Final Exam **Name:** _____
Apr 14, 2020 **Student ID:** _____
CMPT 727: 2020 Spring
Instructor: Maxwell Libbrecht

Instructions:
Take home exam. Electronic media limited to Wikipedia, published textbooks and calculator app. Any physical media is allowed.
Please answer the questions below. Show all your work.

Problem 1. (9 pts) Short answer

1. (3 pts) Order the following KL divergence values from smallest (most similar) to largest (most dissimilar).

   - $p(X) = N(0, 1)$.
   - $q(x) = \text{Uniform}(-1, 1)$.
   - $r(x) = (1/Z)\exp(x^2)$ where Z is chosen such that $\int r(x)dx = 1$.

   (a) $KL(p\|q)$
   (b) $KL(q\|p)$
   (c) $KL(p\|r)$

2. (3 pts) You are developing a model (θ) for diagnosing whether or not a patient has cancer (Y) based on a panel of tests (X). You have a training set ($\mathcal{D}$) of N patients for which you know both X and Y. (Unless stated otherwise, assume our prior distribution $P(\theta)$ is reasonable.) You are considering three potential strategies:

   (1) Use the training set to choose the maximum likelihood estimate $\theta^{MLE} = \text{argmax}_\theta P(\mathcal{D}|\theta)$ and diagnose new patients according $P(Y|X, \theta^{MLE})$.
   (2) Use the training set to choose the maximum a posteriori estimate $\theta^{MAP} = \text{argmax}_\theta P(\theta|\mathcal{D})$ and diagnose new patients according $P(Y|X, \theta^{MAP})$.
   (3) Diagnose new patients according to the posterior predictive distribution $P(Y|X, \mathcal{D}) = \int_\theta P(Y, \theta|X, \mathcal{D})$.

   For each of the following statements, list ALL strategies (1, 2 and/or 3) for which the statement is true, or "none" if all are false.

   (a) A poor choice of prior $P(\theta)$ may lead to unexpected poor results. _____
   (b) For very large N, a smart choice of prior could lead to significantly better results. _____
   (c) For small N, we might overfit, giving a confident incorrect answer on a new patient. _____

3. (3 pts) For each pair of models below, indicate which one allows for more efficient exact inference. Assume the parameters of the models are known. (Circle a or b.)

   i. (a) Naive Bayes-like model where the hidden variable depends on the observed variables. (b) Naive Bayes-like model where the observed variables depend on the hidden variable.

   ii. (a) Ising model. (b) Hidden Markov model.

   iii. Latent representation model with (a) 2 hidden variables with cardinality of 100 each, or (b) 100 hidden variables with cardinality of 2 each.

Problem 2. (16 pts) Classification with naive Bayes

Consider the Naive Bayes model with class variable $C = \{1, 2, 3\}$ and discrete binary evidence variables $\mathbf{X} = \{X1, X2, X3, X4\}$, where each $X_i \in \{-1, +1\}$. The CPDs for the model are parameterized by $P(C = c) = \pi_c$ and $P(X_i = +1|C = c) = \theta_{ic}$.

In addition, we assume that our training data consists of the following:

- $\mathbf{X^{(1)}} =< +1, +1, -1, -1 >$, Class = 1
- $\mathbf{X^{(2)}} =< -1, +1, -1, -1 >$, Class = 3
- $\mathbf{X^{(3)}} =< +1, -1, -1, -1 >$, Class = 1
- $\mathbf{X^{(4)}} =< +1, +1, +1, -1 >$, Class = 2
- $\mathbf{X^{(5)}} =< +1, +1, -1, +1 >$, Class = 3

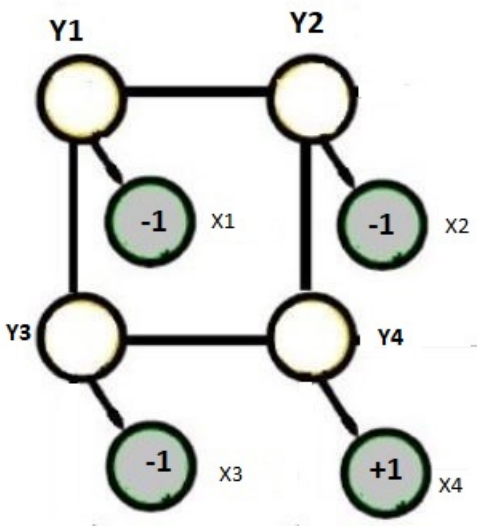
(a) (5 pts) Calculate the MLE for parameters π_1 and θ_{i1} for all $i \in \{1, .., 4\}$

(b) (8 pts) Suppose the data set is NOT labeled with classes. We can still estimate the parameters using the EM algorithm. We use $\boldsymbol{\theta^t}$ to represent all parameters at iteration t. Show the first iteration of EM with uniform initialization $\boldsymbol{\theta^0} = \{\pi_c^0 = \frac{1}{3}$ and $\theta_{ic}^0 = \frac{1}{2}.\}$. That is, calculate π_c^1 and θ_{ic}^1.

