CS 760
Machine Learning
Fall 2013 Exam

Name _____

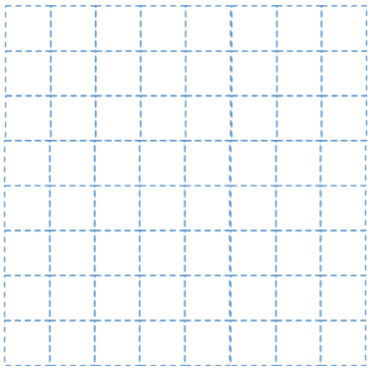
Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, make sure your exam has every page (numbered **1** through **10**).

Problem	**Score**	**Max Score**
1.	_____	20
2.	_____	20
3.	_____	25
4.	_____	15
5.	_____	20
Total		100

1. *k*-Nearest Neighbor and Decision Tree Learning (20 points):.

(a) Suppose we want to learn a *k*-NN model with the following data set, and we are using leave-one-out cross validation (LOOCV) to select *k*. Would LOOCV pick *k*=1 or *k*=3 if we were using Manhattan distance? Show your work.

	A	*B*	*Class*
Instance 1	2	3	pos
Instance 2	4	4	pos
Instance 3	4	5	neg
Instance 4	6	3	pos
Instance 5	8	3	neg
Instance 6	8	4	neg



(b) Show a decision tree (with ID3-like splits) that would perfectly classify this training set.

(c) Suppose this data set also included several irrelevant features. Would you expect *k*-NN or ID3 to be more resistant to these features? Justify your answer.

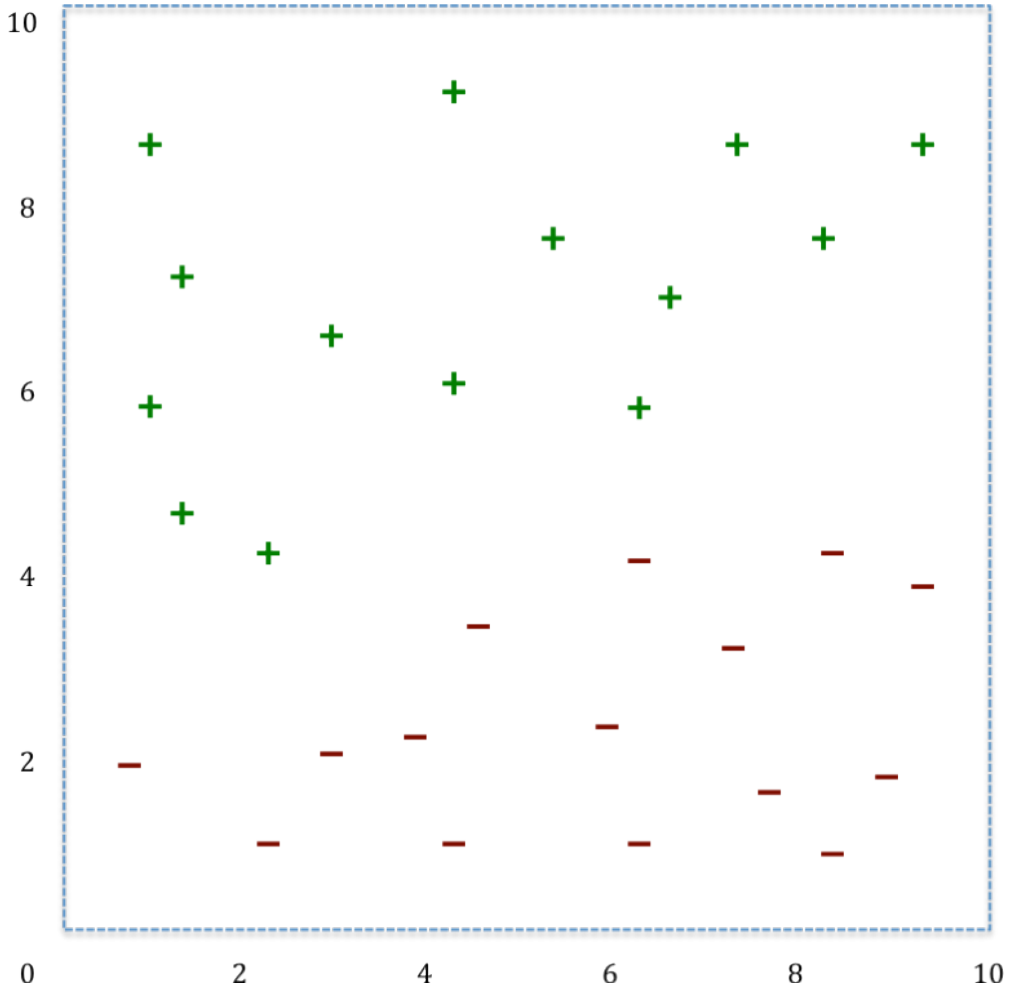
(d) Consider the following training set with two Boolean features and one continuous feature. How much information about the class is gained by knowing whether or not the value of feature *C* is less than 400?

