CS 491 Midterm 1 Name Solution

Short Problems (20 points)

1. For every training data set $(x_1, y_1), \ldots, (x_n, y_n)$, the training error of the 1-nearest neighbor classifier is always zero. You may assume that each x_i is unique. True or False. If true, explain. If false, give a counter example. (4 points)

Every point in the training data is its own neighbor. So it will be classified with 100% accuracy using 1NN.

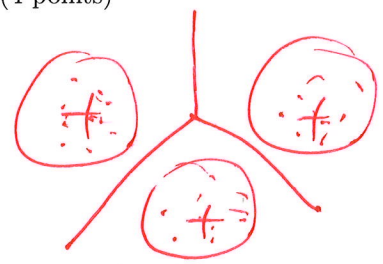
2. When building a decision tree, we never ask about the same feature twice. True or False. Circle one and explain. (4 points)

If our features are non-binary we can ask about them multiple times.

3. I have trained a perceptron on a set of data and achieved 100% accuracy on the training data. Will a decision tree necessarily achieve the same accuracy? Explain your answer. (4 points)

Yes. If a perceptron gets 100% then the data must be linearly separable. So, a DT can get 100% accuracy!

4. K-Means always produces a linear decision boundary. True or False. If true, explain. If false, give a counter example. (4 points)



K=3

5. What purpose do surrogate loss functions serve in gradient descent? (4 points)

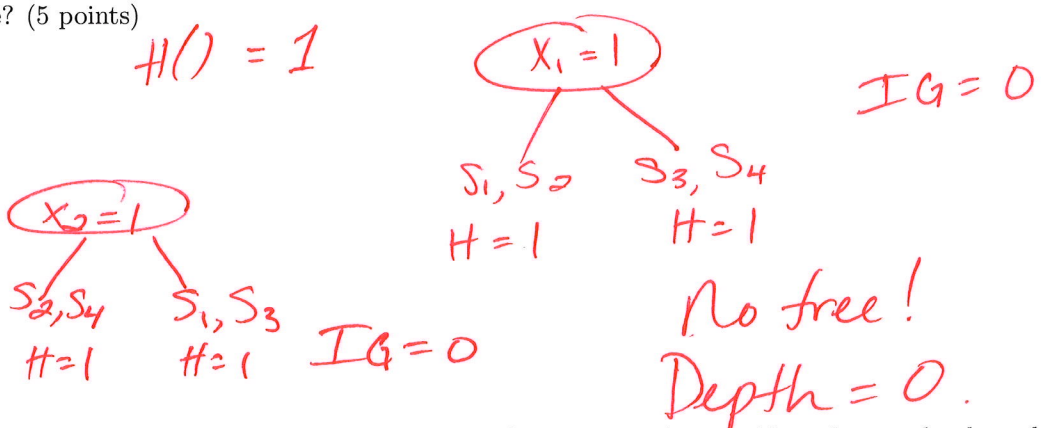
Upper bound on 0/1 loss. Allows us to take derivatives.

Decision Trees (15 points)

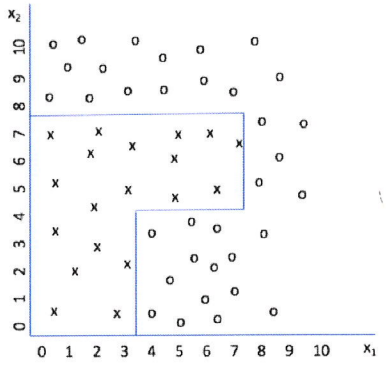
6. Use the following data.

