Homework 3 of CS 291K (Fall 2022)

University of California, Santa Barbara

Due Nov 17, 2022 (Thursday)

Notes:

- There are required parts of the homework and bonus questions.
- The required parts of the homework are required.
- You are welcomed to discuss with your peers / TAs, but you need to write the solutions yourself and declare any help you got.

1 Why should I do this homework?

This homework is given for you to practice what you learned in BayesNet (Problem 1 - 2). In Problem 1, you will practice modeling with BayesNet. In Problem 2, you will practice reading conditional independences from the graph.

Problem 3 and 4 are bonus questions. They help you to connect BayesNet to machine learning. Specifically, Problem 3 asks you to derive the well-known naive Bayes classifier. Problem 4 teaches you something about the notorious Hidden Markov Models (HMM). While Problem 4 is a challenge question, part (a) - (c) are short and highly doable. Problem 5 is about sampling methods. The instructors highly recommend that you try to solve at least one of them. Otherwise you will not appreciate why BayesNet is useful in modeling the world and constructing ML models.

2 Homework problems

Problem 1 (30')

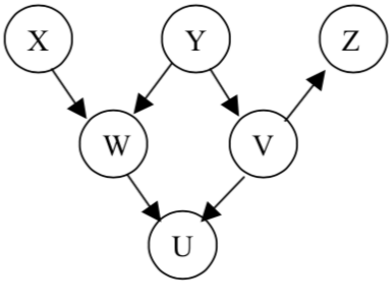
The patient has a probability to recover depending on whether s/he receives the drug, how old s/he is, and which gender the patient has. A doctor gives a patient a drug dependent on their age and gender. Additionally, it is known that age and gender are independent.

(a) (10') Draw the Bayesian network which describes this situation.

(b) (10') Factorize the joint probability distribution into CPTs according to the BayesNet.

(c) (10') Write down the formula to compute the probability that a patient recovers given that you know if s/he gets the drug. Write down the formula using only CPTs.

Problem 2 (70') Consider the Bayes Net below:



(a) (10') Is it true that $P(X|Y, W) = P(X|W)$? Explain.

(b) (10') Write down the expression for computing $P(X|Y)$ using the above Bayes Net.

(c) (10') Are variables X,W conditionally independent of variables V,Z, given Y? Explain.

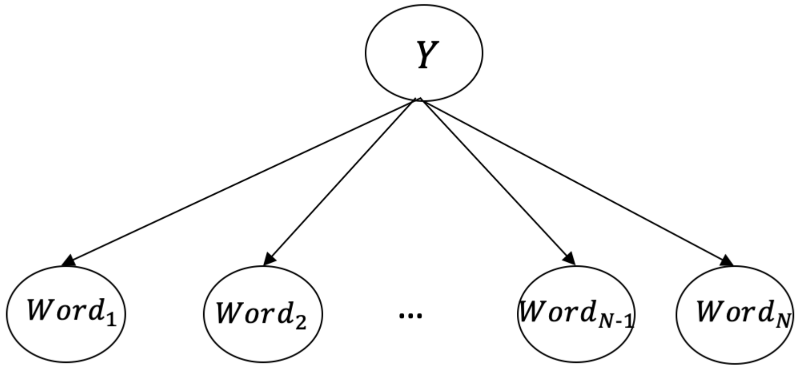
(d) (10') Are variables X,W conditionally independent of variables V,Z, given U? Explain.

(e) (10') Are variables W and Z independent? Explain.

(f) (10') Write down the Markov Blanket of variable W and variable Y.

(g) (10') Assume all the variables are binary, either take value 0 or 1. Write down the expression to compute $P(U = 1, V = 1, W = 1, X = 0, Y = 0, Z = 1)$ using notation like $P(X = 1|W = 0)$.

Problem 3 Multinomial Naive Bayes model (Optional / Bonus) Consider an author-word model. The random variable Y indicates the author and could take k possible values: $\{y_1, \ldots, y_k\}$ (e.g., {Douglas Adams, Tolkien, Xueqin Cao, Andrzej Sapkowski}). Random variables $W_1, ..., W_N$ or $\text{word}_1, ..., \text{word}_N$ indicates the words that the author wrotes in a novel and each word could take any value from a vocabulary of size d — $\{w_1, \ldots, w_d\}$. The joint distribution of these random variables are described by the following BayesNet (or directed graphical model).



