CS 491 Midterm 1 Name Solutions

Short Problems (20 points)

1. Given the following categorical data $\{Noun, Verb\}$, how could you adapt the categorical data such that you could use K-means clustering on it? Give what K value you should use. (4 points)

$\begin{Bmatrix} \text{is noun?} \\ \text{is verb?} \end{Bmatrix}$ cat $\rightarrow \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}$ run $\rightarrow \begin{Bmatrix} 0 \\ 1 \end{Bmatrix}$

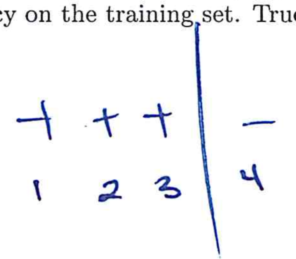
$K = 2.$ now each sample is a point in 2D, so we can do k-means.

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. (4 points)

$\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\sqrt{1^2+1^2} = \sqrt{2} \neq 1$

$u \cdot v = 0$ $\perp$

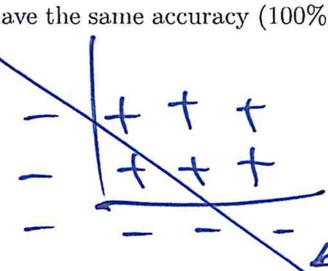
3. If the training data is linearly separable, then the 3-nearest neighbors algorithm will always have 100% accuracy on the training set. True or False. Explain your answer. (4 points)



pt. 4 would be classified as +.

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

no.



decision tree decision boundary

linear boundary.

5. Compute the gradient of the following function at $(1, 1, 1)$: $f(x, y, z) = 2x^2+3y^3+4x^4+xyz+3x^2y+4y^2z$ (4 points)

$\frac{\partial f}{\partial x} = 4x+16x^3+yz+6xy$ $\frac{\partial f}{\partial y} = 9y^2+xz+3x^2+8yz$

$\frac{\partial f}{\partial z} = xy+4y^2$ @ $(1,1,1) = (4+16+1+6, 9+1+3+8, 1+4)$

$(27, 21, 5)$ ✓

Decision Trees (20 points)

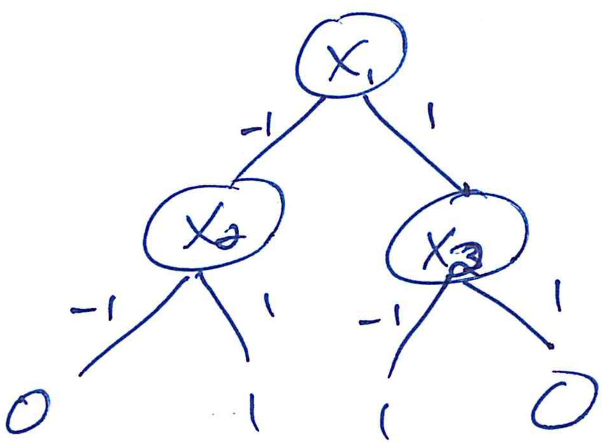
6. Use the following data for the 2D XOR problem.

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

(a) Does it make sense to generate a depth-1 decision tree for 2D XOR? Why or why not? (5 points)

no. accuracy is 50%. if I ask about x_1 or x_2 and if I use depth-0 and just always say yes.

(b) Generate the best depth-2 decision tree for the 2D XOR problem. (10 points)



(c) How would your decision tree from part (b) classify the following sample? (5 points)

Sample	x_1	x_2	x_3	Label
s_t	-1	-1	1	0

↑ ignore.

0

Optimization (15 points)

7. 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.

$$\underset{w,b}{\text{minimize}} \quad \underbrace{1[y(w \bullet x + b) \leq 0]}_{1} + \underbrace{\lambda R(w, b)}_{2}$$

(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. (10 points)

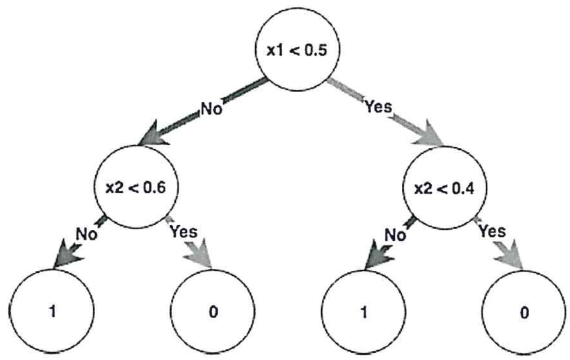
1 is ensuring that we have low training error.
2 is ensuring that the solution is simple.