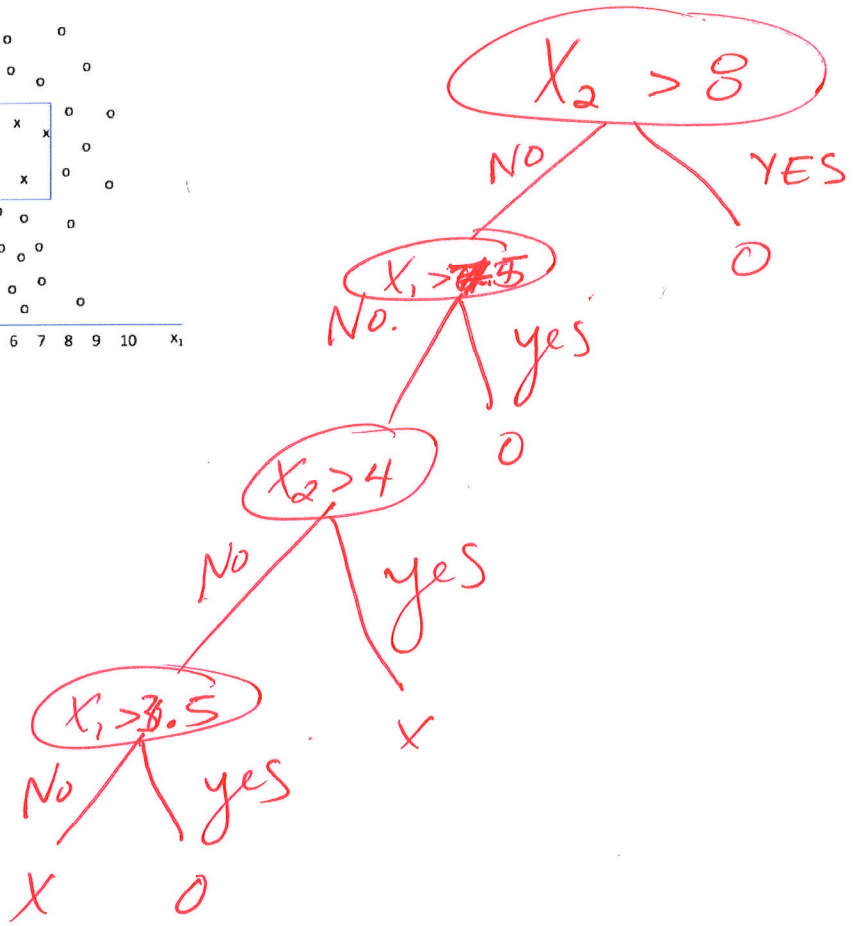
Sample	x_1	x_2	Label
s_1	2	2	0
s_2	2	1	1
s_3	1	2	1
s_4	1	1	0

(a) Using entropy and information gain, at what depth would your algorithm stop building the decision tree? (5 points)



(b) Given the following decision boundary, draw a decision tree that would produce such a boundary. (10 points)





Optimization (20 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.

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

(a) Let us use the following loss function: $L(y, \hat{y}) = (y - \hat{y})^2$ and the following regularization: $R = ||w||^2$. Find $\nabla_w L$ and $\frac{\partial L}{\partial b}$

w/out R: $\nabla_w L = -2(y - (w \cdot x + b)) \cdot x$

$$\frac{\partial L}{\partial b} = -2(y - \hat{y})$$

$\nabla_w L = -2(y - (w \cdot x + b)) \cdot x + 2\lambda w$

(b) Explain how $R(w, b) = \sum_i w_i \neq 0$ enforces a "simpler" solution? (5 points)

This minimizes the # of non-zero weights. So it minimizes the # of features used. This is good because it eliminates noise.