(a) (5') Write down the factorization of the joint distribution according to the graphical model.

(b) (5') What is the smallest number of free parameters needed to represent the joint distribution in the following two situations? List all the unique CPTs and their dimensionality in both cases.

i. (5') When $\text{word}_1|\text{Author}, \text{word}_2|\text{Author}, \text{word}_3|\text{Author}, ..., \text{word}_N|\text{Author}$ are identically distributed, i.e., $\mathbb{P}(\text{word}_i|\text{Author}) = \mathbb{P}(\text{word}_j|\text{Author})$ for all i, j.

ii. (5') When we do not make the above assumption?

(c) (5') Under the assumption in Part (b)i., write down $\mathbb{P}(\text{Author}|\text{word}_1, \text{word}_2, \dots, \text{word}_N)$ using the unique CPTs that you specified above. Simplify the expression as much as possible. (Hint: Apply Bayes Rule)

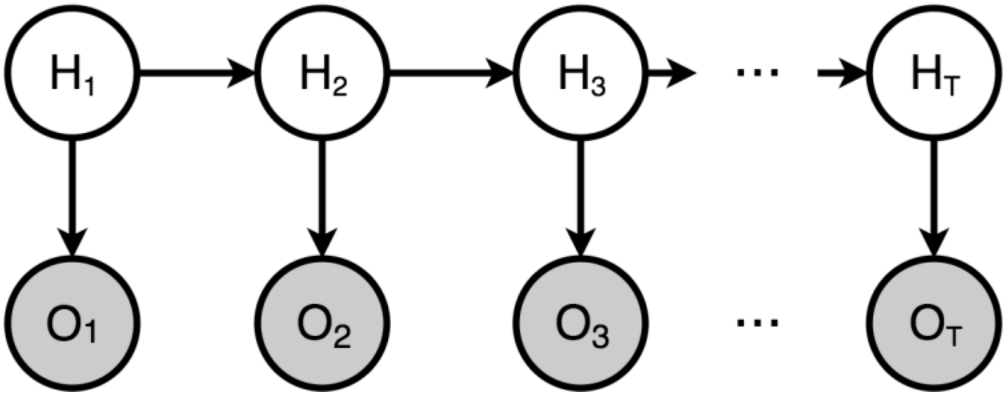
(d) (10') Now let's consider the learning problem. Suppose we observe n data points of the form: $(\mathbf{W}^{(i)}, Y^{(i)})$for i $= 1, \dots, n)$, where the text is $\mathbf{W}^{(i)} = [W_1^{(i)}, \dots, W_N^{(i)}]^T \in \{w_1, \dots, w_d\}^N$, and the label $Y^{(i)} \in \{y_1, \dots, y_k\}$.Derive the maximum likelihood estimates of the model parameters (the CPTs).

(e) (10') Based on the solution you get in Part (c), show that the natural classifier that output

$$\hat{y} = \operatorname*{argmax}_{y \in \{y_1, \dots, y_k\}} P(Y = y|W_1, \dots, W_N)$$

is a linear classifier. Recall that the definition of the linear classifier (you should be familiar with it already from HW1) is $h_\theta(x) = \text{argmax}_{y \in \{y_1, \dots, y_k\}} \Theta^T x$. Be specific on what the feature vector x is and what the weight matrix Θ is; as well as what their dimensionalities are.

Problem 4 Hidden Markov Models (Optional / Bonus / Challenge Problem) In this question, we will learn HMM as is shown in the BayesNet model below. Let all variables be discrete. In particular, let O_i be a discrete random variable that could take d possible values, and H_i be a discrete random variable that could take k possible values.



The parameters of the HMM model are simply the CPTs of the graphical model, i.e.,

$$P(H_1) = \theta \in \mathbb{R}^k,$$
$$P(H_{i+1}|H_i) = A \in \mathbb{R}^{k\times k} \text{ for all } i = 1, 2, 3, ..., T-1,$$
$$P(O_i|H_i) = B \in \mathbb{R}^{d\times k} \text{ for all } i = 1, 2, 3, ..., T.$$

Canonically, parameter θ, A, B are called the "initial state distribution", "transition probabilities" and "emission probabilities" in standard HMM jargon.

