CIS 520, Machine Learning, Fall 2020
Homework 7
Due: Monday, November 23rd, 11:59pm
Submit to Gradescope

Instructions. Please write up your responses to the following problems clearly and concisely. We require you to write up your responses using LaTeX; we have provided a LaTeX template, available on Canvas, to make this easier. **Submit your answers in PDF form to Gradescope. We will not accept paper copies of the homework.**

Collaboration. You are allowed and encouraged to work together. You may discuss the **written homework** to understand the problem and reach a solution in groups. However, **it is recommended that each student also write down the solution independently and without referring to written notes from the joint session.** You must understand the solution well enough to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

Learning Objectives

After completing this assignment, you will be able to:

- understand how HMM models and other belief nets represent probability distributions
- understand the relation between HMMs and RNNs.
- be able to learn belief net stuctures

Deliverable

This homework can be completed individually or in groups of 2. You need to make one submission per group. Make sure to add your team member's name on Gradescope when submitting the homework's written and coding part.

1. **A PDF compilation of** `hw7_template.tex`

Note that there are 93 points in total, and we will scale it out of 100 when calculating your final score.

1 Hidden Markov Models [20 points]

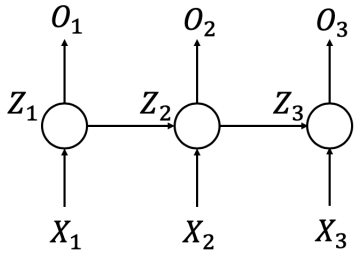
1. **[8 points]** For each of the following problems, is it appropriate to use HMM? Provide a one sentence explanation to your answer.
   (a) Daily precipitation data in Philadelphia
   (b) Optical character recognition
   (c) Netflix challenge: predict how much someone will enjoy a movie based on their and others' movie preferences
   (d) Stock market price data
2. **[4 points]** Are the following statements true or false? Provide brief justifications.
   (a) The HMM directly models the dependency of each hidden state on all previous hidden states.
   (b) When learning an HMM for a dataset, if we do not know the true number of hidden states, we can always increase the training data likelihood by permitting more hidden states.
3. **[4 points]** Consider the problem of predicting whether a person is **asleep** or **awake** at any given moment of the day. Suppose you choose to use HMM. Is the transition matrix consistent? In other words, will the probability of transitioning from **asleep** to **awake** stay constant? How does this example say about the weakness of HMM? Where does this weakness come from? (In reality we have many tricks to overcome this particular obstacle, but we are talking about "vanilla" HMM here.)
4. **[4 points]** Propose another model that could potentially better solve the above problem.

2 Recurrent Neural Networks [20 points]

On any given day, Alice is in one of the two states: happy or sad. You do not know her internal state, but get to observe her activities in the evening. Each evening, she either sings, goes for a walk, or watches TV.

In this problem, we will use RNNs for predicting future activities. Here, you are given a sequence of observations $X_1, X_2, \ldots, X_t$, which will be used to estimate the hidden state Z_{t+1}, which will then give a probability distribution over possible next observations X_{t+1}. Both RNN and HMM are Markovian: Z_{t+1} is an "encoding" of the previous observations such that $p(X_{t+1}|Z_{t+1}) = p(X_{t+1}|Z_{t+1}, X_t, X_{t-1} \ldots X_1)$.

The simple RNN model can be expressed as follows (for a 3 day sequence):



The input X_t is Alice's activity on day t, the hidden state Z_t is the state of Alice on day t, and the output O_t is the probability of Alice's activity on day $(t+1)$ given her activities on day $1, 2, \ldots, t$.

Assume the following:

$$Z_t = \text{ReLU}(UX_t + WZ_{t-1})$$
$$O_t = \text{softmax}(VZ_t)$$

We will represent Alice's actions as one-hot vectors, with the follow entry correspondence: $[\mathsf{sing}, \mathsf{walk}, \mathsf{TV}]^T$ (3 by 1).

Suppose we have the observed that Alice's actions on the first two days are (sing, TV), represented as $X_1 = [1, 0, 0]^T, X_2 = [0, 0, 1]^T$, and assume her (unobservable) initial mood is represented as $Z_1 = [0.5, 0.5]^T$. Also, assume we have estimated model parameters

$$U = \begin{bmatrix} 5.5 & 2.3 & 2 \\ 3.2 & 7.1 & 0.5 \end{bmatrix}$$

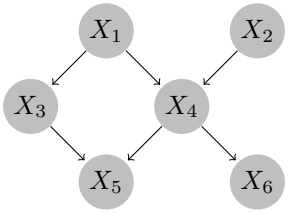
$$W = \begin{bmatrix} 0.5 & 0.2 \\ 2 & 0.9 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.4 & 1.3 \\ 0.6 & 0.9 \\ 1.6 & 0.3 \end{bmatrix},$$

