CS 760
Machine Learning
Fall 2014 Exam

Name ____

Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, make sure your exam has every page (numbered **1** through **8**).

Problem	**Score**	**Max Score**
1.	____________	20
2.	____________	20
3.	____________	15
4.	____________	15
5.	____________	10
6.	____________	20
Total		100

1. Decision tree learning (20 points):.

(a) Suppose we are learning a decision tree using information gain with a lookahead search to choose splits. Using the following training set, show how this procedure would calculate the value of splitting on feature *A* at the root of the tree.

	A	*B*	*C*	*Class*
Instance 1	T	T	F	neg
Instance 2	F	F	T	neg
Instance 3	T	F	F	neg
Instance 4	F	F	F	pos
Instance 5	F	T	F	pos
Instance 6	T	T	F	pos
Instance 7	F	T	T	neg
Instance 8	T	F	T	neg
Instance 9	F	T	F	pos
Instance 10	T	T	T	pos

(b) Notice that in the data set used above, instances 1 and 6 have the same feature vector, but different class labels. Briefly describe two reasons why this might be the case.

2. Neural networks, SVMs, and inductive bias (20 points):

(a) Suppose we are training a neural network with one linear output unit (i.e. its output is the same as its net input) and no hidden units for a binary classification task. Instead of using the squared-error function we considered in class, we want to use the following error function (which is known as cross-entropy error): $E(\mathbf{w}) = -\left[y \log o + (1-y)\log(1-o)\right]$, where y is the target value (0 or 1) and o is the output produced by our current network. What is the update rule we should use for adjusting our weights during learning with this error function?

Hint: the derivative of $\log(a)$ is $\frac{1}{a}$

(b) Describe one specific way in which we could vary the neural net approach described in (a) in order to change its hypothesis space bias?

(c) Describe one specific way in which we could vary the neural net approach described in (a) in order to change its preference bias?

(d) Shown below is a standard primal formulation of the SVM learning task. Briefly explain what tradeoff the parameter *C* controls?

$$\underset{\boldsymbol{w}, b, \xi^{(1)} \ldots \xi^{(m)}}{\text{minimize}} \quad \frac{1}{2} \| \boldsymbol{w} \|_2^2 + C \sum_{i=1}^{m} \xi^{(i)}$$

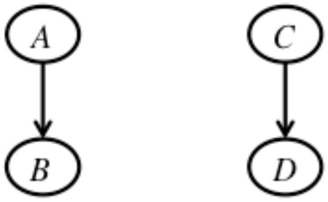
$$\begin{aligned} \text{subject to constraints:} \quad & y^{(i)}(\boldsymbol{w}^\mathsf{T} \boldsymbol{x}^{(i)} + b) \geq 1 - \xi^{(i)} \\ & \xi^{(i)} \geq 0 \\ \text{for } i = 1, \ldots, m & \end{aligned}$$

(e) Briefly describe an experimental methodology for choosing a good value for *C.*

3. Bayesian network parameter learning (15 points): Suppose you are given the training set shown below on left, and the Bayes net structure shown on the right, along with initial parameters for the part of the network that characterizes the variables *C* and *D*. Show the estimated parameters for the entire network when using Laplace estimates and only one iteration of the EM algorithm (where appropriate). The symbol '?' is used to indicate a value that is missing.

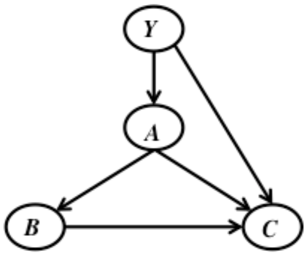
A	*B*	*C*	*D*
t	t	t	f
t	f	t	f
f	f	?	t
f	f	f	t

P(C=t)
0.4



C	*P(D=t)*
t	0.1
f	0.8

4. Bayesian network structure learning (15 points): Suppose you are applying Bayes net learning methods to a problem in which the true (but unknown) dependency structure is as shown below. For each of the Bayes net learning methods listed, show the structure the method could potentially learn that is most similar to the true structure (i.e. that has the smallest number of edge additions/deletions/reversals relative to the true network).



(a) Sparse Candidate algorithm with *k*=2:

(b) TAN (assume that variable *Y* is specified as the class variable):

(c) Naïve Bayes (assume that variable *Y* is specified as the class variable):

5. Learning Theory (10 points):

(a) Consider the concept class *C,* in which each concept is represented by a pair of circles centered at the origin, (0, 0). Let *r* be the radius of the inner circle and *r+a* be the radius of the outer circle (*a* is a positive number). Each training instance is represented by two real-valued features X_1 and X_2, and a binary class label $Y \in \{0, 1\}$. The concept predicts Y=1 for instances that are outside the radius of the inner circle and inside the radius of the outer circle, and Y=0 otherwise. Show that *C* is PAC learnable.

6. Short Answer (20 points): Briefly define each of the following terms.

overfitting

probability estimation trees

deep belief networks

slack variables

values that are missing at random

mistake-bound model

bagging