^T)^2 \implies Y_{ij} = uv_{ij}^T \ \forall (i,j) \in M$. This gives $u_1v_1 = 3, u_1v_2 = 2, u_2v_2 = 2.5, u_2v_3 = 1, u_3v_1 = 1.5$. From these, we can obtain $u_1v_3 = 0.8, u_2v_1 = 3.75, u_3v_2 = 1, u_3v_3 = 0.4$

	R	G	S
A	3	2	0.8
B	3.75	2.5	1
J	1.5	1	0.4

(b) Due to a database wipe, all of James' preferences are now gone. How does the new rating matrix look like now?

	R	G	S
A	3	2	*
B	*	2.5	1
J	*	*	*

Solution: We can't make predictions about James, the other entries look the same.

	R	G	S
A	3	2	0.8
B	3.75	2.5	1
J	*	*	*

2 Big Billion Day

The e-shopping website, EZBuy, has an incomplete user preference matrix for more than 3 million users. It is planning to host a Big Billion Day, 80% off on all products, and offer best matches to its customers. A brute force gradient descent is not possible for Eqn.(2)

$$\min_{U,V} ||P_M(Y) - P_M(UV^T)||_F^2 \tag{2}$$

Here $[P_M(X)]_{ij} = X_{ij}$ if $(i,j) \in M$ else $[P_M(X)]_{ij} = 0$, where M is the set of observed (i,j) pairs for user i and product j. Is there an SGD version for this?

Solution: By definition, $||A||_F^2 = \sum_{1 \leqslant i,j \leqslant n} a_{ij}^2$.
Then, $\min_{U,V} ||P_M(Y) - P_M(UV^T)||_F^2 = \sum_{(i,j) \in M} (Y_{ij} - [UV^T]_{ij})^2$.
Then if we think of M as the dataset D, then due to separability of the loss function the SGD version is easy to derive.

3 Factorization Machines

Factorization machines for an input $\mathbf{x}$, predict output $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \sum_{i=1}^{d} \sum_{j=i+1}^{d} [VV^T]_{ij} x_i x_j$. Here $V \in \mathbb{R}^{d \times k}$, and $\mathbf{w} \in \mathbb{R}^{d \times 1}$, that need to be learned and $k << d$ is a rank hyperparameter. $\mathbf{x}$ is the feature vector, for example: for a record (u,i,r), where $u \in U$, $i \in I$, here U, I are user set and item set respectively, and r is the ranking. Then, $\mathbf{x} \in \mathbb{R}^{|U|+|I|}$ and $x_u, x_i = 1$, and $x_j = 0$ otherwise for the observed user-item pair (u,i). Taken from [1]

(a) What is feature vector for user James and movie The Ring as in the problem 3Flix?

Solution: $x = (x_A, x_B, x_J, x_R, x_G, x_S)$, where $x_k \in \{0,1\}$, then the feature vector looks like $x = (0,0,1,1,0,0)$

(b) Pick

$$V = \begin{bmatrix} A \\ B \end{bmatrix} \in \mathbb{R}^{d \times k} \tag{3}$$

$$w = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{d \times 1} \tag{4}$$

where $A \in \mathbb{R}^{|U| \times k}$, and $a \in \mathbb{R}^{|U| \times 1}$.
Plug this in the Factorization Machine and do you see any similarity to Problem 3Flix?

[1] `http://www.mblondel.org/publications/mblondel-ecmlpkdd2015.pdf`

Solution: Since $A \in \mathbb{R}^{|U| \times k}$, and $a \in \mathbb{R}^{|U| \times 1}$.

$$VV^T = \begin{bmatrix} AA^T & AB^T \\ BA^T & BB^T \end{bmatrix} \tag{5}$$

Then $[VV^T]_{ij}$ for $j > i$, as in the definition, would use only the upper block, AB^T, of VV^T. As a result $y_{ui} = [AB^T]_{ui} + a_u + b_i$, which looks like the matrix factorization problem.

4 Holmes and Watson in LA

Holmes and Watson have moved to LA. Holmes wakes up to find that his lawn is wet. He wonders if it has rained or if he left his sprinkler on. He looks at his neighbor Watson's lawn and sees that it is wet, too. So, he concludes it must have rained.

Use the binary random variables R for rain, S for sprinkler, H for Holmes' lawn being wet and W for Watson's lawn being wet. Assume you are given the following probability distributions:

$$\begin{aligned} P(R = 1) &= 0.2 \\ P(S = 1) &= 0.1 \\ P(W = 1 \mid R = 0) &= 0.2 \\ P(W = 1 \mid R = 1) &= 1.0 \\ P(H = 1 \mid R = 0, S = 0) &= 0.1 \\ P(H = 1 \mid R = 0, S = 1) &= 0.9 \\ P(H = 1 \mid R = 1, S = 0) &= 1.0 \\ P(H = 1 \mid R = 1, S = 1) &= 1.0 \end{aligned}$$

(a) Draw the corresponding directed graphical model?

(b) What is $P(H)$?

(c) What is $P(R \mid H)$?

(d) What is $P(S \mid H)$?

(e) What is $P(W \mid H)$?

(f) What is $P(R \mid W, H)$?

(g) What is $P(S \mid W, H)$?

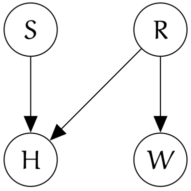
(h) What probability expression corresponds to Holmes' belief that it rained before he goes out? What is its value?

(i) What probability expression corresponds to Holmes' belief that it rained after he sees that his lawn is wet? What is its value?

(j) What probability expression corresponds to Holmes' belief that it rained after he sees that his lawn is wet and that Watson's is wet as well? What is its value?

(k) What probability expression corresponds to Holmes' belief that the sprinkler was on after he sees that his lawn is wet and that Watson's is wet as well? What is its value?

