CS 475 Machine Learning: Homework 5
Graphical Models, Inference, and Structured Prediction
Analytical Problems

Due: Saturday May 2, 2020, 11:59 pm

40 Points Total   Version 1.0

YOUR_NAME (YOUR_JHED)

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not change any formatting in this document, or we may be unable to grade your work. This includes, but is not limited to, the height of textboxes, font sizes, and the spacing of text and tables. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.

We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

1) Probabilistic PCA (10 points)

Draw a directed probabilistic graphical model representing a discrete mixture of probabilistic PCA models in which each PCA model has its own values of $\mathbf{W}$, $\boldsymbol{\mu}$, and σ^2. Then draw a modified graph in which these parameter values are shared between the components of the mixture. The graph should represent the model for a single data point $\mathbf{x}$. You can make these diagrams by hand or in another program and include them here as an image. (Hint: refer to slide 24 of the Dimensionality Reduction lecture as a starting point.)

2) Factorizations of MRFs (10 points)

The probability density function of most Markov Random Fields cannot be factorized as the product of a few conditional probabilities. This question explores some MRFs that can be factorized in this way.

Consider the graph structure in Figure 1. From this graph, we know that X_2 and X_3 are conditionally independent given X_1.

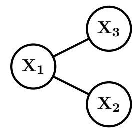

Figure 1: The Original Undirected Graph

We can draw the corresponding directed graph as Figure 2.

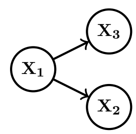

Figure 2: The Converted Directed Graph

This suggests the following factorization of the joint probability:

$$P(X_1, X_2, X_3) = P(X_3|X_1)P(X_2|X_1)P(X_1)$$

Now consider the following graphical model in Figure 3.

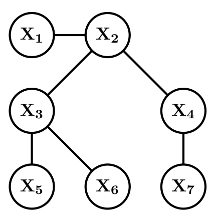

Figure 3: An Undirected Graph

As before, we can read the conditional independence relations from the graph.

(a) Following the example above, write a factorization of the joint distribution into directed, conditional probability distributions:

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7).$$

(b) Is this factorization unique, meaning, could you have written other factorizations that correspond this model? If the factorization is unique, explain why it is unique. If it is not unique, provide an alternate factorization.

(c) What is it about these examples that allows them to be factored in this way?

3) Generative model and Discriminative model (10 points)

Consider the graphical model shown in Figure 4. In this model, $\mathbf{x}$ is a sequence of observations for which we want to output a prediction $\mathbf{y}$, which itself is a sequence, where the size of $\mathbf{y}$ is the same as $\mathbf{x}$. Unlike sequence models we discussed in class, this model has a tree structure over the hidden nodes. Assume that the potential functions have a log-linear form: $\psi(Z) = \exp\{\sum_i \theta_i f_i(Z)\}$, where Z is the set of nodes that are arguments to the potential function (i.e. some combination of nodes in $\mathbf{x}$ and $\mathbf{y}$,) θ are the parameters of the potential functions and f_i is a feature function.

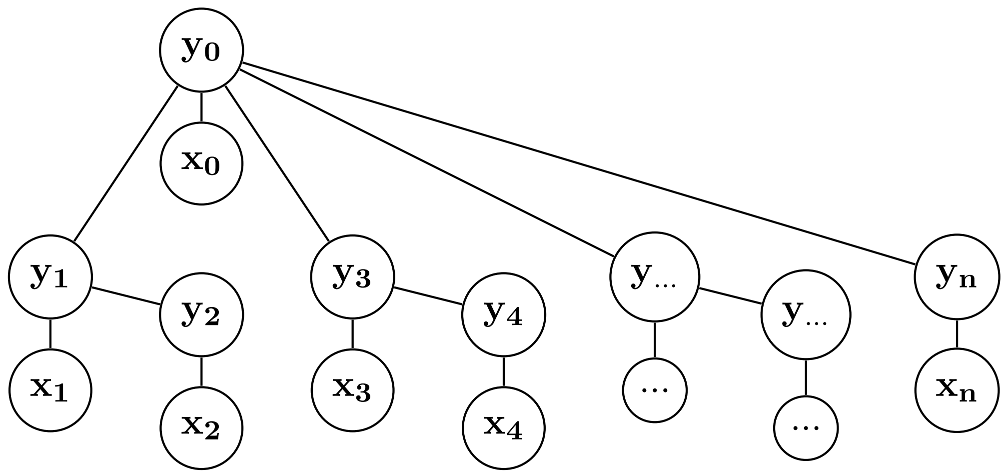

Figure 4: Tree structure model

(a) Write the log likelihood for this model of a single instance $\mathbf{x}$: $\log p(\mathbf{y}, \mathbf{x})$.

(b) Write the conditional log likelihood for this model of a single instance $\mathbf{x}$: $\log p(\mathbf{y}|\mathbf{x})$.

(c) Assume that each variable y_i can take one of k possible states, and variable x_i can take one of k' possible states, where k' is very large. Describe the computational challenges of modeling $\log p(\mathbf{y}, \mathbf{x})$ vs $\log p(\mathbf{y}|\mathbf{x})$.

(d) Propose an efficient algorithm for making a prediction for $\mathbf{y}$ given $\mathbf{x}$ and θ.

4) Sequence Classification (10 points)

Consider the following sequence classification task. We are given sequences of words, and only three words can appear in a sequence: "dog", "cat", "stop". The sequences can be arbitrarily long, but every sequence must end with the word "stop".

We want to fit a sequence model to this data, where there is a hidden state corresponding to each observed word. When we are given a sequence, we infer the hidden states according to the pre-trained model parameters. We use the final state (corresponding to the last word "stop") and use it to predict whether of not the sequence was GOOD or BAD. You can make this prediction by either examining the final state itself, or using the final state as input to a classifier that predicts GOOD or BAD.

A GOOD sequence is one in which the words "dog" and "cat" appear an equal number of times (including 0 times.) A BAD sequence is every other sequence.

For example:

`dog cat dog dog dog dog cat stop`

(BAD)

`cat cat cat cat cat stop cat cat cat cat stop`

(BAD)

`cat cat dog cat cat stop dog dog dog cat cat dog dog stop`

(GOOD)

You may use as much training data as you'd like, or manually set the model parameters yourself.

For each of the following sequence models, state as to whether there exist model parameters that can accurately capture the desired behavior. Explain why.

(a) Hidden Markov Model

(b) Conditional Random Field

(c) Recurrent Neural Network