STUDENT NAME: _____

STUDENT ID: _____

MIDTERM EXAMINATION

Machine Learning - Winter 2016

March 29, 2016

You are allowed one double-sided "cheat sheet". No laptops, calculators or cell phones are allowed.

Read all the questions before you start working. Please write your answer on the provided exam (you can use both sides of each sheet). If you have a question, raise your hand. Partial credit will be given for incomplete or partially correct answers. Please be sure to define any new notation you introduce.

There are 3 multiple-part questions and 8 pages in the exam booklet. You can write on both sides of the sheet. The total number of points is 100.

Good Luck!

1. [10 points] **Kernels**

   Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$.

   (a) [5 points] Let K be defined as:

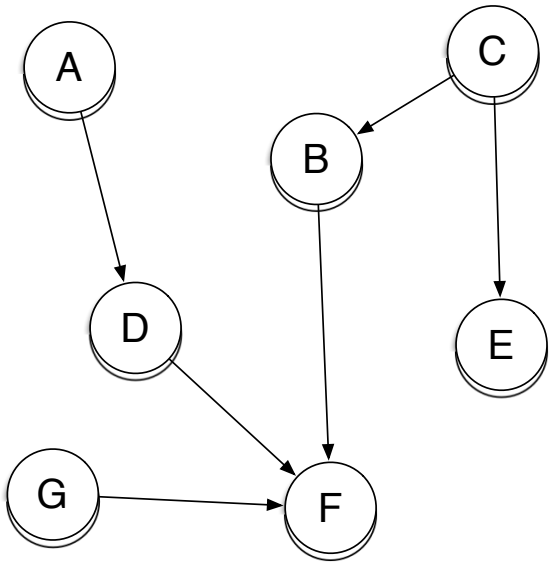
   $$K(\mathbf{x}, \mathbf{z}) = ||\mathbf{x}||_1 ||\mathbf{z}||_1$$

   In other words, K is the result of multiplying the L_1 norm of the two vectors. Prove that K is a kernel.

   (b) [5 points] Would you expect this kernel to be better or worse than a linear kernel ? Justify your answer. You may assume a particular algorithm if it helps you to think about it, but the answer should not depend on this.

2. [30 points] **Graphical models**

   All the questions below refer to the following graphical model:



   (a) [5 points] Write out the joint distribution defined by the model

   (b) [5 points] What is the Markov blanket of node D?

(c) [5 points] Suppose that node B has some evidence, and we want to compute the conditional probability at node A, using Gibbs sampling. Which nodes do *not* need to be sampled in the process? Justify your answer.

(d) [5 points] Suppose A is a continuous variable, whose distribution is parameterized by one Gaussian. Suppose D is a binary variable. Explain how you could parameterize node D.

(e) [5 points] Suppose now C is a latent variable that is never observed. Your boss would like you to remove it from the model, in order to avoid doing EM for learning. Explain how you can change the structure of the Bayes net, if you wanted to make sure no new independencies are introduced.

(f) [5 points] Suppose that you want to learn the structure of the graph from data. Is it possible to tell if the arrow from C to E should have the direction in the picture, or the opposite direction? Justify your answer.

3. [60 points] **Short questions**

   (a) [10 points] Suppose you are doing logistic regression on a classification task using an L_2 regularization penalty. Sketch the average cross-entropy training and testing error you would expect to see when you run the algorithm on a dataset, as well as the norm of the weight vector, as a function of the regularization parameter λ (make one graph for the error curves and another for the weight vector norm). No explanation is necessary.

   (b) [10 points] Suppose you used cross-validation to find a good value for λ, on a dataset of 500 examples with 100 features. But, you have now acquired more data, so instead of the original 500 examples, now you have 50000. Would the same value of λ work best, or would you expect a higher or lower value of λ to work better? Justify your answer.

(c) [10 points] Suppose you try both kernel-based logistic regression and support vector machines, with a linear kernel. Would you obtain the same classifier? Justify your answer.

(d) [10 points] Consider an input data set with 2 continuous features. Is it possible to learn a mixture model with 2 Gaussian components and one uniform component using EM? Justify your answer

(e) [10 points] Suppose you have 100 points in a 2-dimensional continuous space. You are considering fitting mixture models with 1 component, 2, 3, etc. Describe how you can choose the number of components.

(f) [10 points] Suppose you have the values of 10 tests performed on a patient on two different dates, roughly 1 month apart. You want to use this information in order to predict the patient's disease. You are considering two possibilities: (1) feeding the 20 values into a classifier; (2) feeding 10 values representing the difference between the first and second measurement. Explain in two sentences the advantages and disadvantages of each method.