(c) What purpose does λ serve in this optimization problem? (5 points)

λ weights the regularization

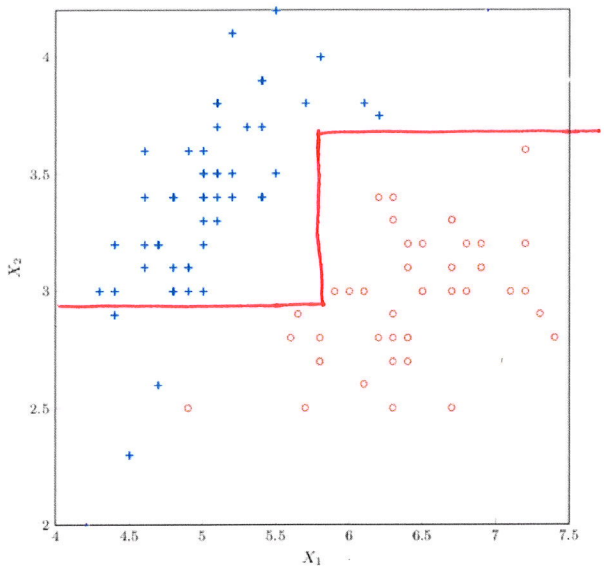
high λ gives simple soln.

Low λ gives exact soln.

prevents overfitting

Decision Boundaries (15 points)

8. Consider the following data.



(a) Explain what the decision boundary looks like for K-NN with K=N? There are 41 + and 38 *o*. (5 points)

K=N means we would always predict +. There is no boundary only a + prediction.

(b) Would a K-NN with K=3 perfectly classify the training data? Explain. (5 points)

No bottom left 3 points. + + o ← classified as +.

(c) Could a depth-2 decision tree perfectly classify the above data? Draw the best (with respect to accuracy) depth-2 decision tree. (5 points)

Shown above.

No. 3 mistakes

Linear Classifier (20 points)

9. Suppose we have the following training data.

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

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

$\vec{w}=\vec{0}$ $b=0$ ① $w=(1,1)$ $b=-1$
② no update
③ no update
④ no update
① no update!

$\vec{w}=(1,1)$ $b=-1$

(b) Would you get the same weights and bias if you iterated from s_4 up through s_1? (5 points)

④ $w=(1,1)$ $b=1$
③ $w=(1,2)$ $b=0$
② no update
① no update
④ no update
③ ~~$w=(1,1)$ $b=1$~~

~~②~~ ~~①~~ ~~$b=-1$~~
② no update
① no update
④ no update
③ "
② "
① " "

no.

$w=(1,2)$
$b=0$.

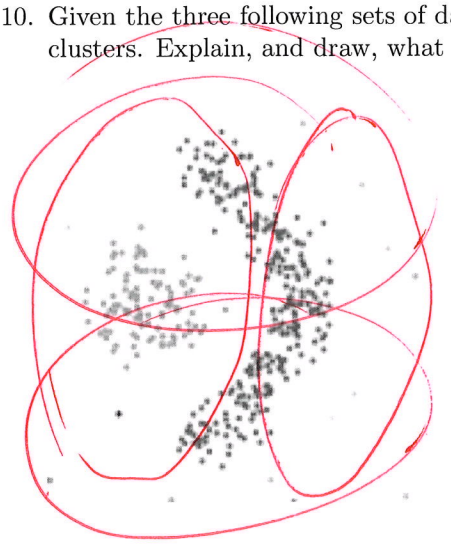
(c) How would your classifier from (a) classify the following test sample? $s_t = (1.5, 1, 1)$ (5 points)

$(1,1)\cdot(1.5,1)-1$

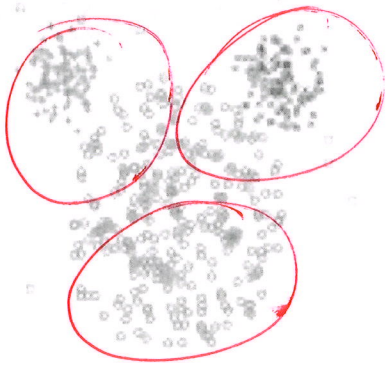
$1.5+1-1=1.5=(+)$

K-Means (10 points)

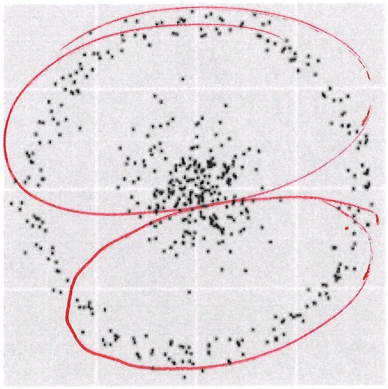
10. Given the three following sets of data (i, ii, and iii). Assume you want to cluster each set of data into clusters. Explain, and draw, what would likely happen with K-Means in each case and why.



