CSE514 Fall 2021 Exam 1

Name: Anh Le **Student ID:** 488493

This exam consists of 13 total pages.

Page 1 is this page, with your name and ID.
Page 2 has the datasets you need to answer the exam questions
Page 3 is your cheat sheet of formulas
Page 4 has a division table, where the values are row/column
Pages 5-13 are the exam questions

You may **not** use any notes or reference materials of your own
You may use a calculator
You may leave answers in fraction form
You may answer questions on scratch paper

Please check that you have all the pages
Page numbers are on the top righthand corner

When you are finished, please check that any scratch paper pages with exam answers are included in your stack before you staple them for submission.

This exam is worth a total of 85 points
This exam includes extra credit opportunities, up to 8 EC points

You have 1hr 30min to finish this exam.
I highly recommend saving the EC points for last

Dataset 1: Wine Samples

ID	Citric Acid	Residual Sugar	Alcohol	Color	Quality
1	High 0.50	H 2.0	L 9.8	Red	Medium
2	L 0.00	H 1.9	L 9.4	Red	Low
3	L 0.00	L 1.8	L 9.4	Red	Low
4	L 0.06	L 1.6	L 9.4	Red	Low
5	H 0.65	L 1.2	H 10.0	Red	High
6	H 0.37	L 1.2	H 10.8	White	Low
7	H 0.40	L 1.5	H 12.4	White	High
8	H 0.62	H 19.3	L 9.7	White	Medium
9	H 0.38	L 1.5	H 11.4	White	High
10	L 0.04	L 1.1	L 9.6	White	Medium

Dataset 2: Tic-tac-toe games

ID	Move1	Move2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	Winner
1	Center	UR	LL	LR	CR	CL	UC	LC	UL	Tie
2	Center	UR	LR	UL	UC	LC	CL	CR	LL	Tie
3	UR	Center	LR	CR	UL	CL	-	-	-	O wins
4	LR	Center	LL	LC	UC	CL	CR	UR	UL	Tie
5	CR	LL	CL	Center	UR	LC	LR	-	-	X wins

Game rules:

1. X goes first, and then X and O take turns filling in a 3 x 3 grid
2. Game ends when either X or O fills in 3 in a row (Win) or the grid is filled (Tie)

Tic-tac-toe board:

Upper Left (UL)	Upper Center (UC)	Upper Right (UR)
Center Left (CL)	Center	Center Right (CR)
Lower Left (LL)	Lower Center (LC)	Lower Right (LR)

Confusion matrix:

	Observed Positive	Observed Negative
Predicted Positive = Red	True Positive (TP)	False Positive (FP)
Predicted Negative = White	False Negative (FN)	True Negative (TN)

Useful formulas:

Distances between vectors A and B of length M:

