CS 273A Midterm Exam

Introduction to Machine Learning: Winter 2021

Thursday February 9th, 2021

Your name: **Row/Seat Number:**

Your ID #(e.g., 123456789) **UCINetID (e.g.ucinetid@uci.edu)**

- **Declaration of Honor.** By submitting this exam you are making the following declaration: I hereby declare, upon my Honor, that this work is my own, and that in reaching an answer I have not assisted any other person, nor have been assisted by any other person in any form. I acknowledge that, had I cheated, I would have been the kind of person who cheats, and I do not wish to be the kind of person who cheats.
- Please put your name and ID **on every page**.
- Total time is 60 minutes. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- Please ensure your final answer is contained in the space provided. We will not consider or grade anything beyond that space.
- If you need any clarification, please ask in our zoom room: `https://uci.zoom.us/j/94903054276`
- You may use **one** sheet containing handwritten notes for reference.
- Turn in your notes and any scratch paper with your exam.

Problems

1 Bayes Classifiers, ***(20 points.)*** **3**

2 Decision Trees, ***(16 points.)*** **5**

3 Linear and Nearest Neighbor Regression, ***(20 points.)*** **7**

4 True/False, ***(20 points.)*** **9**

5 VC-Dimensionality, ***(24 points.)*** **11**

Total, *(100 points.)*

This page is intentionally blank, use as you wish.

Name: ID#:

Problem 1 Bayes Classifiers, *(20 points.)*

x_1	x_2	y
a	b	0
b	c	0
b	c	0
c	c	0
a	c	1
a	b	1
b	a	1
b	b	1

Consider the table of measured data given at right. We will use the two observed features x_1, x_2 to predict the class y. Each feature can take on one of three values, $x_i \in \{a, b, c\}$.
In the case of a tie, we will prefer to predict class $y = 0$.

(1) Write down the probabilities learned by a naïve Bayes classifier: *(8 points.)*

$p(y = 0)$: 1/2 $p(y = 1)$: 1/2

$p(x_1 = a \,|\, y = 0)$: 1/4 $p(x_1 = a \,|\, y = 1)$: 1/2

$p(x_1 = b \,|\, y = 0)$: 1/2 $p(x_1 = b \,|\, y = 1)$: 1/2

$p(x_1 = c \,|\, y = 0)$: 1/4 $p(x_1 = c \,|\, y = 1)$: 0

$p(x_2 = a \,|\, y = 0)$: 0 $p(x_2 = a \,|\, y = 1)$: 1/4

$p(x_2 = b \,|\, y = 0)$: 1/4 $p(x_2 = b \,|\, y = 1)$: 1/2

$p(x_2 = c \,|\, y = 0)$: 3/4 $p(x_2 = c \,|\, y = 1)$: 1/4

(2) Using your naïve Bayes model, compute: *(6 points.)*
$p(y = 0|x_1 = a, x_2 = c)$: 3/5 $p(y = 1|x_1 = a, x_2 = c)$: 2/5

(3) Compute probabilities $p(y = 0|x_1 = a, x_2 = c)$ and $p(y = 1|x_1 = a, x_2 = c)$ for a joint (not naïve) Bayes model trained on the same data. *(6 points.)*

0 1

This page is intentionally blank, use as you wish.

Name: ID#:

Problem 2 Decision Trees, *(16 points.)*

x_1	x_2	x_3	y
0	0	0	1
1	0	1	1
1	0	1	1
1	1	1	0
0	1	0	0
1	0	0	0

Consider the table of measured data given at right. We will use a decision tree to predict the outcome y (one of two classes) using three features, x_1, x_2, x_3, where each can take one of two values: $0, 1$. In the case of ties, we prefer to use the feature with the smaller index (x_1 over x_2, etc.) and prefer to predict class 0 over 1. You may find the following values useful (**do not** leave logs unexpanded):

$\log_2(1) = 0$ $\log_2(2) = 1$ $\log_2(3) = 1.6$ $\log_2(4) = 2$
$\log_2(5) = 2.3$ $\log_2(6) = 2.6$ $\log_2(7) = 2.8$ $\log_2(8) = 3$

