STATS 200A: Homework #4

Professor Yingnian Wu
Assignment: 1, 2

Eric Chuu
UID: 604406828

October 27, 2016

Problem 1

Smoking habit, health, age problem.

(a) Assuming conditional independence of X and Y given Z, show that, marginally, X and Y are not independent.

(b) Not assuming conditional independence, calculate

$$P(Y = y|X = \text{observed at } x) \tag{1}$$
$$P(Y = y|X = \text{do } (x)) \tag{2}$$

Solution
(a) Suppose Ω is the population of smokers. Then let X be a random variable such that

$$X(\omega) = \{\text{pipe, cigarette}\}$$

so that for each smoker $\omega \in \Omega$, he/she either smokes using a pipe or a cigarette. Let Y be a random variable that indicates the health of a smoker:

$$Y(\omega) = \{\text{healthy, unhealthy}\}$$

Finally, let Z be a random variable that maps each smoker to his/her age group:

$$Z(\omega) = \{\text{young, old}\}$$

Assuming conditional independence of X, Y given Z, we can then calculate the probability of the event $\{X = x, Y = y\}$,

$$\begin{aligned}
P(X = x, Y = y) &= \sum_z P(X = x, Y = y, Z = z) &(3)\\
&= \sum_z P(X = x, Y = y|Z = z)P(Z = z) &(4)\\
&= \sum_z P(X = x|Z = z)P(Y = y|Z = z)P(Z = z) &(5)
\end{aligned}$$

where equalities (3), (4), and (5) follow from the law of total probability, the definition of conditional probability, and the assumption of conditional independence. To make a more intuitive conclusion, we consider the following quantities and again invoke conditional independence:

$$\begin{aligned}
&P(Y = \text{healthy } | X = \text{cigarette}, Z = \text{young}) && (6)\\
&= P(Y = \text{healthy } | Z = \text{young}) && (7)\\
&= P(Y = \text{healthy } | X = \text{pipe}, Z = \text{young}) && (8)
\end{aligned}$$

In other words, within the same age group, the proportion of healthy cigarette smokers and healthy pipe smokers is equal, i.e., within the same age group, the choice of smoking a cigarette or a pipe has no effect on health status. Thus, we can conclude that X and Y are not marginally independent. □

(b) If we do not assume conditional independence, then we can calculate (1) and (2) as follows

$$\begin{aligned}
&P(Y = y|X = \text{observed at } x)\\
&= P(Y = y|X = x)\\
&= P(Y = y|X = x)\\
&= p(y|x)\\
&= \sum_z p(y|x, z)p(z|x)\\
&P(Y = y|X = \text{do(x)})\\
&= P(Y = y|X \leftarrow x)\\
&= p(y|do(x))\\
&= \sum_z p(y|x, z)p(z)
\end{aligned}$$

□

Problem 2

Asia problem.

(a) Calculate the joint distribution of the variables

(b) Calculate the conditional probabilities of some queries, i.e., $P(\text{unknown} \mid \text{known})$

Solution
(a) If we define the following Bernoulli random variables

$$\begin{aligned}
V &= \text{Visit to Asia} \\
S &= \text{Smoking} \\
T &= \text{Tuberculosis} \\
L &= \text{Lung Cancer} \\
C &= \text{Tuberculosis or Cancer} \\
B &= \text{Bronchitis} \\
T &= \text{Tuberculosis} \\
X &= \text{X-Ray Result} \\
D &= \text{Dyspnea}
\end{aligned}$$

where each random variable maps to yes/no. Then using the Bayesian network described in lecture, we can calculate the joint distribution of the variables as follows:

$$\begin{aligned}
p(v, s, t, l, b, c, x, d) = p(v) \cdot p(s) \cdot p(t|v) \cdot p(l|s) \cdot p(b|s) \cdot \\
p(c|t, l) \cdot p(x|c) \cdot p(d|c, b)
\end{aligned}$$

□

(b) We can calculate some queries of the form $P(\text{uknown}|\text{known})$

$$\begin{aligned}
p(l|c, x, s) &= \frac{p(l, c, x, s)}{p(c, x, s)} \\
p(c|x, t, v) &= \frac{p(c, x, t, v)}{p(x, t, v)} \\
p(v|t, c, d) &= \frac{p(v, t, c, d)}{p(t, c, d)}
\end{aligned}$$

where the right hand side of all the equations can be calculated from marginalizing out the variables not present. □