6.867: Recitation Handout (Week 9)

November 4, 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	*	*

(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	*	*	*

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?

3 Factorization Machines

Factorization machines for an input $\mathbf{x}$, predict output $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \sum_{i=1}^{d}\sum_{j=i+1}^{d}[VV^T]_{ij}x_ix_j$. 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?

(b) Pick

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

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

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

Plug this in the Factorization Machine and do you see any similarity to Problem 3Flix?

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?

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 $c = 1$.

(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 $p(a,b,c) = p(a)p(c|a)p(b|c)$.

6 Out of gas? (Bishop 8.11)

Consider the example of the car fuel system shown in Figure 2. 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 1.

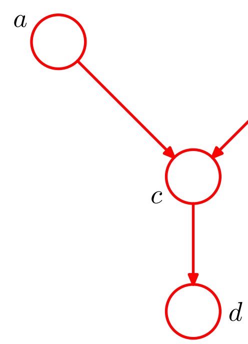


Figure 1: Bayesian Net

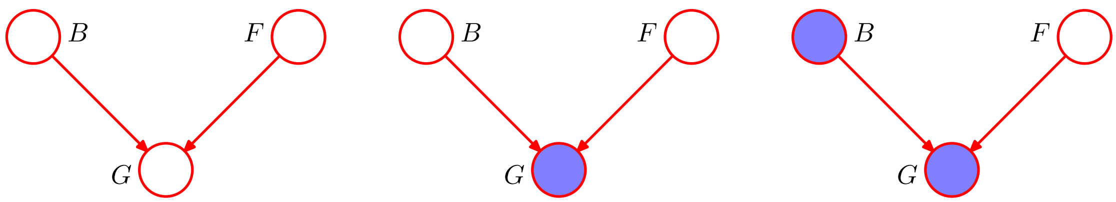
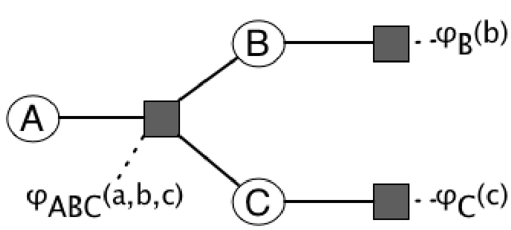


Figure 2: An example of a 3-node graph used to illustrate the phenomenon of 'explaining away'. The three nodes represent the state of the battery (B), the state of the fuel tank (F) and the reading on the electric fuel gauge (G). See the text for details.

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 $p(\hat{x}, \hat{y}) = 0$.

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.

9 Exact inference

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

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

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

(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?

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

(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.

(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.

(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

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

(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?

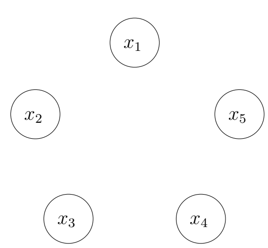
10 Easy exam?

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.

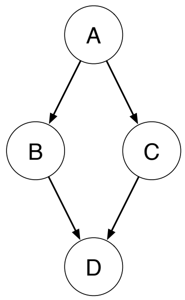


(b) What can you say about the form of the distribution over the five variables?

(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.

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

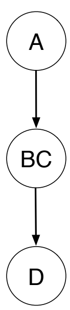
B	C	P(D = 1)
0	0	0.9
0	1	0.1
1	0	0.1
1	1	0.9

P(A = 1)
0.3

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

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

(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.



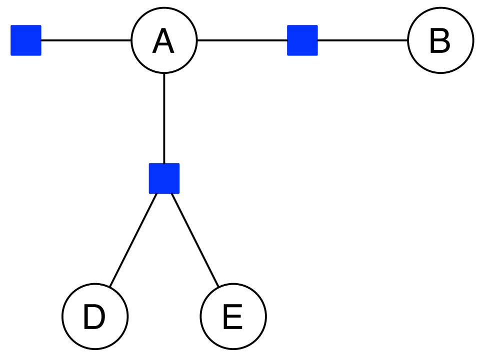
How does it compare in expressive power to the original one?
◯ More ◯ Less ◯ Same Briefly explain your answer.

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

(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.

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

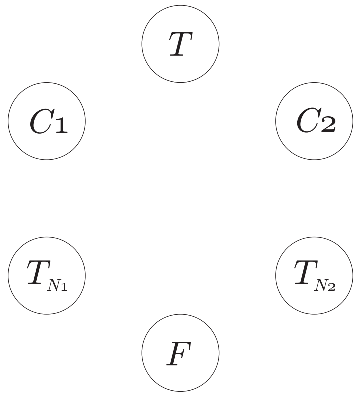
What message does A send to ϕ_{AB}? Your answer should be a table of numbers.

13 Commuting

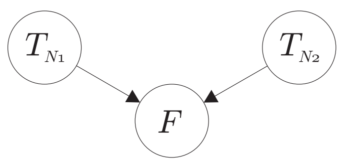
We wish to develop a graphical model for the following transportation problem. A transport company is trying to choose between two alternative routes for commuting between Boston and New York. In an experiment, two identical busses leave Boston at the same but otherwise random time, T_B. The busses take 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.

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

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.

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.