Manhattan

$$\sum_{k=1}^{M} |a_k - b_k|$$

Euclidean

$$\sqrt{\sum_{k=1}^{M} (a_k - b_k)^2}$$

Hamming

$$\sum_{k=1}^{M} \begin{cases} 0 & if\ a_k = b_k \\ 1 & otherwise \end{cases}$$

$$H(x) = \sum p_i \log_2(1/p_i) = -\sum p_i \log_2 p_i$$

$$H(P||Q) = D_{KL}(P||Q) = \sum_k P_k \log \frac{P_k}{Q_k}$$

$$Recall\ = {}^{TP}/_{(TP\ +\ FN)}$$

$$Precision\ = {}^{TP}/_{(TP\ +\ FP)}$$

$$Gini\ Impurity(S_i)\ =\ 1 - \sum_{k \in labels} proportion(k\ in\ S_i)^2$$

$$Variance = \frac{\sum(X-\mu)^2}{N}$$

$$Leaky\ ReLU: \quad f(x) = \max(0.1x, x)$$

$$feature\ map\ size = \frac{N - F + 2P}{S} + 1$$

Division table, row divided by column

	1	2	3	4	5	6	7	8	9
1	1.00	0.50	0.33	0.25	0.20	0.17	0.14	0.12	0.11
2	2.00	1.00	0.67	0.50	0.40	0.33	0.29	0.25	0.22
3	3.00	1.50	1.00	0.75	0.60	0.50	0.43	0.38	0.33
4	4.00	2.00	1.33	1.00	0.80	0.67	0.57	0.50	0.44
5	5.00	2.50	1.67	1.25	1.00	0.83	0.71	0.62	0.56
6	6.00	3.00	2.00	1.50	1.20	1.00	0.86	0.75	0.67
7	7.00	3.50	2.33	1.75	1.40	1.17	1.00	0.88	0.78
8	8.00	4.00	2.67	2.00	1.60	1.33	1.14	1.00	0.89
9	9.00	4.50	3.00	2.25	1.80	1.50	1.29	1.12	1.00

Q1 (10pts): What data type are each of the features in Dataset 1?

Place a single check mark in each row to indicate the data type of the listed features.

		Data types			
		Numerical + Discrete	Numerical + Continuous	Categorical + Nominal	Categorical + Ordinal
Wine Features	Citric Acid		✓		
	Residual Sugar		✓		
	Alcohol		✓		
	Color			✓	
	Quality				✓

Q2 (10pts): A simple "stump" decision tree can be built to predict Color from other features. Start by considering Alcohol > 9.9 for the root, which will split Dataset 1 between a "True" leaf and a "False" leaf

Q2a (2pt): What color label does the "True" leaf get? White

Q2b (2pt): What color label does the "False" leaf get? Red

Q2c (6pts): What is the Total Weighted Gini Impurity of this stump?

Alcohol > 9.9
True: white — 1 Red, 3 White
False: Red — 4 Red, 2 White

$$\text{Gini Impurity}_{True} = 1 - \left(\left(\tfrac{1}{4}\right)^2 + \left(\tfrac{3}{4}\right)^2\right) = \frac{3}{8}$$

$$\text{Gini Impurity}_{False} = 1 - \left(\left(\tfrac{4}{6}\right)^2 + \left(\tfrac{2}{6}\right)^2\right) = \frac{4}{9}$$

$$\hookrightarrow \text{TWGI} = \frac{4}{10} \times \frac{3}{8} + \frac{6}{10} \times \frac{4}{9} = \frac{5}{12}$$

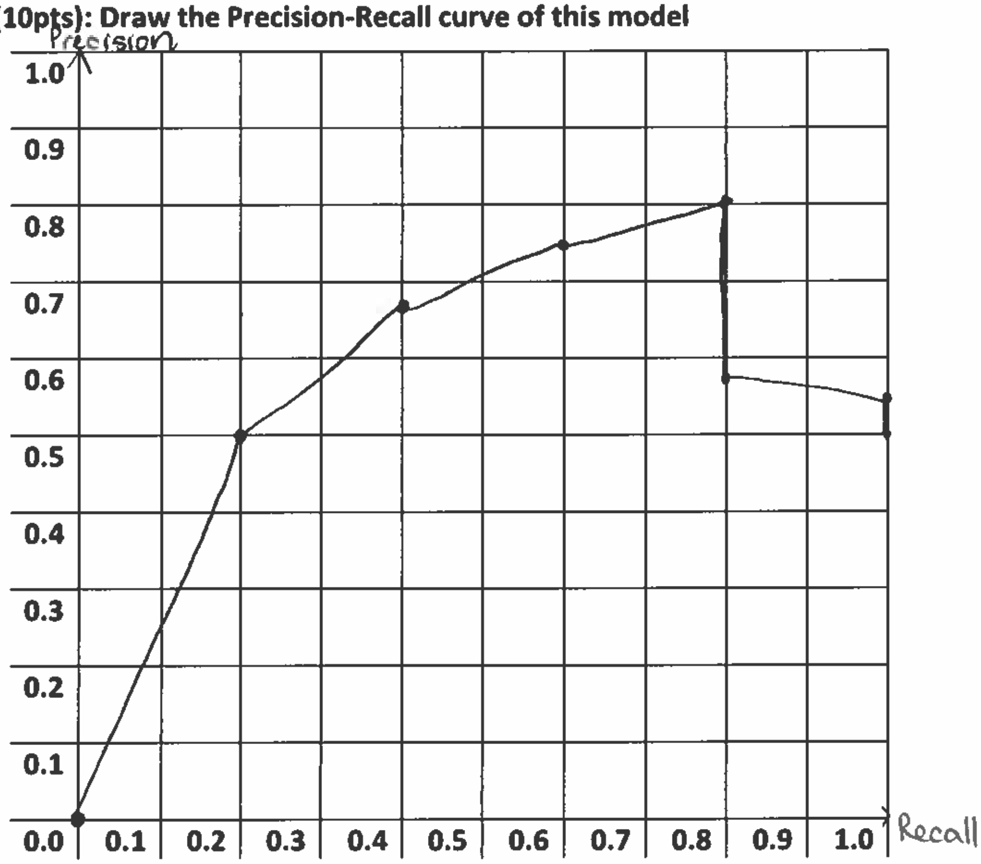
Q3 (10pts + 1EC): Instead of using Alcohol to predict Color, a model is defined to treat the value of a sample's Residual Sugar feature as a score for how confident the model is that the sample's Color is Red.