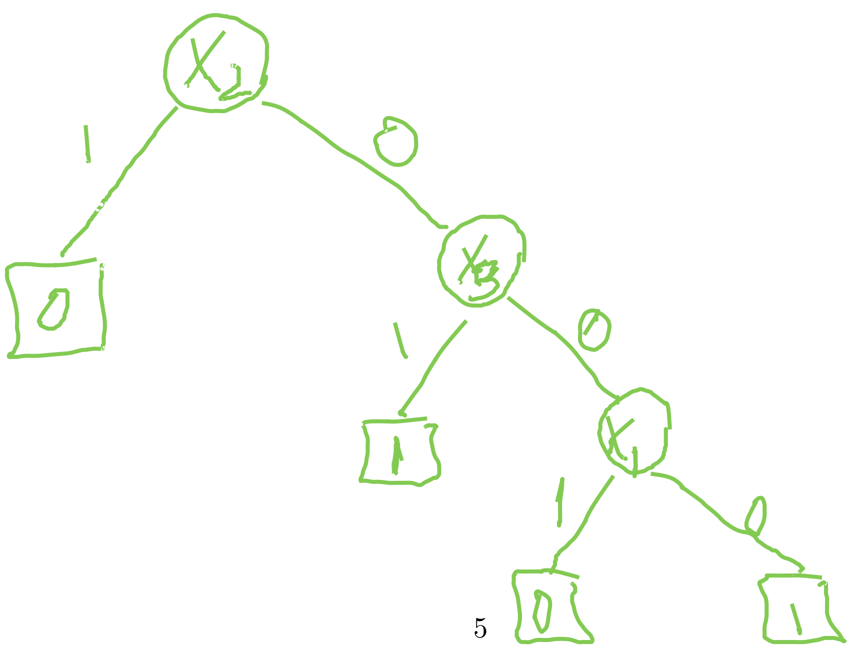
(1) What is the entropy of y? *(4 points.)*

1

(2) What is the information gain of x_1, x_2 and x_3? *(8 points.)*

0 7/15 1/15

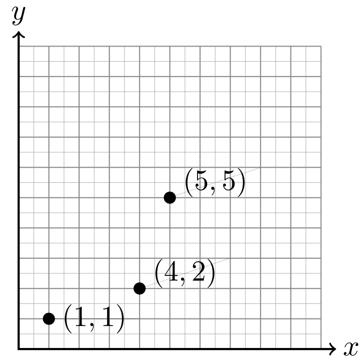
(3) Based on the information gain computed in (2), and any follow-up computations you'll need, build the complete decision tree learned on this data. *(4 points.)*



This page is intentionally blank, use as you wish.

Problem 3 Linear and Nearest Neighbor Regression, *(20 points.)*

Consider the data points shown at right, for a regression problem to predict y given a scalar feature x. As a reminder, leave-one-out with m data points is m-fold cross-validation.



(1) Compute **leave-one-out** cross-validation error of a zero-order (constant) predictor, $f(x) = \theta_0$. *(4 points.)*

$$\frac{1}{3}\left[\left(\frac{5}{2}\right)^2 + 1^2 + \left(\frac{7}{2}\right)^2\right] = 6.5$$

(2) Compute the **leave-one-out** cross-validation error (MSE) of a 1-nearest neighbor predictor. *(4 points.)*

$$\frac{1}{3}\left[1^2 + 3^2 + 3^2\right] = 19/3$$

(3) Compute the **leave-one-out** cross-validation error (MSE) of a 2-nearest neighbor predictor. *(6 points.)*

$$\frac{1}{3}\left[\left(\frac{5}{2}\right)^2 + 1^2 + \left(\frac{7}{2}\right)^2\right] = 6.5$$

(4) Compute the **leave-one-out** cross-validation MSE of a first-order linear regressor, $f(x) = \theta_0 + \theta_1 x$. *(6 points.)*

$$\frac{1}{3}\left[8^2 + 2^2 + \left(\frac{8}{3}\right)^2\right] = \frac{676}{27}$$

This page is intentionally blank, use as you wish.

Name: ID#:

Problem 4 True/False, *(20 points.)*

Here, assume that we have m data points $y^{(i)}, x^{(i)}$, $i = 1 \ldots m$, each with n features, $x^{(i)} = [x_1^{(i)} \ \ldots \ x_n^{(i)}]$. For each of the scenarios below, circle one of "true" or "false" to indicate whether you agree with the statement.

1 **True** or **false**: Applying "early stopping" in SGD by increasing the convergence tolerance (gradient size below which we stop) usually reduces overfitting and increases the bias of the learner.

2 **True** or **false**: With sufficient depth, a decision tree can approximate any Boolean function.

3 **True** or **false**: Increasing the regularization penalty of a linear regression model usually decreases the resulting model's variance.

4 **True** or **false**: When training a linear classifier using the logistic threshold function, negative log-likelihood loss may be preferable to mean square error (MSE) loss, because MSE has gradient 0 and stops updating when all points are classified correctly, even if their margin is small.

5 **True** or **false**: Increasing k in a k-nearest neighbors classifier usually decreases the bias.

6 **True** or **false**: Stochastic gradient descent usually requires more gradient steps than batch gradient descent when the number of data points m is very large.

7 **True** or **false**: For a perceptron, using $2n$ features per data point by adding n random values to each data point usually increases the resulting model's variance.

8 **True** or **false**: For the gradient computed in stochastic gradient descent, increasing the mini-batch size usually decreases both the bias and the variance of a mini-batch gradient as an estimate for the full-batch gradient.

9 **True** or **false**: For a perceptron, increasing the number of training data points m usually increases the resulting model's bias.

10 **True** or **false**: Increasing k in k-fold cross validation usually decreases the bias of the average cross-validation loss as an estimate for the final model's test loss.

This page is intentionally blank, use as you wish.

