

CS 228, Winter 2008

Problem Set #3

We have provided approximate lengths with each of the problems to give you a rough estimate of how long we think each answer might be not including diagrams. These are meant to be guidelines only to make sure that answers are concise and readable (for diagrams, the larger the better).

1. Entanglement in DBNs [20 points]

- (a) [6 points] Prove Proposition 15.2.4:

Let \mathcal{I} be the influence graph for a 2-TBN $\mathcal{B}_{\rightarrow}$. Then \mathcal{I} contains a directed path from X to Y if and only if, in the unrolled DBN, for every t , there exists a directed path from $X^{(t)}$ to $Y^{(t')}$ for some $t' \geq t$.

- (b) [10 points] Prove the entanglement theorem, Theorem 15.2.5:

Let $\langle \mathcal{G}_0, \mathcal{G}_{\rightarrow} \rangle$ be a fully persistent DBN structure over $\mathcal{X} = \mathbf{X} \cup \mathbf{O}$, where the state variables $\mathbf{X}^{(t)}$ are hidden in every time slice, and the observation variables $\mathbf{O}^{(t)}$ are observed in every time slice. Furthermore, assume that, in the influence graph for $\mathcal{G}_{\rightarrow}$:

- there is a trail (not necessarily a directed path) between every pair of nodes, i.e., the graph is connected;
- every state variable X has some directed path to some evidence variable in \mathbf{O} .

Then there is no persistent independence ($\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$) which holds for every DBN $\langle \mathcal{B}_0, \mathcal{B}_{\rightarrow} \rangle$ over this DBN structure.

- (c) [4 points] Is there any 2-TBN (not necessarily fully persistent) whose unrolled DBN is a single connected component, for which ($X \perp Y \mid \mathbf{Z}$) holds persistently but ($X \perp Y \mid \emptyset$) does not? If so, give an example. If not, explain formally why not.

Estimate: 2.5 pages

2. Dirichlet Priors [25 points]

In this problem you will show that the family of mixture of Dirichlet priors is conjugate to the multinomial distribution.

- (a) [9 points] Consider the simple possibly-biased-coin setting described above. Assume we use a prior which is a mixture of two Dirichlet (Beta) distributions: $P(\theta) = 0.95 \cdot \text{Beta}(5000, 5000) + 0.05 \cdot \text{Beta}(1, 1)$; the first component represents a fair coin (for which we have seen many imaginary samples), and the second represents a possibly-biased coin, whose parameter we know nothing about. Show that the expected probability of heads given this prior (the probability of heads averaged over the prior) is $1/2$.

Suppose that we observe the data sequence $(H, H, T, H, H, H, H, H, H, H)$. Calculate the posterior over θ , $P(\theta \mid \mathcal{D})$. Show that it is also a 2-component mixture of Beta distributions, by writing the posterior in the form $\lambda^1 \text{Beta}(\alpha_1^1, \alpha_2^1) + \lambda^2 \text{Beta}(\alpha_1^2, \alpha_2^2)$. Provide actual numeric values for the different parameters $\lambda^1, \lambda^2, \alpha_1^1, \alpha_2^1, \alpha_1^2, \alpha_2^2$.

- (b) [16 points] Now generalize your calculations from part (a) to the case of a mixture of d Dirichlet priors over a k -valued multinomial parameters. More precisely, assume that the prior has the form:

$$P(\boldsymbol{\theta}) = \sum_{i=1}^d \lambda^i \text{Dirichlet}(\alpha_1^i, \dots, \alpha_k^i)$$

and prove that the posterior has the same form.

Estimate: 1-2 pages

3. Bayesian Scores [28 points]

- (a) [6 points]

Show that if \mathcal{G} is I-equivalent to \mathcal{G}' , then if we use table CPDs, we have that $\text{score}_L(\mathcal{G} : \mathcal{D}) = \text{score}_L(\mathcal{G}' : \mathcal{D})$ for any choice of \mathcal{D} .

Hint: consider the set of distributions that can be represented by parameterization each network structure.

- (b) [11 points]

Show that if \mathcal{G} is I-equivalent to \mathcal{G}' , then if we use table CPDs, we have that $\text{score}_{BIC}(\mathcal{G} : \mathcal{D}) = \text{score}_{BIC}(\mathcal{G}' : \mathcal{D})$ for any choice of \mathcal{D} .

- (c) [11 points]

Show that the Bayesian score with a K2 prior in which we have a Dirichlet prior $\text{Dirichlet}(1, 1, \dots, 1)$ for each set of multinomial parameters is not score equivalent.

Hint: construct a data set for which the score of the network $X \rightarrow Y$ differs from the score of the network $X \leftarrow Y$.

Estimate: 1-2 pages

4. Search in Structure Learning [27 points]

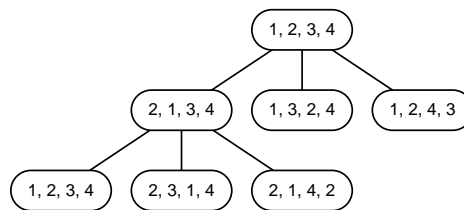


Figure 1: Partial search tree example for orderings over variables X_1, X_2, X_3, X_4 . Successors to $\prec = (1, 2, 3, 4)$ and $\prec' = (2, 1, 3, 4)$ shown.

Consider learning the structure of a Bayesian network for some given ordering, \prec , of the variables, X_1, \dots, X_n . This can be done efficiently as described in Section 19.3.2 of the course reader. Now assume that we want to perform search over the space of orderings, i.e. we are searching for the network (with bounded in-degree k) that has the highest score. We do this by defining the score of an ordering as the score of the (bounded in-degree) network with the maximum score consistent with that ordering, and then searching for the

ordering with the highest score. We bound the in-degree so that we have a smaller and smoother search space.

We will define our search operator, o , to be “Swap X_i and X_{i+1} ” for some $i = 1, \dots, n - 1$. Starting from some given ordering, \prec , we evaluate the BIC-score all successor orderings, \prec' , where a successor ordering is found by applying o to \prec (see Figure 1). We now choose a particular successor, \prec' . Provide an algorithm for computing as efficiently as possible the BIC-score for the successors of the new ordering, \prec' , given that we have already computed the scores for successors of \prec .

Estimate: 1 page