CS 691 Midterm 1 Name ______________________________

Short Problems (20 points)

1. As we increase k, the training error of the K-NN classifier always increases. True or false? Explain. (5 points)

2. If u and v are any two orthogonal unit vectors, then $||u + v||_2 = 1$. Orthogonal=Perpendicular. Unit=Length is 1. True or false. If true, prove it. If false give a counter example. (5 points)

3. If the training data is linearly separable, then the 3-nearest neighbors algorithm will always have 100% accuracy on the training set. True or False. If true, explain how it is true. If false, give a counter-example. (5 points)

4. The decision tree classifier has 100% accuracy on the training set (namely, the data is noise-free). Will a linear classifier necessarily have the same accuracy (100%) on the training set? Explain your answer. (5 points)

691

Decision Trees (20 points)

5. You are given N training samples $S = \{s_1, s_2, \ldots, s_N\}$, (the size of S is N aka $|S| = N$). Each sample s_i in S has D features, $s_i = (x_1, x_2, \ldots, x_D)$, and a binary label $y_i = \{0, 1\}$. Let the set of unique feature vectors in S be F, with $|F| \leq |S|$. For each unique feature vector f_j in F, there are n_j samples in S with that same feature vector. Of these n_j samples, there are k_j $(0 \leq k_j \leq n_j)$ samples with the label 1.

   Give an expression in terms of these variables – *do not use specific values* – for the best accuracy achievable on the training data using a Decision Tree of any depth.

691

Linear Classifier (10 points)

6. Suppose we have the following training data. Give the weights w_1, w_2, and b for a neuron that perfectly classifies the training data.

Sample	x_1	x_2	Label
s_1	0	0	-1
s_2	1	0	-1
s_3	0	1	-1
s_4	2	0	1
s_5	1	1	1
s_6	0	2	1

691

Feature Expansion (10 points)

7. The 2D XOR problem is not linearly separable. Feature expansion can be used to map the 2D XOR problem to a space in which it is linearly separable. That is, given some new features that are a combination of x_1 and x_2, we can find a linear separator between samples from class -1 and samples from class 1. The 2D XOR data is given below.

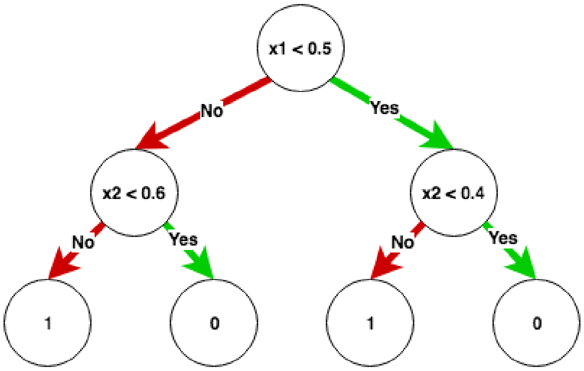
Sample	x_1	x_2	Label
s_1	-1	-1	0
s_2	-1	1	1
s_3	1	-1	1
s_4	1	1	0

Use feature expansion to map the problem to a space in which it is linearly separable. Give the expression for each new feature in terms of x_1 and x_2. Specify which features you will use in your final classification problem.

691

Decision Boundaries (20 points)

8. Consider the following Decision Tree.



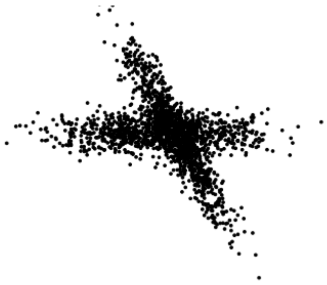
(a) Draw the decision boundary for this tree, labeling all areas with the correct class. (10 points)

(b) Suppose we perform 1-nearest neighbor classification, instead of using the decision tree given above. The training data has four samples from each class. Is it possible that we obtain the same decision boundaries for the 1-NN classifier that we got for the decision tree in part a? If yes, give an example of the location that the points could have. If no, explain why. (10 points)

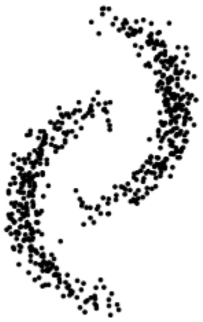
691

K-Means (10 points)

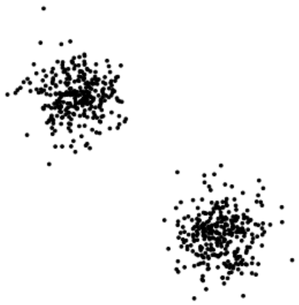
9. Given the three following sets of data (i, ii, and iii). Assume you want to cluster each set of data into two clusters. Explain, and draw, what would likely happen with K-Means (K=2) in each case and why.



i



ii



iii

691

Optimization (10 points)

10. Recall our Regularized Optimization problem to find a linear separator given non-linearly separable data. We are trying to find the w and b that minimize the following objective function.

$$\underset{w,b}{\text{minimize}} \quad 1[y(w \bullet x + b) \leq 0] + \lambda R(w, b)$$

(a) What are the two terms in the above equation doing? Explain them individually. That is, explain what $1[y(w \bullet x + b) \leq 0]$ is doing and explain what $\lambda R(w, b)$ is doing. (5 points)

(b) Assuming R does the "right thing," what value(s) of λ will lead to overfitting? What value(s) will lead to underfitting? (5 points)

691