Solution:

(a) 

(b) We have $\mathrm{P} = \mathrm{P}(\mathrm{R}, \mathrm{S}, W, \mathrm{H}) = \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R})\mathrm{P}(\mathrm{H}|\mathrm{R}, \mathrm{S})$, thus

$$\begin{aligned}
\mathrm{P}(\mathrm{H}=1) &= \sum_{\mathrm{R},\mathrm{S},W} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R})\mathrm{P}(\mathrm{H}=1|\mathrm{R},\mathrm{S}) && (6)\\
&= 0.344 && (7)
\end{aligned}$$

(c)

$$\begin{aligned}
\mathrm{P}(\mathrm{R}=1|\mathrm{H}=1) &= \frac{\sum_{\mathrm{S},W} \mathrm{P}(\mathrm{R}=1)\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R}=1)\mathrm{P}(\mathrm{H}=1|\mathrm{R}=1,\mathrm{S})}{\sum_{\mathrm{R},\mathrm{S},W} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R})\mathrm{P}(\mathrm{H}=1|\mathrm{R},\mathrm{S})} && (8)\\
&= 0.581 && (9)\\
\mathrm{P}(\mathrm{R}=1|\mathrm{H}=0) &= \frac{\sum_{\mathrm{S},W} \mathrm{P}(\mathrm{R}=1)\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R}=1)\mathrm{P}(\mathrm{H}=0|\mathrm{R}=1,\mathrm{S})}{\sum_{\mathrm{R},\mathrm{S},W} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R})\mathrm{P}(\mathrm{H}=0|\mathrm{R},\mathrm{S})} && (10)\\
&= 0 && (11)
\end{aligned}$$

(d)

$$\begin{aligned}
\mathrm{P}(\mathrm{S}=1|\mathrm{H}=1) &= \frac{\sum_{\mathrm{R}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S}=1)\mathrm{P}(\mathrm{H}=1|\mathrm{R},\mathrm{S}=1)}{\sum_{\mathrm{S},\mathrm{R}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(\mathrm{H}=1|\mathrm{R},\mathrm{S})} && (12)\\
&= 0.267 && (13)\\
\mathrm{P}(\mathrm{S}=1|\mathrm{H}=0) &= \frac{\sum_{\mathrm{R}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S}=1)\mathrm{P}(\mathrm{H}=0|\mathrm{R},\mathrm{S}=1)}{\sum_{\mathrm{S},\mathrm{R}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(\mathrm{H}=0|\mathrm{R},\mathrm{S})} && (14)\\
&= 0.012 && (15)
\end{aligned}$$

(e)

$$\begin{aligned}
\mathrm{P}(W=1|\mathrm{H}=1) &= \frac{\sum_{\mathrm{R},\mathrm{S}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W=1|\mathrm{R})\mathrm{P}(\mathrm{H}=1|\mathrm{R},\mathrm{S})}{\sum_{W,\mathrm{R},\mathrm{S}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R})\mathrm{P}(\mathrm{H}=1|\mathrm{R},\mathrm{S})} && (16)\\
&= 0.665 && (17)\\
\mathrm{P}(W=1|\mathrm{H}=0) &= \frac{\sum_{\mathrm{R},\mathrm{S}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W=1|\mathrm{R})\mathrm{P}(\mathrm{H}=0|\mathrm{R},\mathrm{S})}{\sum_{W,\mathrm{R},\mathrm{S}} \mathrm{P}(\mathrm{R})\mathrm{P}(\mathrm{S})\mathrm{P}(W|\mathrm{R})\mathrm{P}(\mathrm{H}=0|\mathrm{R},\mathrm{S})} && (18)\\
&= 0.200 && (19)
\end{aligned}$$

(f)

$$P(R=1|W=1,H=1)=0.874 \tag{20}$$
$$P(R=1|W=1,H=0)=0 \tag{21}$$
$$P(R=1|W=0,H=1)=0 \tag{22}$$
$$P(R=1|W=0,H=0)=0 \tag{23}$$

(g)

$$P(S=1|W=1,H=1)=0.150 \tag{24}$$
$$P(S=1|W=1,H=0)=0.012 \tag{25}$$
$$P(S=1|W=0,H=1)=0.500 \tag{26}$$
$$P(S=1|W=0,H=0)=0.012 \tag{27}$$

(h)

$$P(R=1)=0.2 \tag{28}$$

(i)

$$\begin{aligned} P(R=1|H=1) &= \frac{\sum_{S,W} P(R=1)P(S)P(W|R=1)P(H=1|R=1,S)}{\sum_{R,S,W} P(R)P(S)P(W|R)P(H=1|R,S)} && (29)\\ &= \frac{0.2}{0.2+0.144} = 0.581 && (30)\end{aligned}$$

(j)

$$\begin{aligned} P(R=1|H=1,W=1) &= \frac{\sum_{S} P(R=1)P(S)P(W=1|R=1)P(H=1|R=1,S)}{\sum_{R,S} P(R)P(S)P(W=1|R)P(H=1|R,S)} && (31)\\ &= \frac{0.2}{0.2+0.0288} = 0.874 && (32)\end{aligned}$$

(k)

$$\begin{aligned} P(S=1|H=1,W=1) &= \frac{\sum_{R} P(R)P(S=1)P(W=1|R)P(H=1|R,S=1)}{\sum_{R,S} P(R)P(S)P(W=1|R)P(H=1|R,S)} && (33)\\ &= \frac{0.0344}{0.0344+0.1944} = 0.153 && (34)\end{aligned}$$

5 Turning the tables (Bishop 8.3)

(a)

a	b	c	$p(a,b,c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Consider three binary variables $a, b, c \in \{0,1\}$ having the joint distribution given in the table above. Show by direct evaluation that this distribution has the property that a and b are marginally dependent, so that $p(a,b) \neq p(a)p(b)$, but they become independent when conditioned on c, so that $p(a,b|c) = p(a|c)p(b|c)$ for both $c = 0$ and $c = 1$.

Solution:

The distribution $p(a,b)$ is found by summing the complete joint distribution $p(a,b,c)$ over the states of c so that

$$p(a,b) = \sum_{c\in\{0,1\}} p(a,b,c)$$

and similarly the marginal distributions $p(a)$ and $p(b)$ are given by

$$p(a) = \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} p(a,b,c) \text{ and } p(b) = \sum_{a\in\{0,1\}} \sum_{c\in\{0,1\}} p(a,b,c) \quad (1)$$

Table 2 shows the joint distribution $p(a,b)$ as well as the product of the marginals $p(a)p(b)$, demonstrating that these are not equal for the specified distribution.

