Practice Exam

University of Texas at Dallas

Fall 2021

Name: ____________________________________

NetID: ____________________________________

Question	Topic	Points
1	Short Answers	12
2	Support Vector Machines	18
3	Decision Trees and Ensembles	20
4	Neural Networks	15
5	Maximum Likelihood Estimation	12
6	Linear Regression and Loss Functions	13
7	VC Dimension	10
Total		100

Instructions:

1. This examination contains 21 pages, including this page.
2. You have **two and a half (2.5) hours** to complete the examination.
3. Either you can use this paper or separate set of sheets to fill in your answers. Write clearly so we can understand your handwriting.
4. Please do not search online for answers to the questions. If the answers are similar to something available online, you will get zero points on this examination.
5. The examination has to be done individually by everyone. If someone copies, the entire group of students involved will get a zero.
6. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult one.
7. All the Best!!

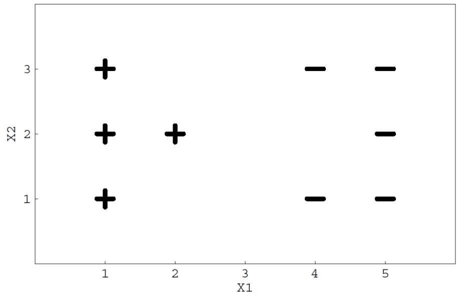
Question 1: Short Answers

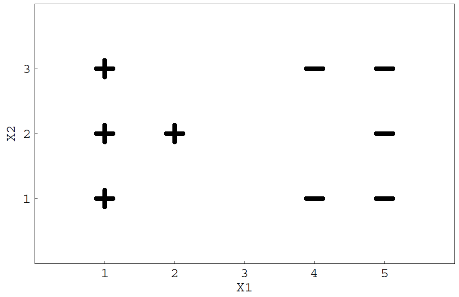
[12 pts] Please provide short and clear answers for the questions below.

(a) (4 points) Consider the 2 dimensional dataset given below. Circle examples having the following property: removing any of the examples and retraining the classifier would yield a different decision boundary than training on the full dataset. Circle examples for the following classifiers:

- Left Figure: Linear Support Vector Machines
- Right Figure: Logistic Regression

Briefly explain why.

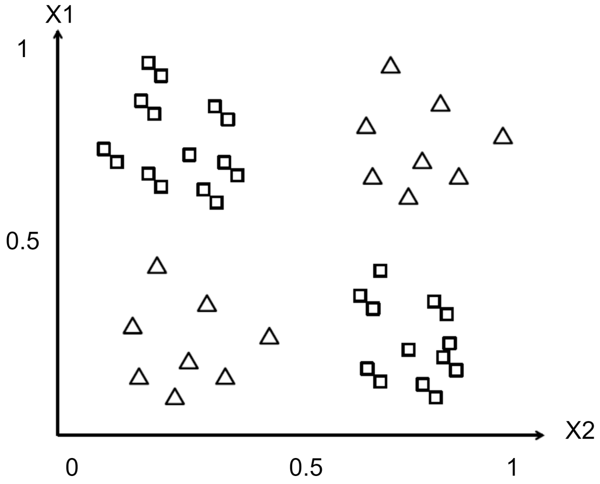




(b) (4 points) Come up with a one dimensional dataset having more than 7 but less than 21 examples such that the leave one out cross validation accuracy for 1-nearest neighbor is 100% but for 3-nearest neighbors, it is 0%.

(c) (4 points) This question has three parts. Refer to the dataset below (triangles are negatives and squares are positives).

