# HOMEWORK 6 471

HW6 is due by Thursday December 3th at 11:59 pm CT . Please upload your solutions to Canvas.

**Attach the codes/command windows that you use to answer the questions.**

(1) (EM for HMMs) For this problem we consider the same set-up as in HW5, the only difference being that now we assume we don't know the values of $q$ and $\sigma^2$. Use the vector $y^*$ from HW5 and the EM algorithm to estimate them.

**Hint:** It will be very useful to write some code whose inputs are $p, \sigma^2, y^*$ and whose outputs are the quantities:

$$\sum_{i=1}^{999} \mathbb{E}_{X|y^*}^{p,\sigma^2}(\mathbb{1}_{X_i=X_{i+1}}),$$

and

$$\sum_{i=1}^{1000} \mathbb{E}_{X|y^*}^{p,\sigma^2}((X_i - y_i^*)^2).$$

In the above, $\mathbb{E}_{X|y^*}^{p,\sigma^2}$ represents conditional expectation given $Y = y^*$ and given that the parameters of the model are $p$ and $\sigma^2$.

(2) Suppose that the random variables $\{X_1, \ldots, X_n\}$ have joint distribution $p$. Does there exist a directed acyclic graph $G = (V, E)$ such that the variables are a BN with respect to $G$? If your answer is yes, then you should be able to write $p$ as

$$p(x_{1:n}) = \prod_{j=1}^{n} p_{j|pa(j)}(x_j|x_{pa(j)}),$$

where $p_{j|pa(j)}$ is the conditional distribution of $X_j$ given $X_{pa(j)}$.

(3) In B. , the city where Pacifica and Solo live, researchers (who studied in the school S and P attend) have determined that the medical status of an individual in the population is described by eight relevant variables: Gender ($G$), Smoking habit ($S$), severity of Angina ($A$), age ($Y$), Weight ($W$), Blood pressure ($B$), Cholesterol level ($C$), and Lung cancer cell type ($L$).

Intense work from the researchers has furthermore showed, that the joint probability distribution of these 8 variables factors as

$$p = p_G \cdot p_A \cdot p_{S|A,G} \cdot p_{W|A} \cdot p_{C|W} \cdot p_{B|C,W} \cdot p_{Y|C} \cdot p_{L|S,Y}$$

where

- $G \in \{0, 1\}$, where $0 = $ female, $1 = $ male.
- $S \in \{0, 1, ..., 60\}$, $S$ represents the number of cigarettes smoked per day.
- $A \in \{0, 1, 2, 3, 4\}$, where $0 = $ no angina, $1 = $ mild angina, . . . , $4 = $ severe angina.
- $Y \in \{16, 17, ..., 110\}$, $Y$ stands for age in years.
- $W \in \{60, 61, ..., 400\}$, $W$ stands for weight in pounds.
- $B \in \{50, 51, ..., 250\}$, $B$ stands for systolic blood pressure in mmHg.
- $C \in \{80, 81, ..., 350\}$, stands for total cholesterol in mg/dL.
- $L \in \{0, 1, 2\}$, where $0 = $ no lung cancer, $1 = $ small-cell carcinoma , $2 = $ non-small-cell carcinoma.

Suppose that you can evaluate each of the functions $p_G, p_A, p_{S|A,G}, p_{W|A}, p_{C|W}, p_{B|C,W}, p_{Y|C}, p_{L|S,Y}$ at no computational cost.

(a) Draw a directed acyclic graph to express the joint distribution of these 8 variables as a BN. (Your DAG should have as few edges as possible.)

(b) What is the associated moral graph respected by the variables?

(c) • Consider the conditional distribution of $(L, Y, C, B, W)$ given that $G = g$, $A = a$, and $S = s$. What is the minimal directed acyclic graph with respect to which this conditional distribution is guaranteed to be a BN? Explain.

• What is the minimal DAG with respect to which $(C, W, B)$, given $G = g$, $A = a$, and $S = s$, is guaranteed to be a BN? Explain.

• What is the minimal DAG with respect to which the marginal (unconditioned) distribution on $(G, A, S, W, C, B)$ is guaranteed to be a BN? Explain.

• What is the minimal DAG with respect to which the marginal (unconditioned) distribution on $(G, A, S, W, B)$ is guaranteed to be a BN? Explain.

(d) How would you compute $\mathbb{P}(L = 2 | S \geq 30)$ exactly?

(e) Pacifica's aunt, who is 50 years old, has serious angina $A \geq 3$, and smokes 20 to 30 cigarettes a day. The doctor wants to know the conditional probability that her cholesterol level is $120mg/dL$ or higher. How would you compute this quantity exactly?

(f) The doctors also wanted to know what is the probability that P 's aunt has lung cancer ( $L \geq 1$ ) given her age, her level of angina and the range of cigarettes she smokes a day. How would you compute this exactly?