The conditional distribution $p(a,b|c)$ is obtained by conditioning on the value of c and normalizing

$$p(a,b|c) = \frac{p(a,b,c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a,b,c)}$$

Similarly for the conditionals $p(a|c)$ and $p(b|c)$ we have

$$p(a|c) = \frac{\sum_{b\in\{0,1\}} p(a,b,c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a,b,c)}$$

and

$$p(b|c) = \frac{\sum_{a\in\{0,1\}} p(a,b,c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a,b,c)} \quad (2)$$

Table 3 compares the conditional distribution $p(a,b|c)$ with the product of the marginals $p(a|c)p(b|c)$, showing that these are equal for the given joint distribution $p(a,b,c)$ for both $c = 0$ and $c = 1$. Note that $p(c = 0) = 0.48$ and $p(c = 1) = 0.52$.

a	b	$p(a,b)$
0	0	0.336
0	1	0.264
1	0	0.256
1	1	0.144
a	b	$p(a)p(b)$
0	0	0.3552
0	1	0.2448
1	0	0.2368
1	1	0.1632

Table 1: Comparison of the conditional distribution $p(a, b|c)$ with the product of the marginals $p(a|c)p(b|c)$ showing that these are equal for the given joint distribution.

a	b	c	$p(a,b|c)$
0	0	0	0.400
0	1	0	0.100
1	0	0	0.400
1	1	0	0.100
0	0	1	0.277
0	1	1	0.415
1	0	1	0.123
1	1	1	0.185

a	b	c	$p(a|c)p(b|c)$
0	0	0	0.400
0	1	0	0.100
1	0	0	0.400
1	1	0	0.100
0	0	1	0.277
0	1	1	0.415
1	0	1	0.123
1	1	1	0.185

Table 2: Comparison of the distribution $p(a, b)$ with the product of the marginals $p(a)p(b)$ showing these are not equal for the given joint distribution $p(a, b, c)$.

a	$p(a)$
0	0.600
1	0.400

a	c	$p(c\|a)$
0	0	0.400
1	0	0.600
0	1	0.600
1	1	0.400

b	c	$p(b\|c)$
0	0	0.800
1	0	0.200
0	1	0.400
1	1	0.600

Table 3: Tables of $p(a)$, $p(c|a)$ and $p(b|c)$ evaluated by marginalizing and conditioning the joint distribution

(b) Evaluate the distributions $p(a)$, $p(b|c)$ and $p(c|a)$ corresponding to the joint distribution given above and show by direct evaluation that $p(a, b, c) = p(a)p(c|a)p(b|c)$.

Solution:

In the previous exercise we already computed $p(a)$ in (1) and $p(b|c)$ in (2). There remains to compute $p(c|a)$ which is done using

$$p(c|a) = \frac{\sum_{b\in\{0,1\}} p(a, b, c)}{\sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} p(a, b, c)}$$

The required distributions are given in Table 3.

Multiplying the three distributions together we recover the joint distribution $p(a, b, c)$ given in Table a, thereby allowing us to verify the validity of the decomposition $p(a, b, c) = p(a)p(c|a)p(b|c)$ for this particular joint distribution. We can express this decomposition using the graph shown in Figure 1.

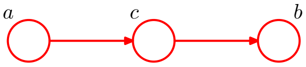


Figure 1: Directed graph representing the joint distribution given in Table a.

6 Out of gas? (Bishop 8.11)

Consider the example of the car fuel system shown in Figure 3. Given:

$$\begin{aligned} p(B = 1) &= 0.9 \\ p(F = 1) &= 0.9. \\ p(G = 1|B = 1, F = 1) &= 0.8 \\ p(G = 1|B = 1, F = 0) &= 0.2 \\ p(G = 1|B = 0, F = 1) &= 0.2 \\ p(G = 1|B = 0, F = 0) &= 0.1 \end{aligned}$$

Suppose that instead of observing the state of the fuel gauge G directly, the gauge is seen by the driver D who reports to us the reading on the gauge. This report is either that the gauge shows full $D = 1$ or that it shows empty $D = 0$. Our driver is a bit unreliable, as expressed through the following probabilities

$$p(D = 1|G = 1) = 0.9 \quad (8.105)$$

$$p(D = 0|G = 0) = 0.9. \quad (8.106)$$

Suppose that the driver tells us that the fuel gauge shows empty, in other words that we observe $D = 0$. Evaluate the probability that the tank is empty given only this observation. Similarly, evaluate the corresponding probability given also the observation that the battery is flat, and note that this second probability is lower. Discuss the intuition behind this result, and relate the result to Figure 2.

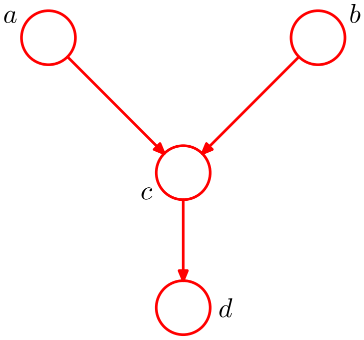


Figure 2: Bayesian Net

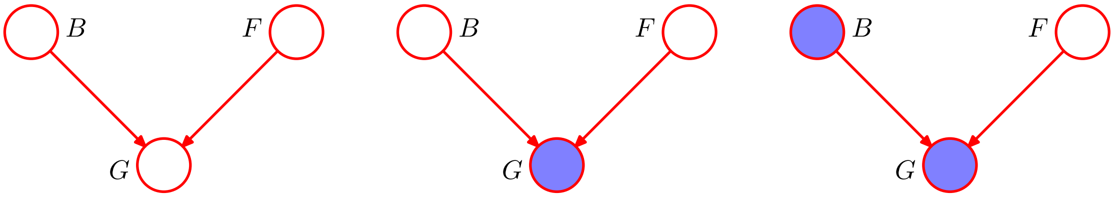


Figure 3: An example of a 3-node graph used to illustrate the phenomenon of 'explaining away'. The three nodes represent the state of the battery (B), the state of the fuel tank (F) and the reading on the electric fuel gauge (G). See the text for details.