(c) (3 pts) Based on your results, why is using a uniform initialization for EM a bad idea?

Problem 3. (20 pts)

In this question, we will be considering a Ising Model with 4 hidden variables, as shown in the following figure.



Recall that in an Ising model, the variables can take on only the values $\{-1, +1\}$. Suppose we observe that $X_1 = X_2 = X_3 = -1$ and $X_4 = +1$.

(a) (8 pts) The factor over every adjacent pairs of Y_i's is: $\psi_{ij}(y_i, y_j) = \exp(y_i y_j)$, i.e., $\psi_{ij} = \begin{pmatrix} e & e^{-1} \\ e^{-1} & e \end{pmatrix}$ and the evidence factor is $\psi_i(y_i) = \exp(x_i y_i)$. Calculate the posterior belief on Y_4 after one iteration of loopy belief propagation. You do not need to normalize your final result.

(b) (8 pts) We'd like to try mean field inference this time. The initial μ_i's are as follows: $\mu_1 = \mu_2 = \mu_3 = -0.245$ and $\mu_4 = 0.245$. After one round of mean field inference, what are the updated values for μ_1 and μ_4? (Assume all values are updated in parallel, and that we use undamped updates.)

(c) (6 pts) Suppose we want to use the Metropolis Hastings Algorithm to sample from this Ising Model. We use a transition function that samples all hidden variables uniformly at random. Our initial sample is: $y_1 = y_2 = +1$ and $y_3 = y_4 = -1$.

Which of the following proposed transitions are guaranteed to be accepted? (Circle all that apply.)

(a) $y_1 = y_2 = y_4 = +1$ and $y_3 = -1$

(b) $y_1 = y_2 = y_3 = y_4 = -1$

(c) $y_1 = y_2 = y_3 = +1$ and $y_4 = -1$

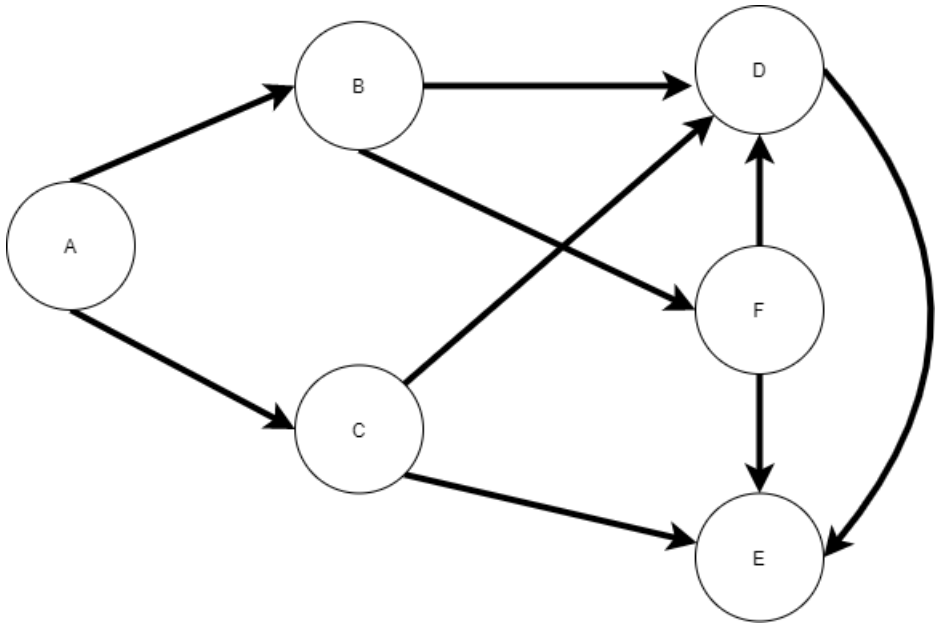
(d) $y_1 = +1$ and $y_2 = y_3 = y_4 = -1$

Problem 4. (8 pts) Given a probability distribution : $p(x) = (1/Z)x$ for $x \in [0, 1]$.

(a) (4 pts) Determine the value for Z.

(b) (4 pts) Show how to use the inverse CDF method to produce samples from $p(x)$. You can assume that you have access to the random() function, which produces values from Uniform(0,1).

Problem 5. (8 pts)



(a) (4 pts) Convert the BN above to MRF.

(b) (4 pts) Fill in the blanks to make a true statement: _____ and _____ are d-separated given _____ in the BN, but not d-separated in the MRF.