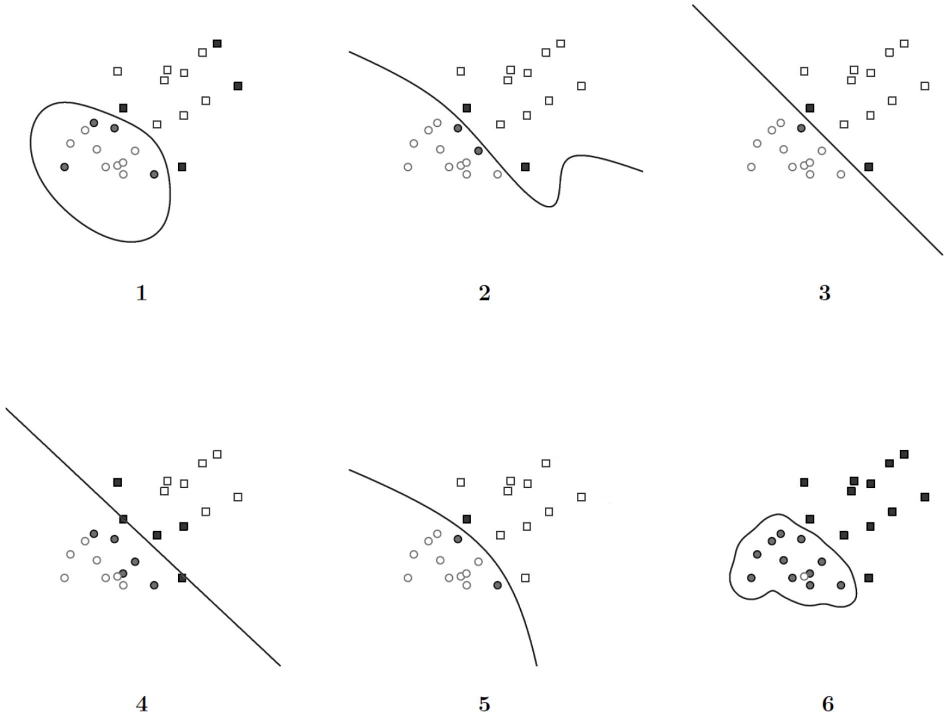
- (1 point) Can this dataset be perfectly classified (i.e. zero training error) with a linear classifier?
- (2 points) Give an example of a kernel such that a kernel SVM will have zero training error.
- (1 point) Draw a decision tree having zero training error.



Question 2: Support Vector Machines

[18 pts] This question on SVMs and Kernels has two parts.

(a) Part 1 (10 Points) The Figure below plots the decision-boundaries using different kernels and/or different slack penalties. There are two classes of data $y_i \in \{-1, 1\}$ corresponding to circles or squares. The solid circles/squares denote the support vectors. For each of the optimization problem below, label the corresponding plot and **explain why** you picked that plot in brief. Each sub-part below is for two points.



- A soft-margin linear SVM with slack penalty $C = 0.1$.

- A soft-margin linear SVM with slack penalty $C = 10$.

- A hard-margin kernel SVM with $K(u, v) = u.v + (u.v)^2$

- A hard-margin kernel SVM with $K(u, v) = \exp(-1/4||u - v||^2)$

- Notice that there are only four classifiers provided above (sub-parts (a)-(d)), but there are six decision boundaries. What are likely classifiers (kernels) which produce the rest of the boundaries.

(b) Part 2 (8 Points) This question is on Kernel operations.

- (3 points) Let k_1, k_2 be two valid kernels, i.e. $k_1(x, y) = \phi_1(x)^T \phi_1(y)$ and $k_2(x, y) = \phi_2(x)^T \phi_2(y)$. Then is $k_1 + k_2$ a valid kernel. Prove it by explicitly constructing a corresponding feature map $\phi(z)$.

- (2 points) Given a constant $c \geq 0$, prove or disprove that ck is a valid kernel if k is a valid kernel.

- (3 points) Let k_1, k_2 be two valid kernels. Then is $k_1 - k_2$ a valid kernel? What about $k_1 - 2k_2$? If it is, prove it and if not, give an example why it is not a valid kernel.

Question 3: Decision Trees and Ensembles

[20 pts] This question consists of three parts. Part 1 is of 8 points, part 2 is of 7 points, and part 3 is for 5 points.