Solution:

The described situation correspond to the graph shown in Figure 2 with $a = B$, $b = F$, $c = G$ and $d = D$ (cf. Figure 3). To evaluate the probability that the tank is empty given the driver's report that the gauge reads zero, we use Bayes' theorem

$$p(F = 0|D = 0) = \frac{p(D = 0|F = 0)p(F = 0)}{p(D = 0)}$$

To evaluate $p(D = 0|F = 0)$, we marginalize over B and G,

$$p(D = 0|F = 0) = \sum_{B,G} p(D = 0|G)p(G|B, F = 0)p(B) = 0.748 \quad (3)$$

and to evaluate $p(D = 0)$, we marginalize also over F ,

$$p(D = 0) = \sum_{B,G,F} p(D = 0|G)p(G|B, F)p(B)p(F) = 0.352 \quad (4)$$

Combining these results with $p(F = 0)$, we get

$$p(F = 0|D = 0) = 0.213$$

Note that this is slightly lower than the probability obtained in (8.32) for $p(F = 0|G = 0)$, reflecting the fact that the driver is not completely reliable. If we now also observe $B = 0$, we no longer marginalize over B in (3) and (4), but instead keep it fixed at its observed value, yielding

$$p(F = 0|D = 0, B = 0) = 0.110$$

which is again lower than what we obtained with a direct observation of the fuel gauge in (8.33) for $p(F = 0|G = 0, B = 0)$. More importantly, in both cases the value is lower than before we observed $B = 0$, since this observation provides an alternative explanation why the gauge should read zero; see also discussion following (8.33).

7 No Independence (Bishop 8.27)

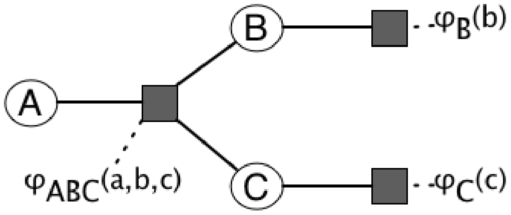
Consider two discrete variables x and y each having three possible states, for example $x, y \in \{0, 1, 2\}$. Construct a joint distribution $p(x, y)$ over these variables having the property that the value $\hat{x}$ that maximizes the marginal $p(x)$, along with the value $\hat{y}$ that maximizes the marginal $p(y)$, together have probability zero under the joint distribution, so that $p(\hat{x}, \hat{y}) = 0$.

Solution: An example is given by

	x=0	x=1	x=2
y=0	0.0	0.1	0.2
y=1	0.0	0.1	0.2
y=2	0.3	0.1	0.0

for which $\hat{x} = 2$ and $\hat{y} = 2$

8 Some product!



Consider the simple factor graph above. The factors are defined as follows:

a	b	c	$\phi_{ABC}(a,b,c)$
0	0	0	10
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	10

b	$\phi_B(b)$
0	2
1	5

c	$\phi_C(c)$
0	1
1	10

Compute $\Pr(A)$ using sum-product.

Solution: In this simple case, we only have to compute one message: $\mu_{\phi_{ABC}\to A}(a)$.

$$\begin{aligned}
\mu_{\phi_{ABC}\to A}(a) &= \sum_{b,c} \phi_{ABC}(a,b,c) \prod_{X_M \in \{B,C\}} \mu_{X_M \to \phi_{ABC}(X_M)} \\
&= \sum_{b,c} \phi_{ABC}(a,b,c) \mu_{B\to\phi_{ABC}}(b) \mu_{C\to\phi_{ABC}}(c) \\
&= \sum_{b,c} \phi_{ABC}(a,b,c) \mu_{\phi_B\to B}(b) \mu_{\phi_C\to C}(c) \\
&= \sum_{b,c} \phi_{ABC}(a,b,c) \phi_B(b) \phi_C(c)
\end{aligned}$$

a	b	c	$\phi_{ABC}(a,b,c)$	$\phi_B(b)$	$\phi_C(c)$	$\phi_{ABC}(a,b,c)\phi_B(b)\phi_C(c)$
0	0	0	10	2	1	20
0	0	1	1	2	10	20
0	1	0	1	5	1	5
0	1	1	1	5	10	50
1	0	0	1	2	1	2
1	0	1	1	2	10	20
1	1	0	1	5	1	5
1	1	1	10	5	10	500

Summing out b and c, we get:

a	$\mu_{\phi_{ABC}\to A}(a)$
0	95
1	527

Normalizing gives us the desired distribution.

a	$\Pr(a)$
0	0.153
1	0.847

9 Exact inference

(a) Draw the factor graph that represents the following probability density:

$$\Pr(X_1, \ldots, X_5) = \frac{1}{z}\phi_{123}(X_1, X_2, X_3)\phi_{345}(X_3, X_4, X_5)$$

where the factors are specified by

X_1	X_2	X_3	$\phi_{123}(X_1, X_2, X_3)$
0	0	0	10
0	0	1	6
0	1	0	1
0	1	1	2
1	0	0	1
1	0	1	2
1	1	0	5
1	1	1	11

X_3	X_4	X_5	$\phi_{345}(X_3, X_4, X_5)$
0	0	0	10
0	0	1	6
0	1	0	1
0	1	1	2
1	0	0	1
1	0	1	2
1	1	0	5
1	1	1	11

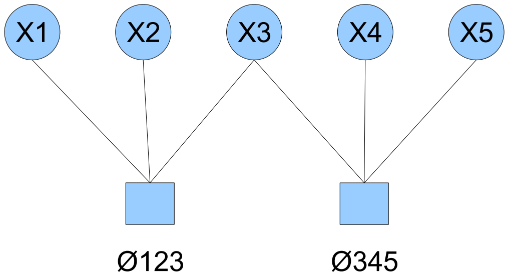
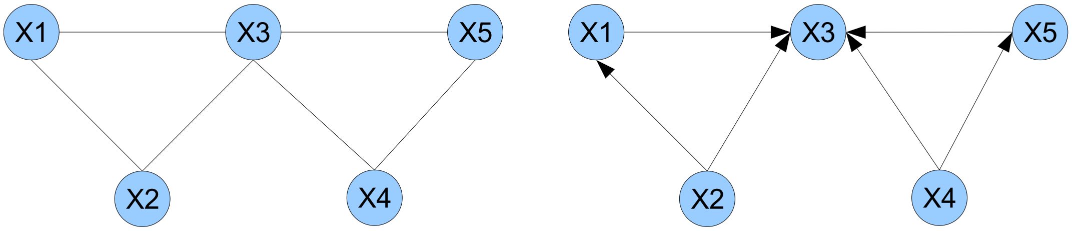


Figure 4: Part (a)

(b) Draw an MRF that could be represented with the same set of factors (factors over the same variables) as those specified above. Then, consider a Bayesian network that could be represented with this factor structure. Hint: What BN involving three nodes is needed to represent all the distributions representable with each of the factors in this example? What would the conditional probability tables (CPTs) in the Bayesian network (directed graphical model) be?



