# CS 228 Problem Set 1

Hugh Zhang

February 5, 2017

## Problem 1

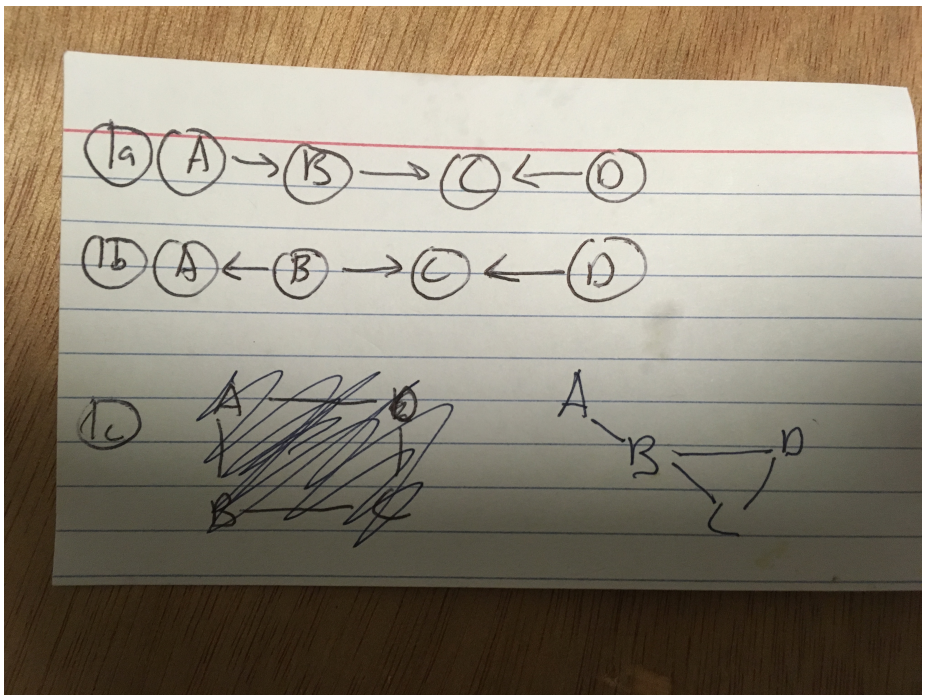


Figure 1: Problem 1

You can't differentiate parts without V structures.

### 1.1

We got this by moralizing the Bayesian net. This is not a perfect map because $B \perp\!\!\!\perp D$ but it doesn't appear that way on the Markov network. In general, moralizing doesn't give you perfect maps.

# Problem 2

## 2.1.1

$$P(C) = \sum_c P(A*, B*, c, D*)$$
$$P(C = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$
$$P(C = 1 \mid D = 0) = \frac{\frac{1}{4}}{\frac{1}{8} + \frac{1}{8} + \frac{1}{4}} = \frac{1}{2}$$
$$P(C = 1 \mid D = 1) = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{1}{2}$$

Thus, C and D are independent.

## 2.1.2

$$P(C = 1) = \frac{1}{2}$$
$$P(C = 1 \mid B = 0) = \frac{\frac{1}{4}}{\frac{1}{8} + \frac{1}{4}} = \frac{1}{3}$$

Thus, C and B are not independent.

## 2.2

There is only one possible I map for this. We can deduce this as following.

Since $C \perp\!\!\!\perp D$, we can deduce that it must be a V structure, since no other group of 3 can have independence without conditioning on anything. Thus, the arrow goes from C to A and D to A. Continuining on this, $C \not\perp\!\!\!\perp B$, we can further deduce there is NOT a v strucutre, so the arrow goes from A to B, completing a cascade. From this, we can then conclude the final arrow goes from D to B, otherwise there would be a cycle and its not a DAG. Thus, this I-map is unique.

If the underlying map G is correct, then this is a perfect map since it is unique.

## 2.3

C has no parents.

$$P(C = 0) = \frac{1}{8} + \frac{1}{8} + \frac{1}{4} = \frac{1}{2}$$

D has no parents.

$$P(D = 0) = \frac{1}{8} + \frac{1}{8} + \frac{1}{4} = \frac{1}{2}$$

A has two parents (C and D).

$$P(A = 0 \mid C = 0, D = 0) = \frac{P(A = 0, C = 0, D = 0)}{P(C = 0, D = 0)} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{8}} = \frac{1}{2}$$
$$P(A = 0 \mid C = 1, D = 0) = 0$$
$$P(A = 0 \mid C = 0, D = 1) = 0$$
$$P(A = 0 \mid C = 1, D = 1) = \frac{1/4}{1/4} = 1$$

B has two parents (A and D).

$$P(B = 0 \mid A = 0, D = 0) = \frac{P(B = 0, A = 0, D = 0)}{P(A = 0, D = 0)} = \frac{1/8}{1/8} = 1$$
$$P(B = 0 \mid A = 1, D = 0) = 0$$
$$P(B = 0 \mid A = 0, D = 1) = 0$$
$$P(B = 0 \mid A = 1, D = 1) = \frac{1/4}{1/4} = 1$$

