CPE/EE/AAI-695 Midterm Exam

The is an open-book and open-note exam.

By taking this exam, I pledge to abide by the Graduate Student Code of Academic Integrity.

Signature: _____ Printed Name: _____

I. Concept Questions (10 pts)

1. Circle True or False.
   a. When a decision tree is grown to full depth, it is more likely to fit the noise in the data. [**True**, **False**]

   b. Lasso Regression model does not have a close-formed solution (i.e., normal equation) [**True**, **False**]

2. What is the bias-variance trade-off? How can we address bias and variance respectively?

3. What is overfitting? List several techniques that can reduce overfitting.

II. Logistic Regression (16 pts)

Given a dataset $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$ where the target variable $y^{(i)}$ is either 0 or 1 and the feature $x^{(i)}$ is an n- dimensional vector, consider a **binary classification** problem. Let $p(x) = \Pr(y = 1|x)$ represent the probability that $y = 1$ when given feature x.

1) Given a linear regression model $h(x) = w_0 + w_1x_1 + \cdots + w_nx_n$. Derive the logistic regression model using Sigmoid function. That is, show how the Sigmoid function can be used to map $h(x)$ to $p(x)$ and derive the formula for $p(x)$ in terms of $h(x)$.

2) Explain why Sigmoid function is suitable for modeling the probability $p(x)$.

3) Show the log loss function for this binary classification problem. Show how to use Stochastic Gradient Descent to train the logistic regression model with the log loss function.

III. Decision Tree (14 pts)

Consider the following dataset for commute listed in Table 1 for a classification problem. We will use a decision tree learner based on information gain.

X_1 = "Own a car"	X_2 = "Distance to campus"	Y = "Commute Way"
Yes	Far	Drive
Yes	Far	Drive
Yes	Far	Bus
Yes	Close	Walk
No	Close	Bus
No	Close	Walk

1) Calculate the information gain for both attributes. You can use the approximation $\log_2 3 \approx 19/12$. Report the information gain as fraction or as decimals with the precision of three decimal digits. Show your calculations.

2) Report which attribute is used for the first split. Draw the decision tree resulting from using this split alone. Make sure to label the split attribute (which attribute the branch corresponds to, what is the predicted label for the leaf)

3) If a student doesn't own a car and is far from campus, how would this tree classify this example?

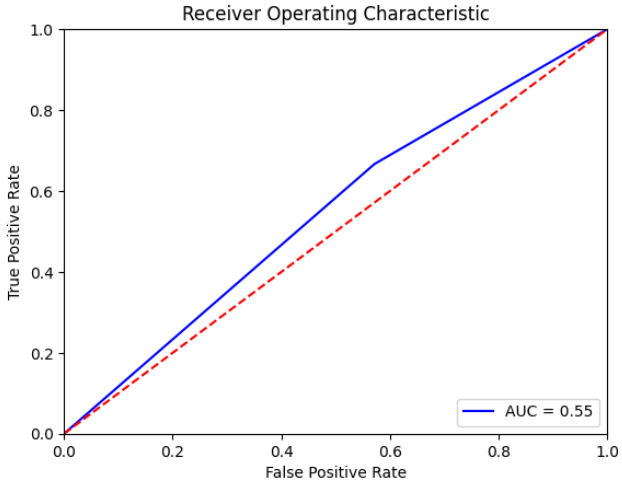
IV. Model Evaluation (10 pts)

Assume there is a dataset with 10 data samples in which the label vector is given as $y = [1,1,1,0,0,0,1,1,1,1]$. A binary classifier gives the predicting result $\hat{y} = [1,1,0,1,0,0,1,0,0,0]$. Take 1 as positive class and 0 as negative class.

1) Show the confusion matrix.

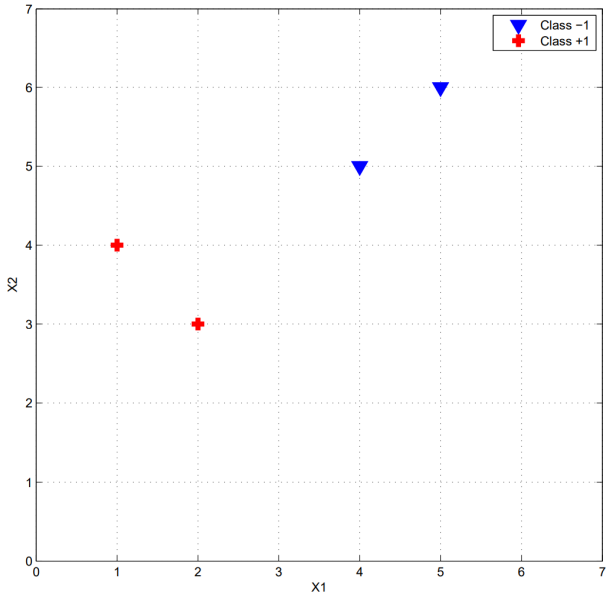
2) What are the precision, recall and F-1 score, respectively? Show your calculations.

3) Here is the ROC curve generated for this binary classifier, briefly discuss your observation, and how AOC/RUC reflects a model's performance.



V. SVM (8 pts)

Assume you are training SVM on a dataset with 4 data samples as shown below. This dataset consists of two examples with class label -1 and two examples with class label +1.



1) Find the weight vector w and bias b. What's the equation corresponding to the hyperplane (decision boundary)

2) Which point is the support vector? Write down all the answers.

VI. Ensemble Learning (12 pts)

Consider three binary classifiers that make the following predictions for a sample X. We want to make the final decision using the ensemble of the three classifiers.

	Result
Classifier 1	Class 1
Classifier 2	Class 2
Classifier 3	Class 1

1) If we use the majority vote as the fusion function, what is our final decision.

2) If we use the weighted majority vote as the fusion function, what is the final decision given the weights as follows:
Classifier 1: 0.2
Classifier 2: 0.4
Classifier 3: 0.4

3) If we use the Naïve Bayes method as the fusion function, what is our final decision? Given the confusion matrix as follows. Show your calculations.

(a) Classifier 1

	Class1	Class2
Class1	80	40
Class2	20	60

(b) Classifier 2

	Class1	Class2
Class1	60	10
Class2	40	90

(c) Classifier 3

	Class1	Class2
Class1	70	30
Class2	50	50

VII. Naïve Bayes (14 pts)

1. Prove Bayes' Theorem. Explain why it is useful for machine learning problems.

2. Consider the "PlayTennis'' dataset with 12 samples. Using the Naïve Bayes algorithm to predict a new instance given as <*Outlook* = sun, *Temperature* = Cool, *Humidity* = High, *Wind* = Strong>

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes

VIII. Neural Network (16 pts)

1. Can you represent the Boolean function shown in Table 2 with a single logistic threshold unit (i.e., a single unit from a neural network)? If yes, show the weights. If not, explain why.

A	B	F(A,B)
1	1	0
0	0	0
1	0	1
0	1	0

2. Consider a 3-layer feedforward neural network with two inputs: x_1, x_2 and one output unit y. The network has one hidden layer with two units h_1 and h_2. All the functions are linear without bias. Assume a learning rate $\eta = 0.4$, momentum $\alpha = 0.8$ and all the weights are initialized to 0.1, derive **the first training iteration** of neural networks using **Stochastic Gradient Descent** with **back-propagation** algorithm, and the following training samples.

x_1	x_2	y
1	0	1