STUDENT NAME: _____

STUDENT ID: _____

MIDTERM EXAMINATION

Machine Learning - Winter 2016 - Solutions

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.

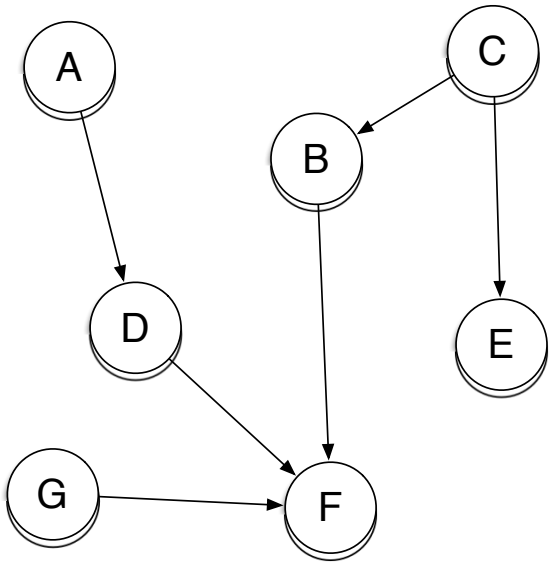
   Solution: Consider the function $\phi : \mathbb{R}^n \to \mathbb{R}, \phi(\mathbf{x}) = ||\mathbf{x}||_1$. It is immediate that $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ so K is a kernel (by using the definition).

   (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.

   Solution: This kernel loses a lot of information (all components of the vector get boiled down to the norm) so it will be much more biased than the linear kernel. It is unlikely that this kernel would ever perform better.

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

   Recall that each node has a conditional distribution of the node dependent on its parents, and the joint is the product of the factors at all nodes, so we have:

   $$P(A,B,C,D,E,F,G) = P(A)P(B|C)P(C)P(D|A)P(E|C)P(F|B,D,G)P(G)$$

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

Solution: The Markov blanket is the parents, children and spouses, so for node D it is $\{A,F,B,G\}$.

(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.

Solution: In general, in Gibbs sampling we sample all nodes, in some random fashion, each node being re-sampled from its Markov blanket. Having evidence at B means this node has some value $B = b$ that is known and fixed. This would influence the value of node F. If F is sampled and its value is given, information from it propagates to D and G, and if D is re-sampled, A will be re-sampled too. Hence, we need to sample $\{A,D,G,F\}$, nodes $\{C,E\}$ are separated by the known and fixed value of B, so they are not needed.

A different way to see this would be to write out

$$P(A|B=b) = \frac{P(A.B=b)}{P(B=b)} = \frac{P(A.B=b)}{\sum_a P(A=a,B=b)}$$

$$\begin{aligned}
P(A=a,B=b) &= \sum_{c,d,e,f,g} P(A=a,B=b,C=c,D=d,E=e,F=f,G=g) \\
&= \sum_{c,d,e,f,g} P(A=a)P(B=b|C=c)P(C=c)P(D=d|A=a)P(E=e|C=c) \\
&\quad P(F=f|B=b,D=d,G=g)P(G=g) \\
&= P(A=a) \sum_{f,d,g} P(F=f|B=b,D=d,G=g)P(D=d|A=a) \\
&\quad \sum_c P(C=c)P(B=b|C=c) \sum_e P(E=e|C=c)
\end{aligned}$$

Notice that the last sum is 1, and the previous sum (over c) is the same in all the terms, so can be factored out and will simplify in the conditional probability fraction. Hence, only the first terms need to be evaluated (in Gibbs, as well as any other inference method in fact)

(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.

Solution: As in assignment 2, you could make D a sigmoid function over the value of A: $P(D|A=a) = \sigma(w_1 a + w_0)$. For any given evidence $A = a$, we can then treat $P(D|A=a)$ as the probability of D being 1. Note that other solutions are possible as well.

(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.

Solution: E and B will have to be connected through an arc, as they will become dependent. The direction of the arc does not matter, asn the same conditional independencies will be preserved either way.

(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.

Solution: If we only have observational data, we will not be able to tell, as the two directions lead to the same conditional independencies. We would require the ability to *set* the values of C or E and see what happens to the other variable. In oter words, we need the ability to perform *interventions* on the system.

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.

Solution: The norm of the weight vector will continually decrease, tending to 0 as $\lambda \to \infty$. The testing error will be usually u-shaped, the training error will usually be increasing with λ as we put in more bias (recall the interpretation of the regularized as a prior). We would typically expect testing error to be higher than training error, though this is not always the case. As the amount of data increases to ∞, the testing and training error will tend to coincide and become best at $\lambda = 0$.

(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.

Solution: No, we would expect a lower λ to be needed. This is because λ gives a bias-variance trade-off. With a lot of data, variance poses less of a problem, so the strength of the regularizer can be allowed to diminish (which will also allow less bias).

(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.

Solution: No, they optimize a different objective (cross-entropy for logistic regression, margin for SVM). Note also that if we introduce regularization in an SVM, only some of the points in the training set (the support vectors) will matter to the decision boundary, whereas for logistic regression, the decision boundary always depends on all points in the training set.

(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

Solution: Yes, EM is a general framework that only requires having the ability to compute the components through maximum likelihood. It is immediate to see how to set up hard EM in this case - one would start with some assignment of data to the 3 components, then use exactly the formulas discussed in class for mixture of Gaussians to compute new parameters

for the parent node and the Gaussian components. For the uniform distribution, the minimum and maximum of the points assigned to the component give the ML parameter estimates.

(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.

Solution: Cross-validation (using data log-likelihood measured on a subset of the data).

(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.