(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)

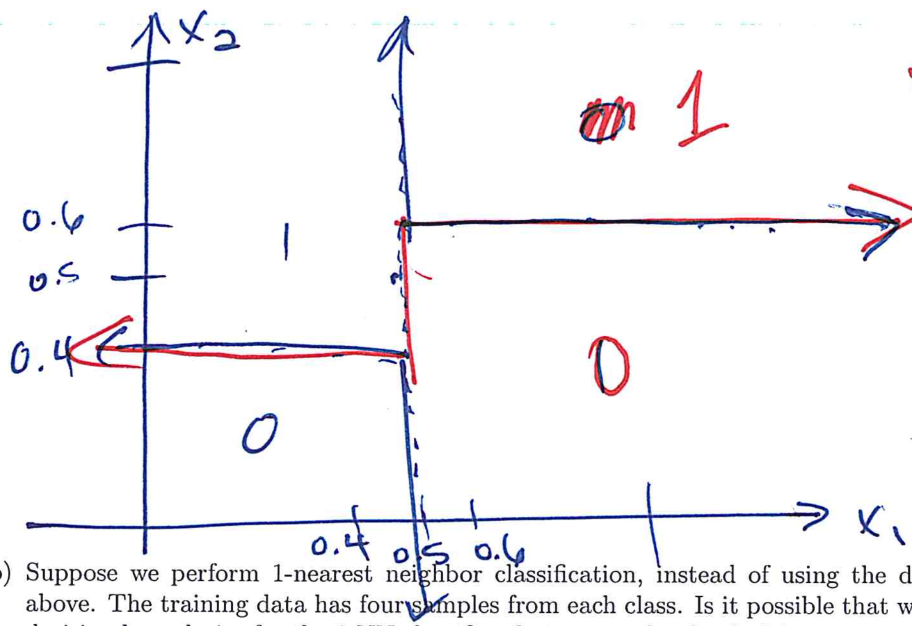
Large λ gives underfitting
Small λ gives overfitting

Decision Boundaries (15 points)

8. Consider the following Decision Tree.



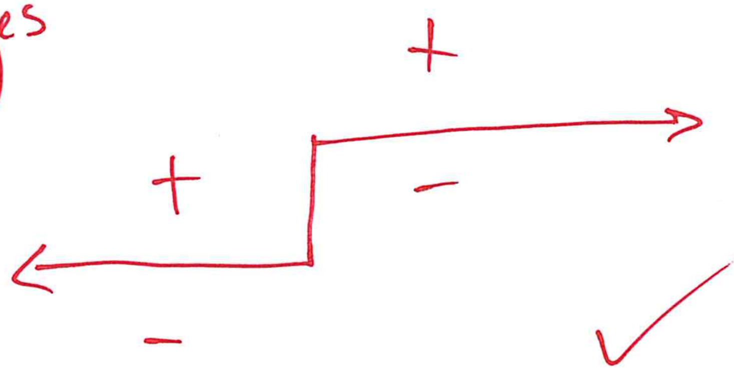
(a) Draw the decision boundary for this tree, labeling all areas with the correct class. (10 points)



red is final boundary.

(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. (5 points)

yes



Linear Classifier (20 points)

9. Suppose we have the following 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

(a) Give the weights w_1, w_2, and b for a neuron that perfectly classifies the training data. (10 points)