1. Then, estimate the most likely activity on day 3 as follows:
   (a) **[4 points]** Give the formula to calculate Z_2 and O_2.
   (b) **[4 points]** Find the values of Z_2 and O_2.
   (c) **[2 points]** What is the most likely predicted activity on day 3?
2. Now estimate the most likely activity on day 4. To do this, use $X_3 = O_2$ and proceed as follows:
   (a) **[4 points]** Give the formula to calculate Z_3 and O_3.
   (b) **[4 points]** Find the values of Z_3 and O_3.
   (c) **[2 points]** What is the most likely predicted activity on day 4?

3 Bayesian Networks [20 points]

Consider the Bayesian network over 6 random variables $X_1, X_2, X_3, X_4, X_5, X_6$ shown below (assume for simplicity that each random variable takes 2 possible values):



1. **[2 points]** Write an expression for the joint probability mass function $p(X_1, X_2, X_3, X_4, X_5, X_6)$ that makes the same (conditional) independence assumptions as the Bayesian network above.

2. **[3 points]** Consider a joint probability distribution satisfying the following factorization:

$$p(X_1, X_2, X_3, X_4, X_5, X_6) = p(X_1)p(X_2)p(X_3)p(X_4)p(X_5 \mid X_3)p(X_6 \mid X_3)\,.$$

   Is this distribution included in the class of joint probability distributions that can be represented by the Bayesian network above? Briefly explain your answer.

3. **[3 points]** If the edge from X_3 to X_5 is removed from the above network, will the class of joint probability distributions that can be represented by the resulting Bayesian network be smaller or larger than that associated with the original network? Briefly explain your answer.

4. **[12 points]** Given the above figure, determine whether each of the following is true or false. Briefly justify your answer.

   (a) $p(X_1, X_2) = p(X_1)p(X_2)$

   (b) $p(X_3, X_6|X_4) = p(X_3|X_4)p(X_6|X_4)$

   (c) $p(X_1, X_2 \mid X_6) = p(X_1 \mid X_6)p(X_2 \mid X_6)$

   (d) $p(X_2, X_5 \mid X_4) = p(X_2 \mid X_4)p(X_5 \mid X_4)$

4 Belief Net Construction [33 points]

Given the following observed counts for all different combinations of the binary random variables A, B, C and D (each variable can be true (T) or false (F)), construct a belief net using the algorithm described in class, where variables are added sequentially to the network. Note that there are 3400 observations in total. *Consider the variables in the order A, B, C, D*, and make sure to give *both* the graph and the conditional probability tables.

A	B	C	D	Count
T	T	T	T	600
T	T	T	F	200
T	T	F	T	0
T	T	F	F	800
T	F	T	T	400
T	F	T	F	0
T	F	F	T	0
T	F	F	F	200
F	T	T	T	200
F	T	T	F	400
F	T	F	T	0
F	T	F	F	200
F	F	T	T	200
F	F	T	F	0
F	F	F	T	0
F	F	F	F	200

In this problem, you will be comparing many conditional probabilities to assess whether some variables are independent of others. In modeling situations like this we need some criteria for deciding whether one variable depends on the other based on the observed conditional probabilities. For this problem, let's suppose if the range of conditional probabilities for a set of variables is larger than 0.05, then the variables are dependent. (In reality, we might evaluate dependence by either i) doing a statistical significance test, or ii)

regularizing the Bayes net by adding a cost for each new connection).

Remember to check for both single and joint dependencies between variables. The key question is always to ask for each possible link that one might add to the graph: can it be removed? In building your belief net, follow the algorithm discussed in class:

1. Add A
2. **[5 points]** Add B and decide whether to add a link from A to B. Report if you add the link (1 point) and show steps (4 points).
3. **[5 points]** Add C and decide whether to add a link from A to C. Report if you add the link (1 point) and show steps (4 points).
4. **[5 points]** Decide whether to add a link from B to C. Report if you add the link (1 point) and show steps (4 points).
5. **[5 points]** Add D and decide whether to add a link from A to D. Report if you add the link (1 point) and show steps (4 points).
6. **[5 points]** Decide whether to add a link from B to D. Report if you add the link (1 point) and show steps (4 points).
7. **[5 points]** Decide whether to add a link from C to D. Report if you add the link (1 point) and show steps (4 points).
8. **[3 points]** Draw the final constructed network.

Again, be sure to give the conditional probability tables for downstream variables when appropriate.