(a) MRF (b) BN

Figure 5: Part (b)

Solution: In an MRF, the factors apply to cliques, so we need two three-node cliques that share a node. There are many possible Bayesian networks for the given factor graph. The crucial point is to recognize that for the given factors, $\{X_1, X_2, X_3\}$ form a clique when converted to an undirected graph, and likewise for $\{X_3, X_4, X_5\}$ (the directionality is not particularly important). The example figure above shows a directed clique, which can represent any factor over $\{X_1, X_2, X_3\}$. Depending on the factor, it would also be acceptable to have a v-structure such as $X_1 \rightarrow X_3 \leftarrow X_2$, since when converted to an undirected graph this would also be a clique. These are the minimum necessary edges in the graph to represent the dependencies present in the factors; we have to be able to represent a general joint distribution among the three variables in the factor. Depending on the network structure, the conditional probability tables can be found by normalizing the appropriate values for the factors. For example, $\Pr(X_2 = 0|X_1 = 0) = (10+6)/(10+6+1+2) = 16/19, \Pr(X_2 = 1|X_1 = 0) = 3/19$, etc.

(c) Compute the distribution $\Pr(X_3)$ by multiplying the factors and summing out unwanted variables (take advantage of conditional indepependences to reduce work).

Solution: To compute the distribution of X_3, we marginalize out all other variables in the joint:

$$\begin{aligned}
\Pr(X_3) = \sum_{X_1,X_2,X_4,X_5} \Pr(X_1,\ldots,X_5) &= \sum_{X_1,X_2,X_4,X_5} \left[\frac{1}{z}\phi_{123}(X_1,X_2,X_3)\cdot\phi_{345}(X_3,X_4,X_5)\right] \\
&= \frac{1}{z}\left[\sum_{X_1,X_2}\phi_{123}(X_1,X_2,X_3)\right]\left[\sum_{X_4,X_5}\phi_{345}(X_3,X_4,X_5)\right]
\end{aligned}$$

Specifically, for $X_3 = 0$:

$$\begin{aligned}
\Pr(X_3 = 0) &= \frac{1}{z}\left[\phi_{123}(0,0,0) + \cdots + \phi_{123}(1,1,0)\right]\left[\phi_{345}(0,0,0) + \cdots + \phi_{345}(0,1,1)\right] \\
&= \frac{1}{z}\left[10+1+1+5\right]\left[10+6+1+2\right] = \frac{1}{z}(17\cdot 19) = \frac{1}{z}(323)
\end{aligned}$$

Similarly, for $X_3 = 1$:

$$\Pr(X_3 = 1) = \frac{1}{z}\left[6+2+2+11\right]\left[1+2+5+11\right] = \frac{1}{z}(21\cdot 19) = \frac{1}{z}(399)$$

Normalizing these two values, $\Pr(X_3 = 0) = 17/38, \Pr(X_3 = 1) = 21/38$.

(d) Compute the distribution $\Pr(X_3)$ using the sum-product message-passing algorithm. You may do it by hand or with a computer program.

Solution: In our factor graph, the nodes X_1, X_2, X_4, X_5 are leaves. Their messages are initialized to $\mathbf{1}$, the all-ones vector (in this case, the vector is of length two, corresponding to the 0 and 1 variable assignments):

$$\mu_{X_1\to\phi_{123}}(X_1) = \mathbf{1}, \quad \mu_{X_2\to\phi_{123}}(X_2) = \mathbf{1}, \quad \mu_{X_4\to\phi_{345}}(X_4) = \mathbf{1}, \quad \mu_{X_5\to\phi_{345}}(X_5) = \mathbf{1}$$

Both factor nodes now have all incoming messages except that of X_3, hence the next set of messages to be passed is from each factor to X_3. We demonstrate the computation for $\mu_{\phi_{123}\to X_3}(X_3)$ in detail:

$$\begin{aligned}\mu_{\phi_{123}\to X_3}(X_3) &= \sum_{X_1,X_2} \phi_{123}(X_1,X_2,X_3)\cdot\mu_{X_1\to\phi_{123}}(X_1)\cdot\mu_{X_2\to\phi_{123}}(X_2)\\ \mu_{\phi_{123}\to X_3}(X_3=0) &= \left[\phi_{123}(0,0,0)\cdot\mu_{X_1\to\phi_{123}}(0)\cdot\mu_{X_2\to\phi_{123}}(0)\right]+\\ &\quad\cdots+\left[\phi_{123}(1,1,0)\cdot\mu_{X_1\to\phi_{123}}(1)\cdot\mu_{X_2\to\phi_{123}}(1)\right]\\ &= (10\cdot1\cdot1)+(1\cdot1\cdot1)+(1\cdot1\cdot1)+(5\cdot1\cdot1)=17\\ \mu_{\phi_{123}\to X_3}(X_3=1) &= (6\cdot1\cdot1)+(2\cdot1\cdot1)+(2\cdot1\cdot1)+(11\cdot1\cdot1)=21\end{aligned}$$

A similar computation yields $\mu_{\phi_{345}\to X_3}(X_3) = [19,19]^T$ for the other factor's message. Finally, to obtain the marginal, we first multiply all incoming messages to X_3, giving $[17\cdot19, 21\cdot19]^T = [323,399]^T$, and normalizing this result gives the actual X_3 marginal $[17/38, 21/38]^T$, which is the same as question 3.

(e) Compute the distribution $\Pr(X_3|X_1=0,X_4=1)$, by direct computation (products of factors and summing out) and by message passing.