# Problem 3

## 3.1

$$\begin{aligned} & P(h_i \mid v) = \frac{P(h_i, v)}{\sum_{h_k} P(h_k, v)} \\ = & \frac{\sum_{h_1 \dots h_{k-1}, h_{k+1} \dots h_n} P(h_k, v, h_1 \dots h_{k-1}, h_{k+1} \dots h_n)}{\sum_{h_1 \dots h_n} P(h, v)} \\ = & \frac{\sum_{h_1 \dots h_{k-1}, h_{k+1} \dots h_n} e^{-\alpha^T v - \beta^T h - v^T W h}}{\sum_{h_1 \dots h_n} e^{-\alpha^T v - \beta^T h - v^T W h}} \end{aligned}$$

v's are constant, so they can be pulled out and cancelled. All h's except for $h_i$ are also all factorable out on both the top and can be cancelled too. This is not the most easy to see, but you can break the $e^{abc}$ up into products $e^a e^b e^c$, and then factor the constants out of the sum. What's left is:

$$\frac{e^{-\beta_i^T h_i - (v^T W)_i h_i}}{\sum_{h_1} e^{-\beta_i^T h_i - (v^T W)_i h_i}}$$

This can be computed tractible. Just compute $v^T W$ and then summing over $h_1$ only requires two calculations since it is binary.

## 3.2

Notice that the observed units form a markov blanket on any given hidden unit (look at the structure of the Restricted Bolzmann Machine. Thus, we can compute $P(h_i \mid v)$ independently as above, then multiply them all together.

$$P(h \mid v) = \prod_{h_i} P(h_i \mid v)$$

Since the hidden units are all independent conditioned on the observed units, we can just factor the probabilities that way.

### 3.3

Notice that you can rewrite

### 3.4

This is equivalent to 3.3, because h and v are symmetrical.

### 3.5

I think not. Papers seem to indicate it is not so easy, and the factoring methods we used above don't work.

## Problem 4

Initution easy tells us that adding another edge gives us strictly more to work with, and thus will have a better likelihood estimation on the model.

We want to prove

$$\max_{\theta'} l_{G'}(\theta', D) \geq \max_{\theta} l_G(\theta, D)$$

Notice that since the only difference between $G$ and $G'$ is that $G'$ has an extra edge, e.g. one node in $G'$ has an extra parent. We'll focus on this node N and its parent P, thus we can ignore the first sigma $[\sum_{i=1}^{n}]$. Further, we will prove this for each example in the data set, thus it will clearly hold for the sum over all the data and we can ignore the third sigma $[\sum_{x_i}]$.

Thus, our goal can be simplified to showing

$$\sum_u \sum_P M[x, u, P] \log \theta_{x|(u,P)} \geq \sum_u M[x, u] \log \theta_{x|u}$$

Using the $\theta$ equation:

$$\sum_u \sum_P M[x, u, P] \log \theta_{x|(u,P)} = \sum_u \sum_P M[x, u, P] \log \frac{M[x, u, P]}{\sum_x M[x, u, P]}$$

Letting $k = M[x, u, P]$, $f(k) = k \log \frac{k}{\sum_x k} = k \log k - k \log \sum_x k$, we can rewrite:

$$\sum_u \sum_P M[x,u,P] \log \frac{M[x,u,P]}{\sum_x M[x,u,P]} = \sum_u \sum_P f(k)$$

Notice now, that $f(k)$ is concave since $f''(k) = \frac{1}{k} - \frac{c}{k}$, which is always positive since $k >= 0$ (its a frequency count). Thus, we can apply Jensen's inequality, which claims that

$$\begin{aligned}
&\sum_u \sum_P f(k) \geq \sum_u f(\sum_P k) \\
=&\sum_u (\sum_P k) \log \frac{\sum_P k}{\sum_P \sum_x k} \\
=&\sum_u (\sum_P M[x,u,P]) \log \frac{\sum_P M[x,u,P]}{\sum_P \sum_x M[x,u,P]}
\end{aligned}$$

Then, for the final step we realize that we are marginalizing over P, and can thus remove it, finishing the proof.

$$\sum_u (\sum_P M[x,u,P]) \log \frac{\sum_P M[x,u,P]}{\sum_P \sum_x M[x,u,P]} = \sum_u (M[x,u]) \log \frac{M[x,u]}{\sum_x M[x,u]}$$

# Problem 5

## 5.1

Test error rate: 0.9138

## 5.2

Test error rate: 0.9569

## 5.3

$$P(Democrat \mid votes) = 0.9999986$$
$$P(educationvote = 1 \mid votes) = 0.100840336134$$

## 5.4

Naive Bayes error rate on smaller data: 0.9052 TANB error rate on smaller data: 0.8836

Since TANB is a more complex model and we trained on a much smaller amount of data, it is possible that TANB overfit the data and thus is not as good as a simple model.