Name: ID#:

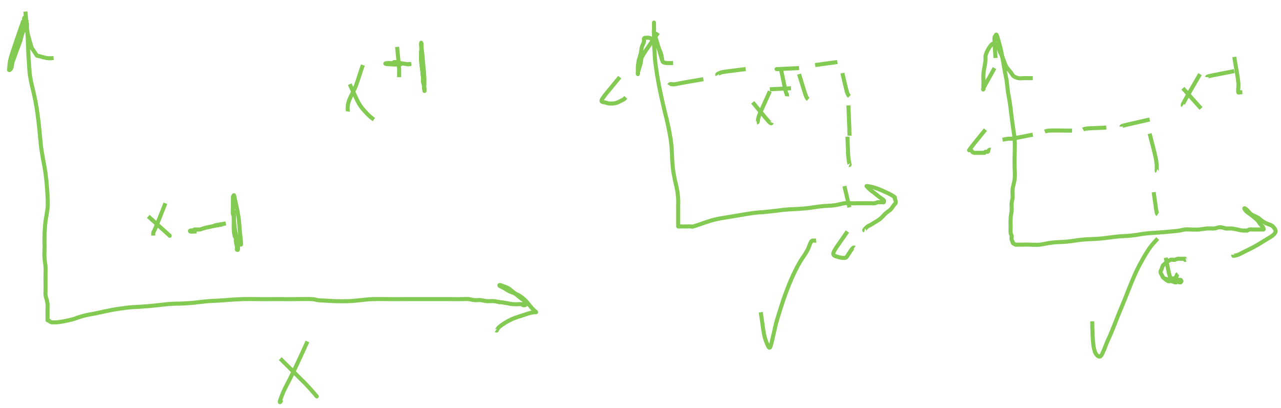
Problem 5 VC-Dimensionality, *(24 points.)*

We will be considering a family of "box-shaped" classifiers on a two-dimensional feature space (x_1, x_2), such that the region inside the box is clasified as +1.

(1) First, consider a simple classifier f_0 that uses a square which has one of the points at the origin, and c as the parameter that defines the edge size, i.e.

$$f_0(x) = \begin{cases} +1 & (0 < x_1 < c) \wedge (0 < x_2 < c) \\ -1 & \text{otherwise} \end{cases}$$

Show that this classifier has a VC-dimensionality of 1. *(6 points.)*

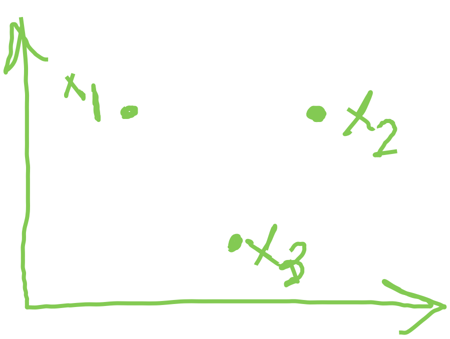


(2) Now consider an extension of this classifier with two additional parameters, f_1 that uses a point (a_1, a_2) and c as parameters to describe this square. Specifically:

$$f(x) = \begin{cases} +1 & (a_1 < x_1 < a_1 + c) \wedge (a_2 < x_2 < a_2 + c) \\ -1 & \text{otherwise} \end{cases}$$

Show that there exist 3 points that f_1 can shatter. What does it say about the VC-dimensionality of f? *(6 points.)*

Choose x_1, x_2 with same y



S_0

$VC \geq 3$

Continued on next page

This page is intentionally blank, use as you wish.

Name: ID#:

(3) Either show that f_1 can shatter 4 points, or argue, informally, why it cannot. What does this say about the VC-dimensionality of f_1? *(6 points.)*

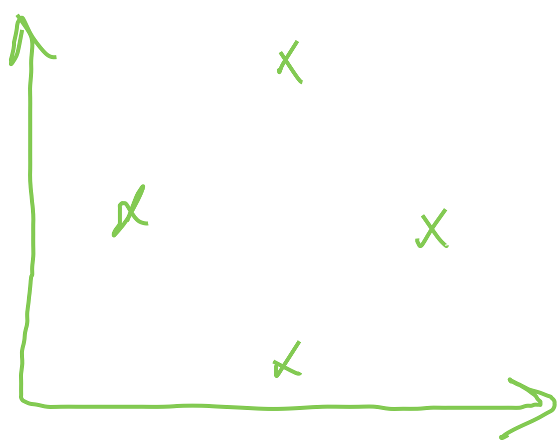
Cannot shatter 4 points

Just label them alternatively

(4) Now consider yet another extension f_2 where the region is a rectangle bounded by points (a_1, a_2) and (b_1, b_2), a total of 4 parameters:

$$f_2(x) = \begin{cases} +1 & (a_1 < x_1 < b_1) \wedge (a_2 < x_2 < b_2) \\ -1 & \text{otherwise} \end{cases}$$

Either show that f_2 can shatter 4 points, or argue, informally, why it cannot. What does this say about the VC-dimensionality of f_2? *(6 points.)*



- Since it is now a rectangular it can shatter 4 points forming a convex shape

This page is intentionally blank, use as you wish.

This page is intentionally blank, use as you wish.

This page is intentionally blank, use as you wish.