CS4269/6362 Machine Learning, Spring 2016: Homework 5

Zejian Zhan

Question 1:

Bayesian network:

(a) Yes. Note that from A to B we'll find X_9 is between X_5 and X_{14} . X_5 is a tail-to-tail node on each path, and X_{14} is a head-to-tail node on the path. So if X_5 and X_{14} are observed, all paths from A to B will be blocked. So we can say that A and B are d-separated given C.

(b) No. For the same reason that X_{15} is a head-to-head node on a path from A to B. So if we observe X_{15} , then there'll be at least a path from A to B. So A and B are not d-separated given C.

(c) Yes. Note that all paths from A to B will pass through X_{15}. X_{15} is head-to-head node and it is not observed. So all paths are blocked. And A and B are d-separated.

(d) No. X_{16} is a head-to-head node on a path from A to B. With the same reason explained in (b), we can see that A and B are not d-separated given C.

Markov random field:

(a) Yes. We can see that paths from A to B pass through nodes in C. So A and B are d-seperated.

(b) Yes. All paths from A to B pass through X_{15}. So A and B are d-seperated.

(c) No. There is a path from A to B ($4 \to 6 \to 11 \to 15 \to 12 \to 8 \to 5$) that does not pass any node in C. So A and B aren't d-seperated.

(d) Yes. All paths from A to B pass through X_{15}. So A and B are d-seperated.

Question 2: $A = \{X_5\}$, as $\{X_5\}$ is markov blanket of X_2.

Question 3:

1. We can derive the formula of Log-likelihood function $P(X)$ on both labeled and unlabeled data:

$$L(\theta) = \sum_{i:x^i \in L} \log \left(P(y^i) \prod_{j=1}^{m} P_j(x_j^i | y^i) \right) + \sum_{i:x^i \in U} \log \sum_{y \in Y} P(x^i, y) \tag{1}$$

m is the dimension of features, θ is the set of all parameters to be estimated: probability of class: $P(y)$, and the conditional probability of jth feature taking value x given class label: $P_j(x|y)$. Because we have some labeled data L, we can use maximum-likelihood estimation to estimate $P^0(y)$ and $P_j^0(x|y)$, which can be used later.

2. By using $\theta^0 = \{P^0(y), P_j^0(x|y)\}$, we can equivalently maximize: $\hat{L}(\theta) = \sum_{i:x^i \in U} \log \sum_{y \in Y} P(x^i, y)$

Since the above equation is hard to optimize as there is a summation inside logarithm, we need to find a way to move summation to the outside of logarithm:

$$\log \sum_{y \in Y} P(x^i, y) \geq \sum_{y \in Y} \delta(y|x^i) \log P(x^i, y) \tag{2}$$

(since $log(x)$ is a concave function, and $\sum_{y \in Y} \delta(y|x^i) = 1,\ 0 \leq \delta(y|x^i) \leq 1$), we derive an auxiliary function:

$$Q(\theta, \theta^{t-1}) = \sum_{i:x^i \in U} \sum_{y \in Y} \delta(y|x^i) \log P(x^i, y) \tag{3}$$

where:

$$\delta(y|x^i) = \frac{P^{t-1}(y) \sum_{j=1}^{m} P_j^{t-1}(x_j^i|y)}{\sum_{y \in Y} P^{t-1}(y) \sum_{j=1}^{m} P_j^{t-1}(x_j^i|y)} \tag{4}$$

$Q(\theta, \theta^{t-1})$ is conditional expectation of complete-data log-likelihood conditioned on posterior distribution $\delta(y|x^i)$, so we finish E step.

For M step:

$$P^{t-1}(y) = \frac{1}{|U|} \sum_{i:x^i \in U} \delta(y|x^i) \tag{5}$$

$$P_j^{t-1}(x|y) = \frac{\sum_{i:x^i \in U \,\&\, x_j^i = x} \delta(y|x^i)}{\sum_{x^i \in U} \delta(y|x^i)} \tag{6}$$