For example, the model is more confident that sample 1 is Red than sample 6 because sample 1 has a higher Residual Sugar value than sample 6.
At a threshold of 1.6, this model predicts samples 1, 2, 3, 4, and 8 to be Red, while the remaining samples are predicted to be White.

For this model, Red is the positive class, while White is the negative class

Q3a (10pts): Draw the Precision-Recall curve of this model

	Precision	Recall
9.3	0	0
.0	1/2 = 0.5	1/5 = 0.2
.9	2/3	2/5
.8	3/4	3/5
.6	4/5	4/5
5	4/7	4/5
2	5/9	1
.1	1/2	1



Q3b (+1 EC): How would the shape of the PR-curve change if the model had perfect accuracy?

The curve would be plat

Q4 (10pts): A linear regression model is fit to Dataset 1 to predict Alcohol:

$$Alcohol = 10 + x_1 - 0.01x_2 - 1.1x_3 + 0.5x_4$$

x_1: Citric Acid x_2: Residual Sugar

x_3: 1 if Red, 0 if White x_4: 2 if High, 1 if Medium, 0 if Low

Parameter B = 10   M = [1, -0.01, -1.1, 0.5]

Q4a (2pts): What is the model's prediction of alcohol for sample 1 in Dataset 1?

Alcohol = $10 + 0.5 - 0.01 \times 2 - 1.1 \times 1 + 0.5 \times 1 = 9.88$

To improve this linear model's fit, stochastic gradient descent is implemented:

the loss function is sum of squared errors
the learning rate in 0.1, and
only samples 5 and 10 were randomly selected to calculate the gradients

The current model predicts alcohol levels of sample 5 = 10.54, and sample 10 = 10.53

Q4b (6pts): Calculate the gradient for parameter B:

$$\frac{\partial L}{\partial b} = \sum -2(y_i - \hat{y}) = -2\left[(10 - 10.54) + (9.6 - 10.53)\right] = 2.94$$

Q4c (2pts): Calculate the new value for parameter B:

$$b = b - 0.1\frac{\partial L}{\partial b} = 10 - 0.1 \times 2.94 = 9.706$$

Q5 (5pts + 1EC): Use k-nearest neighbors to predict a new wine sample's quality.