$a_1 = 0+0+0 = 0 \quad w_1 = 0 \quad w_2 = 0 \quad b = -1$

$a_2 = -1$ ✓

$a_3 = -1$ ✓

$a_4 = -1 \quad w_1 = 2 \quad w_2 = 0 \quad b = 0$

$a_5 = 2$ ✓

$a_6 = 0 \quad w_1 = 2 \quad w_2 = 2 \quad b = 1$

$a_1 = 1 \quad w_1 = 2 \quad w_2 = 2 \quad b = 0$

$a_2 = 2 \quad w_1 = 1 \quad w_2 = 2 \quad b = -1$

$a_3 = 2 \quad w_1 = 1 \quad w_2 = 1 \quad b = -2.$

$a_4 = 0 \quad w_1 = 3 \quad w_2 = 1 \quad b = -1$

$a_5 = 3$ ✓

$a_6 = 0 \quad w_1 = 3 \quad w_2 = 3 \quad b = 0$

$a_1 = 0 \quad w_1 = 3 \quad w_2 = 3 \quad b = -1$

$a_2 = 2 \quad w_1 = 2 \quad w_2 = 3 \quad b = -2$

$a_3 = 1 \quad w_1 = 2 \quad w_2 = 2 \quad b = -3$

$a_4 = 1$ ✓ $\quad a_1 = -3$

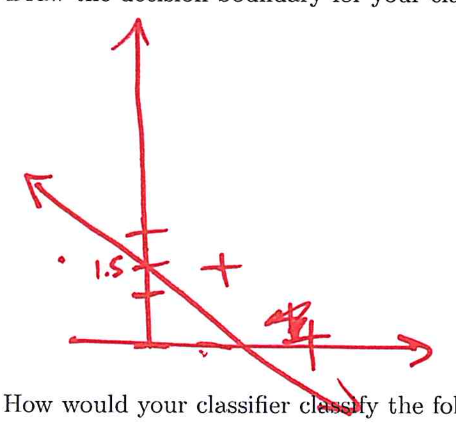
$a_5 = 1 \quad a_2 = -1$

$a_6 = 1 \quad a_3 = -1$ ✓

$2x_1 + 2x_2 - 3 = 0$

$x_2 = -x_1 + 1\frac{1}{2}$

(b) Draw the decision boundary for your classifier. (5 points)



(c) How would your classifier classify the following test sample? $s_t = (1.5, 1, 1)$ (5 points)

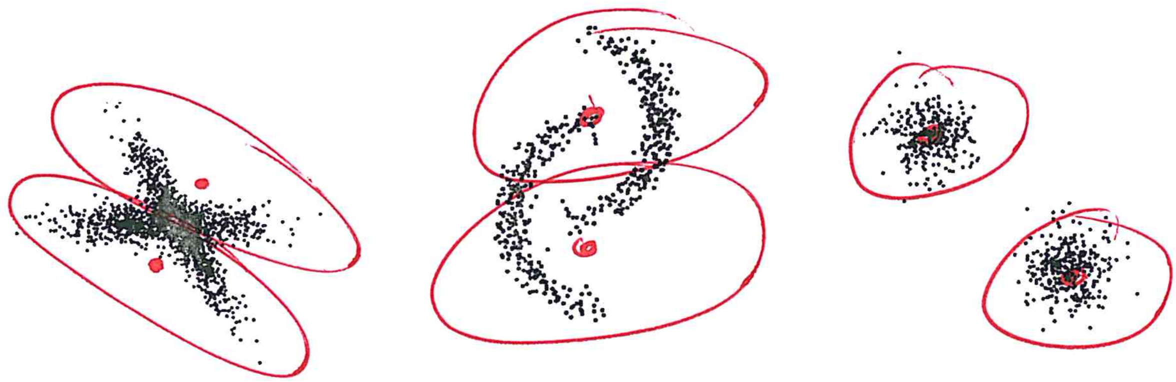
$2 \cdot 1.5 + 2 \cdot 1 + -3 =$

$3 + 2 - 3 = 2$

1

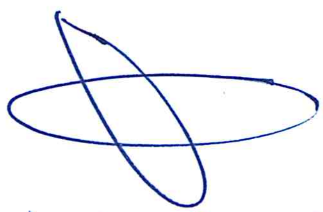
K-Means (10 points)

10. 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.

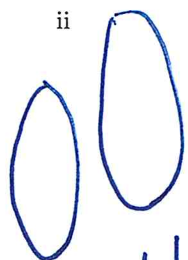


i ii iii

We expect



but k-means can't handle two clusters with same mean.



want this but k-means clusters based off of distances to mean.

this is perfect for kmeans

CS 691 Midterm 1   Name Solution

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)

Accepted multiple answers for this depending on the justification

K=1   + + + +   − − − −   K=3

error=0   error=0. doesn't always increase.