Convince yourself the dimensionality of these CPTs are correct.

Note that the transition and emission probabilities are *the same* for all $i = 1, ..., T$.

(a) (5') Write down the joint probability of $P(H_1, ..., H_T, O_1, ..., O_T)$ in factorized form as function of the CPTs θ, A, B.

(b) (5') Write down the probability distribution of the observed variables $P(O_1, ..., O_T)$ as a function of the CPTs θ, A, B.

(hint: This is identical to expressing $P(O_1, ..., O_T)$ using CPTs, but the parameters are shared. The final expression (if you use a matrix form, will be quite clean))

Remark: The above probability distribution $P(O_1, ..., O_T)$ is jointly parametrized by values of $O_1, ..., O_T$, and the values of θ, A, B. When we view it as a function of θ, A, B, while keeping $O_1, ..., O_T$ fixed, Then this function is known as the likelihood function: $L(O_1, ..., O_T; \theta, A, B)$. This measures the likelihood of observing $O_1, ..., O_T$ when the data generating distribution is specified by θ, A, B.

Given a sequence of observation $[O_1, ..., O_T] = [o_1, ..., o_T]$, the parameters A, B, θ that maximizes the likelihood, i.e.

$$[\hat{\theta}, \hat{A}, \hat{B}] = \operatorname*{argmax}_{A,B,\theta} L(O_1 = o_1, ..., O_T = o_t; \theta, A, B)$$

is called the maximum likelihood estimator.

Solving the optimization for this MLE is not easy. It is not a convex optimization problem and we will have to use the EM algorithm to find a local optimal solution. The E-step alone requires using dynamic programming — a Forward-Backward algorithm (closely related to the more famous Viterbi algorithm). The EM solution itself is known as the Baum-Welch algorithm. Rest assured. You are *not* going to derive that in this homework.

We will take an alternative route using only things that we have learned from the class.

(c) (5') Show (using the rules of d-separation or otherwise) that for $2 \leq i \leq T-1$, O_{i-1}, O_i, O_{i+1} are conditionally independent given H_i.

(d) (5') Use the conditional independence in (c) to show that:

$$P(O_1, O_2, O_3) = \sum_{i=1}^{k} P(H_2 = i) P(O_1 | H_2 = i) P(O_2 | H_2 = i) P(O_3 | H_2 = i). \tag{1}$$

(e) (10') Let O_1, O_2, O_3 be discrete random variables with d possible values and H_2 be a discrete random variable with k possible values.

- What is the total number of independent numbers to describe $P(H_2)$, $P(O_2|H_2)$, $P(O_1|H_2)$, $P(O_3|H_2)$ in terms of k and d?

- Let us enumerate all combinations of O_1, O_2, O_3 in (1), how many equations do we get in total?

- Note that the LHS of (1) can be estimated from the data directly and the RHS are all unknown parameters. By solving the system of (nonlinear) equations, we can potentially identify the unknowns: $P(H_2), P(O_2|H_2), P(O_1|H_2), P(O_3|H_2)$. What is a condition on k, d such that we have enough equations to identify all unknowns variables? (Assume that we need one equation for one unknown.)

  (Hint: the number of unknown variables are the same as the number of independent parameters)

(f) (10') If we can solve the nonlinear equations about, we can then identify

$$P(H_2), P(O_2|H_2), P(O_1|H_2), P(O_3|H_2).$$

But these are not the CPTs. If the CPTs are ultimately what we want to learn, then we need an set of equations to convert these quantities back to CPTs.

Write $P(O_2|H_2)$, $P(O_3|H_2)$ and $P(O_1|H_2)$ in terms of the model parameters (the CPTs): θ, A, B.

Problem 5 Sampling (Optional/Bonus) (30') We want to generate samples from Cauchy distribution, defined as

$$p(x) = \frac{1}{\pi 1 + x^2}$$

1. (10') Please implement the code to generate 100 samples from Cauchy distribution using the importance sampling method with standard Normal distribution as the proposal. Plot the histogram of generated samples.

2. (20') Please implement the code to generate 100 samples from Cauchy distribution using Metropolis-Hastings algorithm with the standard Normal distribution as the proposal. Plot the histogram of generated samples.

You need to implement it in Jupyter Notebook and submit the PDF file.