(i) K=2



(ii) K=3



(iii) K=2

or

i) we can't get that structure ● because the cluster centers would move samples from the) to the ●.

ii) we can't get ○○ because more points would move from the large circle to the small ones, shifting the mean.

iii) we can't get (○) because the centers would be the same and K-means cannot handle this situation.

CS 691 Midterm 1 Name Solution.

Short Problems (20 points)

1. We can only have a tie in the case of KNN when K is even. True or false. If true explain. If false, give a counter example. (4 points)

K=1

- ? +

equidistant training pts.

2. We are trying to apply machine learning to a particular problem. We have a labeled dataset consisting of 10 million samples, each containing 10 real-valued features. What would be the best algorithm/model for this problem? What would be the worst? Explain. (4 points)

Best: several, Perceptron (online algorithm) 10 weights at a time. look @ 1 sample

Worst: K-NN hold all data in memory.

3. I have trained a perceptron on a set of data and achieved 100% accuracy on the training data. Will a decision tree necessarily achieve the same accuracy? Explain your answer. (4 points)

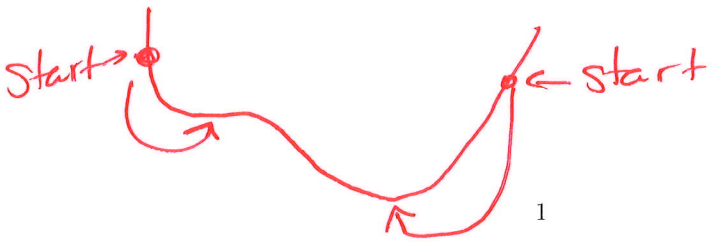
Same as 491 #3

4. K-Means always produces a linear decision boundary. True or False. If true, explain. If false, give a counter example. (4 points)

Same as 491 #4

5. Gradient descent is guaranteed to give you a global minimum. True or False. If true, explain. If false, give a counter example. (4 points)

local.



dependent on starting point

Decision Trees (15 points)

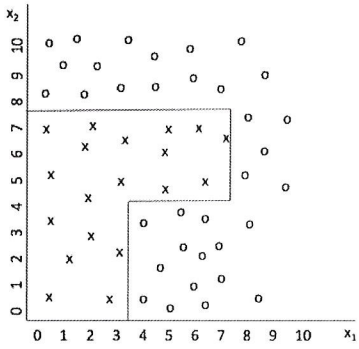
6. Use the following data.

Sample	x_1	x_2	Label
s_1	2	2	0
s_2	2	1	1
s_3	1	2	1
s_4	1	1	0

(a) Using entropy and information gain, at what depth would your algorithm stop building the decision tree? (5 points)

Same as 491 #6

(b) Given the following decision boundary, draw a decision tree that would produce such a boundary. (10 points)



Optimization (20 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.

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

(a) Let us use the following loss function: $L(y, \hat{y}) = (y - \hat{y})^2$ and the following regularization: $R = ||w||^2$. Compute the $\nabla_w L$ and $\frac{\partial L}{\partial b}$. (10 points)

Same as 491 #7a.