(a) Part 1 (8 points): Consider the following two 2D datasets (with labels +1 and -1):

Dataset 1: $\{[(-1,-1),+1],[(-1,1),-1],[(1,-1),-1],[(1,1),+1]\}$ and

Dataset 2: $\{[(-1,-1),-1],[(-1,1),-1],[(1,-1),-1],[(1,1),+1]\}$.

For both datasets, provide a decision tree classifier and a ensemble classifier[1] having at most four weak classifiers that achieve 0 training error. If no classifier can be found to be consistent with the data, clearly **explain why**.

[1] Recall that an ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples.

(b) Part 2 (7 points): Suppose that for each data point, the feature vector $x \in \{0,1\}^m$, i.e. x consists of m binary valued features, the class label $y \in \{-1,1\}$, and the true classifier is a majority vote over the features, i.e. $y = sign(\sum_{i=1}^{m}(2x_i - 1))$, where x_i is the ith component of the feature vector.

- Describe (either in words or by drawing a schematic/picture) a binary decision tree having minimum depth that is consistent with the data. How many leaves does it have?
- Describe an ensemble classifier with the minimum number of weak classifiers that is consistent with the data. Precisely specify the weak classifiers and their weights.

(c) Part 3 (5 points): In class we used decision trees and ensemble methods for classification, but we can use them for regression as well (i.e. learning a function from features to real values). Let us imagine that our data has 3 binary features A, B, C, which take values 0/1, and we want to learn a function which counts the number of features which have value 1.

- Draw a decision tree which represents this function. How many leaf nodes does it have?
- Represent this function as a sum of decision stumps. How many decision stumps do we need?

Question 4: Neural Networks

[15 pts] Consider two kinds of Neural Network activation functions. A linear function $y = w_0 + \sum_i w_i x_i$ and a hard threshold: $y = 1$ if $w_0 + \sum_i w_i x_i \geq 0$ and 0 otherwise.

Which of the following functions can be exactly represented by a neural network with one hidden layer which uses linear and/or hard threshold activation functions? For each case, **justify** your answer. Also, if the answer is no, specify some simple non-linearities (activation functions) which might represent them. The question has five sub-parts: (a)-(e). Each sub-part is worth three points.

(a) Polynomials of degree one.

(b) Hinge loss: $h(x) = \max(1 - x, 0)$

(c) polynomials of degree two

(d) Piecewise constant functions[2].

[2] A function is said to be piecewise constant if it is locally constant in connected regions separated by a possibly infinite number of lower-dimensional boundaries.

(e) Piecewise Linear functions[3].

[3] A piecewise linear function is a function defined on a (possibly unbounded) interval of real numbers, such that there is a collection of intervals on each of which the function is an affine function.

Question 5: Maximum Likelihood Estimation

[12 pts] Suppose we know a continuous random variable X is uniformly distributed between values 0 and a positive number c, but c is unknown. In other words, $P(X = x|c) = \frac{1}{c} I(0 \leq x \leq c)$. To help estimate c, we observe N independent samples $x_1, x_2, \ldots, x_N$ of X. This question has three sub-parts: (a)-(c).

(a) (2 points) Write the expression for the joint likelihood of $P(x_1, \cdots, x_N|c)$

(b) (6 points) Find the maximum likelihood estimate of c. Your answer should be a closed form solution for the estimate $\hat{c}$. **Show your work.**

(c) (4 points) How many independent samples N are required in order to obtain an estimate $\hat{c}$ which is at most a fraction $1+\epsilon$ from the true value c (denoted by c^*) with probability $1-\delta$? Provide answers for $\epsilon = \delta = 0.05$.

Hint: Use the Chernoff bound: $P(|\hat{c} - c^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$.