Solution: For direct computation, we perform a similar computation as question 3, except we can exclude some terms in the sums because X_1 and X_4 values are given and do not need to be marginalized.

$$\begin{aligned}\Pr(X_3=0|X_1=0,X_4=1) &= \frac{1}{z}\left[\phi_{123}(0,0,0)+\phi_{123}(0,1,0)\right]\left[\phi_{345}(0,1,0)+\phi_{345}(0,1,1)\right]\\ &= \frac{1}{z}\left[10+1\right]\left[1+2\right]=\frac{1}{z}(11\cdot3)=\frac{1}{z}(33)\\ \Pr(X_3=1|X_1=0,X_4=1) &= \frac{1}{z}\left[6+2\right]\left[5+11\right]=\frac{1}{z}(8\cdot16)=\frac{1}{z}(128)\end{aligned}$$

Normalizing these values gives an X_3 marginal of $[33/161, 128/161]^T$.

For message passing, the evidence can be modeled as extra factors ϕ_1, ϕ_4 where they have a value of 1 for their observed values and 0 for everything else. This eventually leads to the following messages:

$$\mu_{X_1\to\phi_{123}}(X_1)=[1,0]^T,\quad \mu_{X_2\to\phi_{123}}(X_2)=\mathbf{1},\quad \mu_{X_4\to\phi_{345}}(X_4)=[0,1]^T,\quad \mu_{X_5\to\phi_{345}}(X_5)=\mathbf{1}$$

When computing the messages from the factors to X_3, the zero values in the X_1 and X_4 messages will make some terms in the summation zero (like in the direct computation case above) because the product will be zero. (Compare the following equations with the ones in question 4.)

$$\begin{aligned}\mu_{\phi_{123}\to X_3}(X_3=0) &= (10\cdot1\cdot1)+(1\cdot1\cdot1)+(1\cdot0\cdot1)+(5\cdot0\cdot1)=11\\ \mu_{\phi_{123}\to X_3}(X_3=1) &= (6\cdot1\cdot1)+(2\cdot1\cdot1)+(2\cdot0\cdot1)+(11\cdot0\cdot1)=8\end{aligned}$$

A similar computation gives $[3, 16]^T$ for the other factor's message. Taking the product of these two messages and normalizing them gives the same X_3 marginal as the direct computation case.

(f) Draw the factor graph that represents the following probability density:

$$\Pr(X_1, \ldots, X_5) = \frac{1}{z}\phi_{123}(X_1, X_2, X_3)\phi_{345}(X_3, X_4, X_5)\phi_{14}(X_1, X_4)$$

Where ϕ_{123} and ϕ_{345} are as before, and ϕ_{14} is

X_1	X_4	$\phi_{14}(X_1, X_4)$
0	0	1
0	1	10
1	0	10
1	1	0

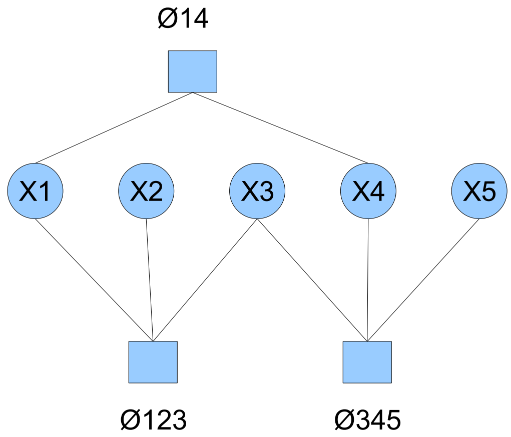


Figure 6: Part (f)

(g) Compute the distribution $\Pr(X_3)$ by direct computation.

Solution: Normalizing the values below, we get $\Pr(X_3 = 0) = 1466/3160, \Pr(X_3 = 1) = 1694/3160$.

$$\Pr(X_3) = \sum_{X_1, X_2, X_4, X_5}\left[\frac{1}{z}\phi_{123}\cdot\phi_{345}\cdot\phi_{14}\right] = \frac{1}{z}\sum_{X_1, X_4}\left[\sum_{X_2}\phi_{123}\right]\left[\sum_{X_5}\phi_{345}\right]\phi_{14}$$

$$\begin{aligned}\Pr(X_3 = 0) &\propto [10+1]\,[10+6]\cdot 1 + [10+1]\,[1+2]\cdot 10 + [1+5]\,[10+6]\cdot 10 + [1+5]\,[1+2]\cdot 0\\ &= (11\cdot 16\cdot 1) + (11\cdot 3\cdot 10) + (6\cdot 16\cdot 10) + (6\cdot 3\cdot 0) = 176 + 330 + 960 + 0 = 1466\\ \Pr(X_3 = 1) &\propto [6+2]\,[1+2]\cdot 1 + [6+2]\,[5+11]\cdot 10 + [2+11]\,[1+2]\cdot 10 + [2+11]\,[5+11]\cdot 0\\ &= (8\cdot 3\cdot 1) + (8\cdot 16\cdot 10) + (13\cdot 3\cdot 10) + (13\cdot 16\cdot 0) = 24 + 1280 + 390 + 0 = 1694\end{aligned}$$

(h) Compute the distribution $\Pr(X_3)$ using the sum-product message-passing algorithm, starting from an initialization of all the messages with ones. Simulate two iterations of the algorithm. Are we guaranteed to get the answer computed by direct computation?

Solution: Unlike the case in trees, there is no well-defined ordering for passing messages in the general loopy case. There has been work on characterizing the conditions and methods under which it can work; search for "loopy belief propagation" for more. For this problem, one possible strategy was:

- Initialize all messages to arbitrary non-zero values (e.g., 1)
- Repeat until convergence (of the X_3 marginal):

  Pick a random node X (to send the message from)

  Send the appropriate message from X to a randomly chosen neighbor of X

Our implementation of the sum-product message-passing algorithm converged to an X_3 marginal of $[0.4741, 0.5259]^T$ from multiple different initializations. Convergence took about 50 iterations (one iteration is an update of all messages at once), although it only took at most a few iterations to reach roughly the correct numbers. Comparing with the answer computed by direct computation, which is $[0.4639, 0.5361]^T$, we see that the message-passing algorithm converges to a close but incorrect answer. Note that, especially if you did this by hand, we do not expect you to have reached convergence by hand calculation. This exercise was to allow you to understand the general message passing algorithm.