ID	Citric Acid	Residual Sugar	Alcohol	Color	Quality
11	0.36 H	20.7 H	8.8 L	White	-

Q5a (5pts): What is the new wine sample's quality?
Use Hamming distance, and k = 3
For the numerical feature of Citric Acid, use Low for <0.2 and High for >=0.2
For the numerical feature of Residual Sugar, use Low for <1.9 and High for >=1.9
For the numerical feature of Alcohol, use Low for <10 and High for >=10
The categorical feature of Color can be used as it is

$d(1,11) = 0+0+0+1 = 1$ ✓

$d(2,11) = 1+0+0+1 = 2$

$d(3,11) = 1+1+0+1 = 3$

$d(4,11) = 1+1+0+1 = 3$

$d(5,11) = 0+1+1+1 = 3$

$d(6,11) = 0+1+1+0 = 2$ ✓

$d(7,11) = 0+1+1+0 = 2$

$d(8,11) = 0+0+0+0 = 0$ ✓

$d(9,11) = 0+1+1+0 = 2$

$d(10,11) = 1+1+0+0 = 2$

→ New sample's quality = Medium

Q5b (+1 EC): What is the new wine sample's quality?
Use Manhattan distance, and k = 1
Ignore the categorical feature of Color

$d(1,11) = (0.5-0.36) + (20.7-2.0) + (9.8-8.8) = 19.84$

$d(2,11) = (0.36-0) + (20.7-1.9) + (9.4-8.8) = 19.76$

$d(3,11) = (0.36-0) + (20.7-1.8) + (9.4-8.8) = 19.86$

$d(4,11) = (0.36-0.06) + (20.7-1.6) + (9.4-8.8) = 20$

$d(5,11) = \ldots$

$d(8,11) = (0.62-0.36) + (20.7-19.3) + (9.7-8.8) = 2.56$ ✓

→ New sample's quality = Medium

(Ignore the rest because Δresidual sugar is very large compared to $d(8,11)$)

Q6 (5pts + 1EC): An artificial neural network is requested with the goal of **predicting the winner of tic-tac-toe from the first five moves** of the games in Dataset 2.

Q6a (2pt): Is this a Classification problem or a Regression problem? Why?

Classipication because the output is [1, 0, -1] = [X wins, tie, O wins]

Q6b (3pts): Explain why a Recursive Neural Network is suited to solving this problem

Because RNN is good for sequential data, which here is a series of moves by the player

Q6c (+1 EC): The samples of Dataset 2 have no numerical values, but you can only use numerical values as inputs for a Neural Network. **How would you solve this problem?**

Utilize one-hot encoding for those numerical input values

Q7 (5pts): A convolutional neural network is requested with the goal of identifying the winner of a tic-tac-toe game from a 24 x 24 pixels image of the game after the last move is made.

Q7a (2pts): One of the filters has a 4 x 4 receptive field.
Using stride = 2 and padding = 1 on all sides, **what are the feature map dimensions?**

$$\frac{24 - 4 + 2 \times 1}{2} + 1 = 12 \quad \rightarrow 12 \times 12$$

Q7b (3pts): Explain the purpose of convolutional layers in this network

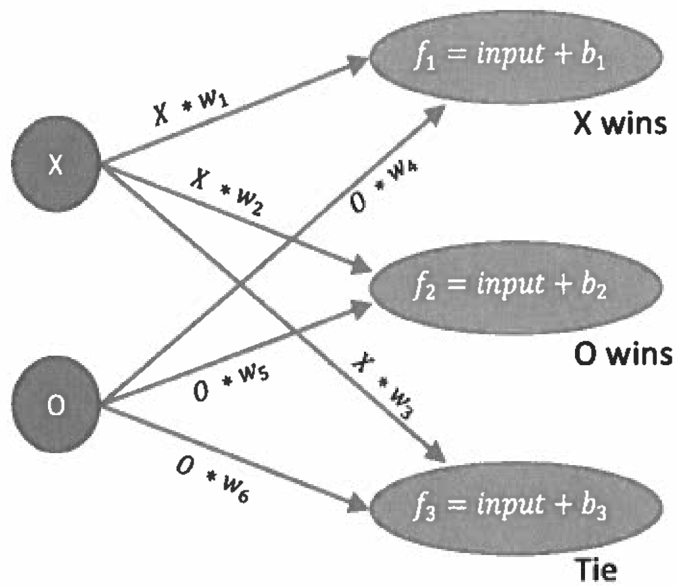
The convolutional layers search for overlapping batches of input values for feature using the pilter

Q8 (10pts): After the convolutional and pooling layers for the network from **Q7**, the end of the network looks like this:

Parameter values:

W = [3, -3, 1, -3, 3, 1]

B = [10, 10, -20]



Given the following:

the X node = 24

the O node = 16

Q8a (3pts): What is the value of f1?