Question 6: Linear Regression and Loss Functions

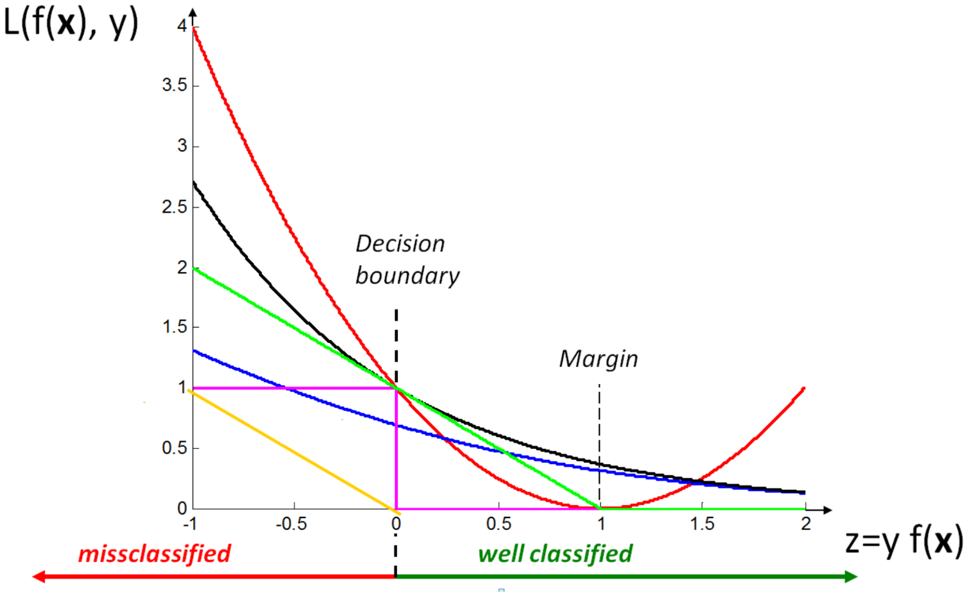
[13 pts] Part one of this question is on Linear regression and part two is on loss functions.

(a) Part 1 (7 points): Consider a regression problem where we want to predict y from a single feature x. We are given n training data points $(x_i, y_i)_{i=1}^n$. Consider two possible models to be estimated using linear regression: $y_i = w_0 + w_1 x_i + \epsilon_i$ and second as $y_i = w_0 + w_1 x_i + w_2 x_i^2 + \epsilon_i$. Assume that the error terms ϵ_i are independent and identically distributed from a normal distribution with zero mean.

- (4 points) Derive the expression for estimating the parameters of the two models. **Show your work**.

- (3 points) Will one model fit the training data better than the other, will they fit equally well, or is it impossible to say? Explain your reasoning. What about the test data?

(b) Part 2 (6 points) The plot below compares six loss functions: a) 0/1 Loss, b) Square Loss $1 - z^2$, c) Exponential (adaboost) loss, d) Perceptron Loss, e) Hinge Loss, and f) Logistic Loss. Identify the loss functions by their colors (yellow, black, red, green, pink, blue).



Question 7: VC Dimension

[10 pts] Given a hypothesis class $\mathcal{H}$, the VC dimension $VC(\mathcal{H})$ is defined to be the size of the largest set shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter arbitrarily large sets, then its VC dimension is ∞. This question consists of three sub-parts: (a)-(c).

(a) Part 1 (4 points): It is sometimes useful to think of the VC dimension as being related to the number of parameters needed to specify an element of $\mathcal{H}$. What is the VC dimension of the set of hypotheses of the following form: $h_\alpha(x) = 1$ if $\alpha_d x^d + \alpha_{d-1} x^{d-1} + \cdots + \alpha_0 > 0$ and 0 otherwise? **Justify your answer.**

(b) Part 2 (3 points): Despite the above, the VC dimension is not always related to the number of parameters. Provide an example of a hypothesis class which has M parameters, but its VC dimension is 1?

(c) Part 3 (3 points): Conversely, provide an example of a hypothesis class, where the number of parameters is constant (or say, even 1) but the VC dimension is unbounded (i.e. ∞).