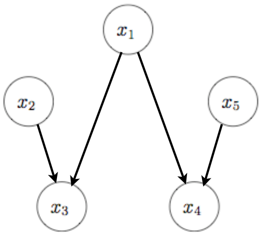
10 Easy exam?

We administered a short exam for $n = 60$ students. There were only five exam questions and all of them were simple true/false questions. We were interested in finding out how the answers might depend on each other. To this end, we collected all the answers into a dataset $D = \{(x_{t1}, ..., x_{t5}), t = 1, \ldots, n\}$, where x_i^t is student t's answer (T/F) to question i. From this data we were able to estimate a Bayesian network, graph G and the associated distribution, over the five variables x_1, . . . , x_5 (answers to questions).

So, of course we misplaced the graph. But we do remember a few relevant properties that may help reconstruct the graph. In particular,

- x_1, x_2, and x_5 were all marginally independent of each other,
- knowing the answer to x_1 made x_2 independent of x_4 and x_3 independent of x_5 (regardless of any other observations).

(a) Draw a Bayesian network that you can infer from the above constraints. Draw the edges in the figure below.



Solution: The arrangement $x_2 \to x_3 \leftarrow x_1$, gives us that x_1 is conditionally independent of x_2 given no observations, that is, x_1 and x_2 are marginally independent. Similarly for (x_1, x_5) and (x_2, x_5). Note also that given x_1, the path between x_2 and x_4 is blocked, even if we know x_3. x_1 also blocks the path between x_3 and x_5, even if we know x_4.

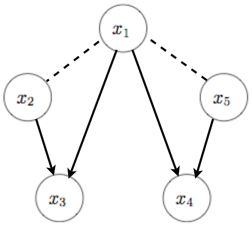
(b) What can you say about the form of the distribution over the five variables?

Solution: Based on the graph we now the distribution has to factor according to

$$\Pr(x_1)\Pr(x_2)\Pr(x_5)\Pr(x_3 \mid x_1, x_2)\Pr(x_4 \mid x_1, x_5)$$

(c) Consider only students who answered x_3 = T and x_4 = T. If we looked at their answers to x_2 and x_5, would we expect these answers to be independent of each other? Briefly justify your answer.

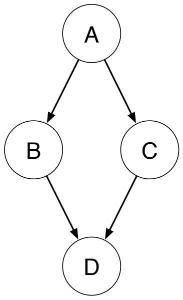
Solution: The answers would be dependent. We could derive this formally by asking whether x_2 is independent of x_5 given x_3 and x_4. The moralized ancestral graph is:



Alternatively, we can simply note that if we know x_3, x_1 and x_2 are dependent. Similarly, if we know x_4, x_1 and x_5 become dependent.

11 The Deciding Factor

Consider the following directed graphical model:



Assume the variables are all binary and the CPTs are as follows:

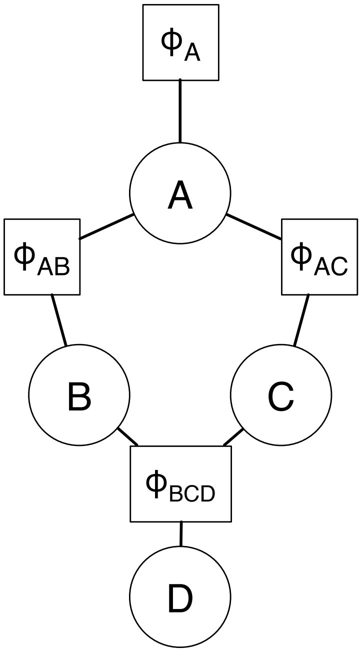
A	P(B = 1)
0	0.3
1	0.6

A	P(C = 1)
0	0.9
1	0.2

B	C	P(D = 1)
0	0	0.9
0	1	0.1
1	0	0.1
1	1	0.9

P(A = 1)
0.3

(a) Draw its associated factor graph and specify the factors in terms of the CPTs given above.



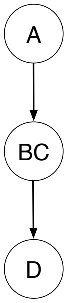
Solution:

The factors are just the CPTs.

(b) Is belief propagation appropriate for exact inference on this model?

Solution: No.

(c) Jody suggests converting the original directed graph to the following one, where BC is a random variable that can take on four possible values: 00, 01, 10, 11 which correspond to joint assignments to variables B and C from the original model.

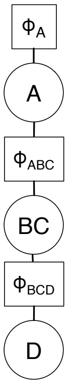


How does it compare in expressive power to the original one?
√ **More** ○ Less ○ Same Briefly explain your answer.

Solution: B and C do not have to be conditionally independent given A any more.

(d) Draw its associated factor graph and provide tables for any factors that differ from part a.

Solution:



Factor ϕ_{ABC} is

A	B	C	
0	0	0	.07
0	0	1	.63
0	1	0	.03
0	1	1	.27
1	0	0	.32
1	0	1	.08
1	1	0	.48
1	1	1	.12

(e) Show how to use belief propagation on this graph to compute $P(A \mid D = 0)$, by supplying formulas for each of the messages that is computed. Your expressions may use factor values and values of any previously computed messages. You do not need to do numeric computation.

Solution: Depends on their factor graph.

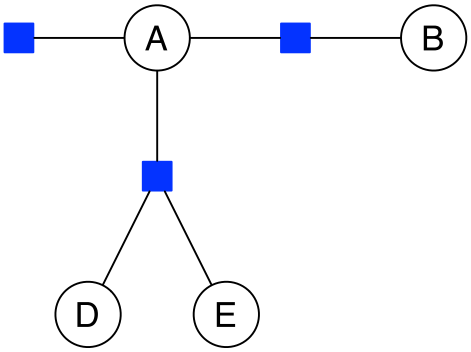
- Send message $\mu_{D \to \phi_{BCD}}$

D	$\mu_{D \to \phi_{BCD}}$
0	1
1	0

- Send message $\mu_{\phi_{BCD} \to BC}(b, c) = \sum_D \phi_{BCD}(b, c, d)\mu_{D \to \phi_{BCD}}(d)$
- Send message $\mu_{BC \to \phi_{ABC}}(b, c) = \mu_{\phi_{BCD} \to BC}(b, c)$
- Send message $\mu_{\phi_{ABC} \to A}(a) = \sum_{b,c} \mu_{BC \to \phi_{ABC}}(b, c)\phi_{ABC}(a, b, c)$
- Send message $\mu_{\phi_A \to A}(a) = \phi_A(a)$
- Final result is $\Pr(A = a) = \mu_{\phi_A \to A}(a)\mu_{\phi_{ABC} \to A}(a)$

12 Belief propagation

Consider the sum-product algorithm on this factor graph:



We will refer to factors by the set of variables they are connected to.