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)

Same as #2 on 491

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)

Same as #3 on 491

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)

Same as #4 on 491

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.

<u>Conflicting Features & Labels</u>

how many can I get right for a particular unique feature vector f_j ?

$\max(k_j, n-k_j)$

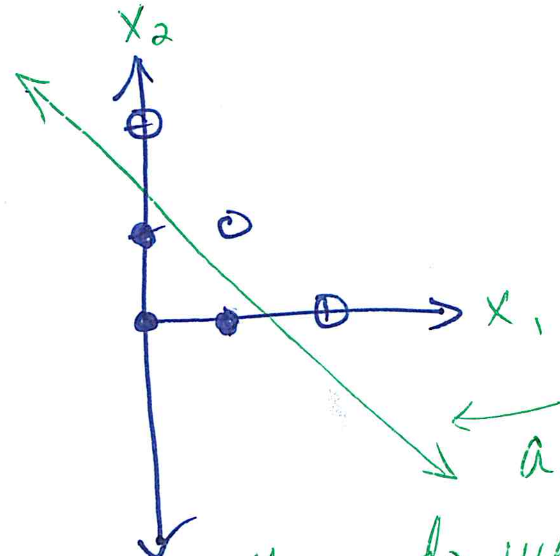
$$\frac{1}{N} \sum_j \max(k_j, n-k_j)$$

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

another way to solve #9 ~~@~~ 491:



this is a nice separator!

equation for line:

$X_2 = \underbrace{-}_{\text{slope}}X_1 + \underbrace{1.5}_{\text{y-intercept}}$

How do we get it in the form we want?

Let's get it in $w_1x_1 + w_2x_2 + b = 0$.

$X_1 + X_2 - 1.5 = 0$ $\quad w = [1, 1] \quad b = [-1.5]$

$w_1 \quad w_2 \quad b$

Feature Expansion (10 points)

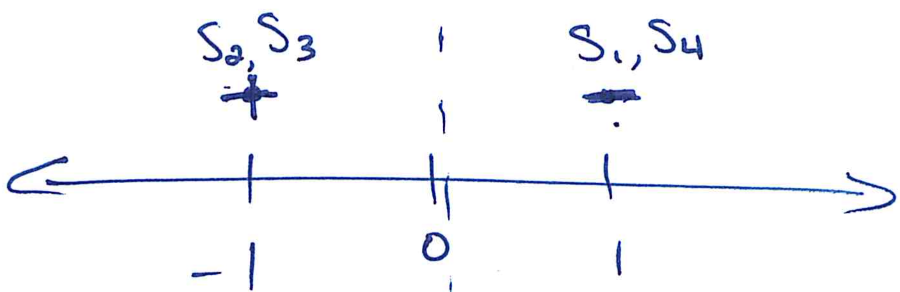
7. The 2D XOR problem is not linearly separable. Feature expansion can be used to map the 2D XOR problem to a space in which it is linearly separable. That is, given some new features that are a combination of x_1 and x_2, we can find a linear separator between samples from class -1 and samples from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

$X_3 = X_1 \cdot X_2$

Only use X_3
problem becomes 1D

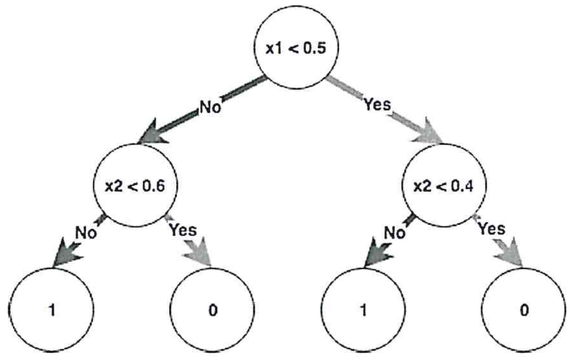


done ✓

many possible answers for this.

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)

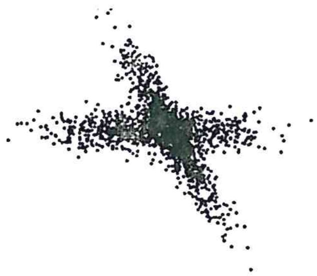
Same as problem #8

491

(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)

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

Same as problem #10

491

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)

Same as problem #7
491

(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)