$f_1 = (x \cdot w_1 + 0 \cdot w_4) + b_1 = 34$

Q8b (3pts): What is the value of f2?

$f_2 = (x \cdot w_2 + 0 \times w_5) + b_2 = -14$

Q8c (3pts): What is the value of f3?

$f_3 = (x \cdot w_3 + 0 \cdot w_6) + b_3 = 20$

Q8d (1pt): Does the CNN predict X wins, O wins, or a tie?

X wins

Q9 (5pts): A friend is trying to fit a multivariate linear regression model, but the error doesn't decrease from one iteration of gradient descent to the next. What would you recommend your friend should try? Give a brief explanation of why that would help.

My one suggestion would be lowering the stepsize because this would avoid updating the parameter by a large amount, leading to higher loss function.

Q10 (5pts): Summarize how to implement one of the following algorithms:

1. Bagging (Bootstrap Aggregating)
2. Boosting
3. Cross-Validation

First we train a model using the original training dataset. Then, we use that model's performance to re-weight the training datapoints so that the falsely predicted ones are worth more. Then, train a new model using the newly re-weighted dataset. Repeat previous steps. We get final predictions from weighted votes from all models where better performed models get bigger votes.

Q11 (10pts + 2.5EC): To better understand their product, a wine company collects data on 100,000 wine samples using the same features as Dataset 1. Now they want to know which model they should be using to analyze this dataset. Their goal is to predict quality from the other features.

Fill in 5 blocks of the following table by giving a pro or a con for each of the models. (+0.5 EC) for every additional block filled

The first row for linear regression has been filled in as an example

Model	Pro	Con
Linear Regression	Fast and easy to fit	The response value is categorical, not numerical, so it's necessary to also optimize thresholds before the model can make predictions
Logistic Regression	**11a** Fast & able to give categorical response	**11b** Cannot fit more complicated data.
Decision tree	**11c** Fast & easy to fit	**11d** Not accurate since the data is very large
Random Forest	**11e** Works with larger dataset	**11f** Not as fast compared to others
kNN Classifier	**11g** Might be easy to explain	**11h** Might not be very accurate compared to others; large dataset
Artificial Neural Network	**11i** Works for larger dataset	**11j** Have to make sure of different layers for each feature /input

TP 1	FP 3
FN 4	TN 2

Q12 (2.5 EC): Given the stump from **Q2** that splits Dataset 1 on Alcohol > 9.9 to predict Red or White, fill in values for the following performance measures.

Use the label of the "True" leaf as positive, and the label of the "False" leaf as negative.

(+0.25 EC) for every performance metric

1: True positive rate

$$\frac{TP}{TP+FN} = \frac{1}{1+4} = \frac{1}{5}$$

2: False positive rate

$$\frac{FP}{FP+TN} = \frac{3}{3+2} = \frac{3}{5}$$

3: False negative rate

$$\frac{FN}{FN+TP} = \frac{4}{5}$$

4: True negative rate

$$\frac{TN}{TN+FP} = \frac{2}{2+3} = \frac{2}{5}$$

5: Sensitivity

$$\frac{1}{5}$$

6: Specificity

$$\frac{2}{5}$$

7: Accuracy

$$\frac{TP+TN}{TP+FP+FN+TN} = \frac{3}{10}$$

8: F-score

$$\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2}{\frac{1}{1/5}+\frac{1}{1/4}} = \frac{2}{9}$$

9: Type I error (mistaken rejection of the null hypothesis)

3

10: Type II error (mistaken acceptance of the null hypothesis)

4