	A	*B*	*C*	*Class*
Instance 1	F	T	115	neg
Instance 2	T	F	890	neg
Instance 3	T	T	257	pos
Instance 4	F	F	509	pos
Instance 5	T	T	733	pos

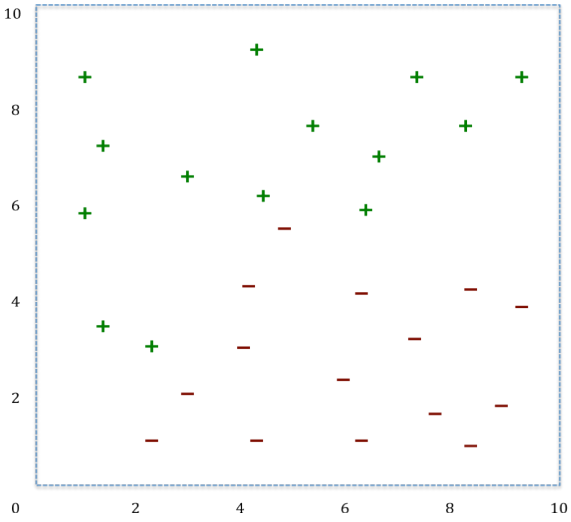
(e) How much information about the class is gained by knowing whether or not features *A* and *B* have different values?

2. Neural networks and Support Vector Machines (20 points):

(a) Suppose we are applying a hard-margin linear SVM to the two-class data set illustrated in the figure below. On the figure, draw the margin and indicate which instances are the support vectors.



(b) The data set below cannot be handled by a hard-margin, linear SVM. Briefly describe how you could modify <u>two</u> different aspects of this SVM to make it better suited to this training set, and explain why each is appropriate.



(c) 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): $E(\boldsymbol{w}) = -\left[y \log o + (1-y)\log(1-o)\right]$, 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 $\frac{1}{a}$

(d) Suppose we want to use backward elimination to select which features will be included in our final neural network. Assume that we have n features, only r of which are relevant, and the backward elimination search does not consider feature sets with fewer than r features. If we are using k-fold cross validation to evaluate each feature set, how many neural networks will be trained in the process of finding a feature subset of size r?

3. Bayesian networks (25 points): Consider the following training set with two Boolean features, and one 3-valued feature, *C*, that has possible values {red, blue, green}.

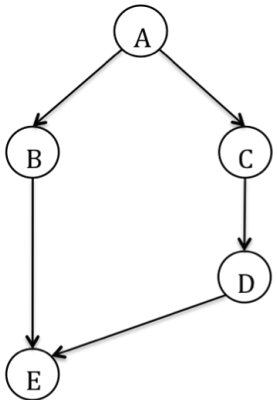
	A	*B*	*C*	*Class*
Instance 1	F	T	red	neg
Instance 2	T	F	blue	neg
Instance 3	T	T	red	pos
Instance 4	F	F	blue	pos
Instance 5	T	T	green	pos

(a) Draw the structure of a naïve Bayes network for this task.

(b) Using *m*-estimates with $m = 6$ and uniform priors for each variable, estimate the parameters in this naïve Bayes network.

(c) Consider the Bayes net that would result from reversing every edge in the naïve Bayes model in part **(a)**. Assume that we are using tables to represent conditional probability distributions. How many parameters are in this model? Explain your answer.

(d) Assume that we are doing a hill-climbing structure search using the *add-edge*, *delete-edge* and *reverse-edge* operators. List or show all of the operator applications that we could consider for the current network shown below.



(e) Now assume that we are using the Sparse Candidate algorithm for our structure search, and the candidate parents for each node are those listed below. List or show all of the operator applications that we could consider for the first change in the maximize step that starts with the network from part **(d)**.

node	candidate parents
A	{B, E}
B	{A, C}
C	{A, D}
D	{C, E}
E	{B, D}

4. Learning Theory (15 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.

(b) Briefly describe the relationship between the Halving algorithm and the Weighted Majority algorithm?

(c) Whereas the Halving algorithm is only of theoretical interest, the Weighted Majority algorithm can be applied in practice to real problems. Briefly describe <u>two</u> limitations of the Halving algorithm that make it impractical, and describe these limitations are, or can be addressed in Weighted Majority.

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

error-correcting output codes

inductive bias

confusion matrix

autoencoder

values that are missing systematically

single-link clustering

internal cross validation