(b) Assume we have two nearby samples: $s_1 = [x_1, x_2, ..., x_d]$ and $s_2 = [x_1+\epsilon, x_2, ..., x_d]$. s_2 is the same as s_1 except that the first feature is off by a small number ϵ. Show how adding $R(w, b) = ||w||^2$ to our optimization problem helps to ensure that neighboring samples have similar predictions. (10 points)

$a_1 = s_1 \cdot w$ $\quad a_2 = s_2 \cdot w$

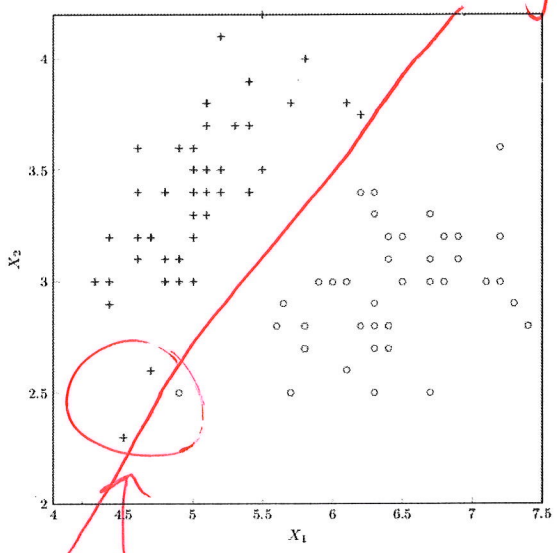
$= x_1 w_1 + \cdots$ $\quad = (x_1 + \epsilon) w_1 + \cdots$

$a_2 = a_1 + \epsilon w_1$

If w_1 is large, the activation will be significantly different for nearby points. Minimizing $||w||^2$ ensures neighbors have similar activations.

Decision Boundaries (15 points)

8. Consider the following data.



ugly line, but you get it.

(a) Could K-NN perfectly separate the data? If so, for what values of K? If not, explain. (5 points)

K=1 only.
K=3 would mess up that area.
Higher K values would continue to make such mistakes.

(b) On the figure, draw the decision boundary for a perceptron. Could you get a different boundary with a perceptron? (5 points)

yes. You can get many ∞ lines

(c) Could a depth-2 decision tree perfectly classify the above data? Draw the best (with respect to accuracy) depth-2 decision tree. (5 points)

Same as 491 # 8c.

Linear Classifier (20 points)

9. Suppose we have the following training data.

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

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

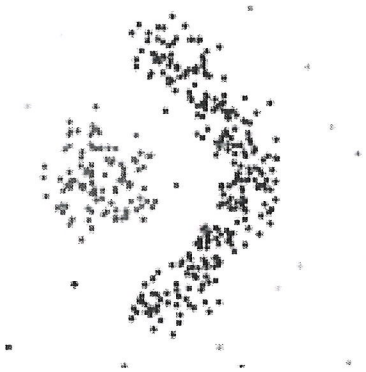
Same as 491 #9

(b) Would you get the same weights and bias if you iterated from s_4 up through s_1? (5 points)

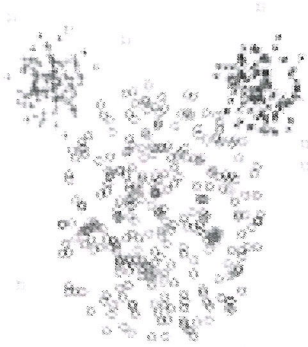
(c) How would your classifier from (a) classify the following test sample? $s_t = (1.5, 1, 1)$ (5 points)

K-Means (10 points)

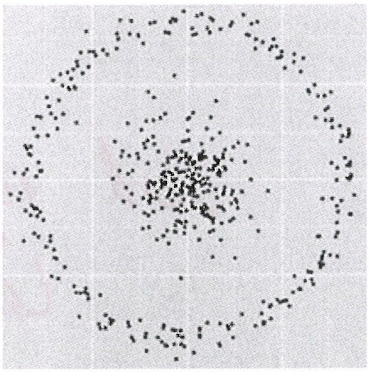
10. Given the three following sets of data (i, ii, and iii). Assume you want to cluster each set of data into clusters. Explain, and draw, what would likely happen with K-Means in each case and why.



(i) K=2



(ii) K=3



(iii) K=2

Same as #10 491