(a) What messages would need to be computed, and in what order?

(b) Assume the factors have the following numerical values:

$\phi_{ADE} =$

A	D	E	
0	0	0	1
0	0	1	2
0	1	0	1
0	1	1	3
1	0	0	2
1	0	1	1
1	1	0	4
1	1	1	3

$\phi_{AB} =$

A	B	
0	0	1
0	1	1
1	0	2
1	1	4

$\phi_A =$

A	
0	2
1	1

What message does A send to ϕ_{AB}? Your answer should be a table of numbers.

Solution:

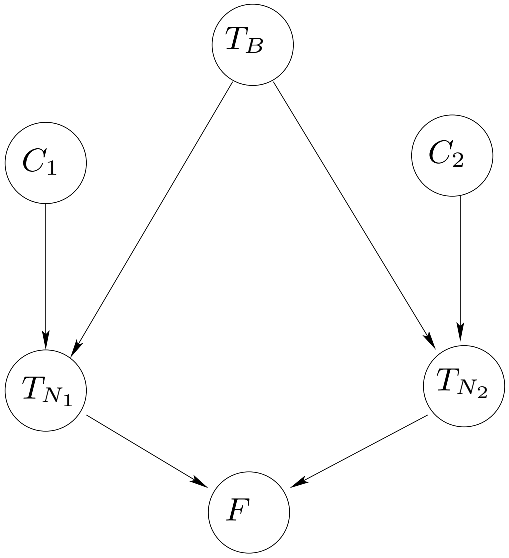
$\mu_{D\to\phi_{ADE}} = 1, \mu_{E\to\phi_{ADE}} = 1(leaves)$
$\mu_{\phi_{ADE}\to A}(a) = \sum_d \sum_e \phi(a, d, e,)\mu_{D\to\phi_{ADE}}\mu_{E\to\phi_{ADE}} = [(1+2+1+3), (2+1+4+3)] = [7, 10]$
$\mu_{\phi_A\to A}(a) = [2, 1]$
$\mu_{A\to\phi_{AB}}(a) = [7, 10]. * [2, 1] = [14, 10]$

13 Commuting

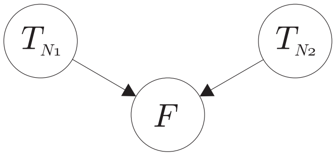
We wish to develop a graphical model for the following transportation problem. A transport company is trying to choose between two alternative routes for commuting between Boston and New York. In an experiment, two identical busses leave Boston at the same but otherwise random time, T_B. The busses take different routes, arriving at their (common) destination at times T_{N1} and T_{N2}.

Transit time for each route depends on the congestion along the route, and the two congestions are unrelated. Let us represent the random delays introduced along the routes by variables C_1 and C_2. Finally, let F represent the identity of the bus which reaches New York first. We view F as a random variable that takes values 1 or 2.

(a) Complete the following directed graph (Bayesian network) with edges so that it captures the relationships between the variables in this transportation problem.



(b) Consider the following directed graph as a possible representation of the independences between the variables T_{N1}, T_{N2} and F only:



Which of the following factorizations of the joint are consistent with the graph? Consistency here means: whatever is implied by the graph should hold for the associated distribution.

$\surd$ $\Pr(T_{N1}) \Pr(T_{N2}) \Pr(F \mid T_{N1}, T_{N2})$

$\surd$ $\Pr(T_{N1}) \Pr(T_{N2}) \Pr(F \mid T_{N1})$

$\surd$ $\Pr(T_{N1}) \Pr(T_{N2}) \Pr(F)$

14 A marginal joint (Bishop 8.26)

Consider a tree-structured factor graph over discrete variables, and suppose we wish to evaluate the joint distribution $p(x_a, x_b)$ associated with two variables x_a and x_b that do not belong to a common factor. Define a procedure for using the sum- product algorithm to evaluate this joint distribution in which one of the variables is successively clamped to each of its allowed values.

Solution: We start by using the sum-product rule to write

$$p(x_a, x_b) = p(x_b|x_a)p(x_a) = \sum_{x_{\backslash ab}} p(x)$$

where $x_{\backslash ab}$ denotes the set of all variables in the graph except x_a and x_b. We can use the sum-product algorithm (see Bishop section 8.44) to first evaluate $p(x_a)$ by marginalizing over all other variables (including x_b). Next, we successively fix x_a at each of its allowed values; for each value, we evaluate $p(x_b|x_a)$ by marginalizing over $x_{\backslash ab}$, with x_a taking its current, fixed value. Finally, we can use $p(x_a)$ and $p(x_b|x_a)$ to evaluate the joint distribution $p(x_a, x_b)$.

15 In times out(Bishop 8.23)

We showed that the marginal distribution $p(x_i)$ for a variable node x_i in a factor graph is given by the product of the messages arriving at this node from neighbouring factor nodes. Show that the marginal $p(x_i)$ can also be written as the product of the incoming message along any one of the links with the outgoing message along the same link.

Solution: The outgoing message that node x_i sends to factor f_s is the product of the incoming messages to x_i from all other links. Using this fact, we can show that the product of the incoming message to x_i along a link and the outgoing message along that same link is equivalent to the product of all incoming messages from factor nodes neighboring x_i.

$$\begin{aligned} p(x_i) &= \mu_{f_s \to x_i}(x_i)\mu_{x_i \to f_s}(x_i) \\ &= \mu_{f_s \to x_i}(x_i) \prod_{t \in ne(x_i) \setminus f_s} \mu_{f_t \to x_i}(x_i) \\ &= \prod_{s \in ne(x_i)} \mu_{f_s \to x_i}(x_i) \end{aligned} \tag{35}$$