CSM148 Homework 3

Due date: Friday, March 17 at 11:59PM PST
Instructions:

All work must be completed individually.

Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

1 Decision Tree [20 points]

The following table contains training examples that help predict whether a patient is likely to have a heart attack. Suppose we want to build a decision tree based on the given data using entropy gain.

PATIENT ID	CHEST PAIN?	MALE?	SMOKES?	EXERCISES?	HEART ATTACK?
1.	yes	yes	no	yes	yes
2.	no	no	yes	no	yes
3.	yes	yes	yes	no	yes
4.	no	yes	no	yes	no
5.	yes	no	yes	yes	yes
6.	no	yes	yes	yes	no
7.	no	no	no	yes	no
8.	yes	no	yes	no	yes

(a) **(3 points)** What is the overall entropy of whether or not a patient is likely to have a heart attack, without considering any features?

(b) **(8 points)** Suppose we want to build a decision tree by selecting **one** feature to split. What are the information gains for the four given features, and which feature do you want to choose at the first step?

(c) **(3 points)** To construct a full decision tree, we will need to re-calculate the Information Gain for the remaining features and continue the splits, until a stop criterion is met. What are possible stop criterions for this process?

(d) **(3 points)** Should we standardize or normalize our features?

(e) **(3 points)** Are decision trees robust to outliers?

2 Perceptron [15 points]

Consider training a Perceptron model $y = \text{sgn}(\mathbf{w}^\top \mathbf{x})$, $\mathbf{w} \in \mathbb{R}^d$ on a dataset $D = \{(\mathbf{x}_i, y_i)\}, i = 1 \dots 5$. Both $\mathbf{w}$ and $\mathbf{x}_i$ are vectors of dimension d, and $y \in \{+1, -1\}$ is binary. Assume that the bias term is already augmented in $\mathbf{x}_i$: $\mathbf{x}_i = [1, x_1, \dots, x_{d-1}]$. The activation function is a sign function where $\text{sgn}(x) = 1$ for all $x > 0$ and $\text{sgn}(x) = -1$ otherwise. The Perceptron algorithm is given below,

Algorithm 1 Perceptron

```
Initialize w = 0
for i = 1 ... N do
    if y_i ≠ sgn(w^T x_i) then
        w ← w + y_i x_i
    end if
end for
return w
```

(a) **(5 points)** The Perceptron model is trained for 1 epoch, i.e. iterated over the entire dataset once, and made mistakes on the following three data points: $\{(\mathbf{x}_1, y_1), (\mathbf{x}_3, y_3), (\mathbf{x}_4, y_4)\}$. What will be the weight vector $\mathbf{w}$ after this training epoch? Express $\mathbf{w}$ in using variables $\mathbf{x}_i$ and y_i where you will figure out what values of $i \in \{1, 2, 3, 4, 5\}$ will be used from the algorithm.

(b) **(5 points)** Let $d = 3$ and the data points be given as follows

i	$\mathbf{x}_{i,1}$	$\mathbf{x}_{i,2}$	$\mathbf{x}_{i,3}$	y_i
1	1	1	0	+1
2	1	2	-1	+1
3	1	1	-3	-1
4	1	3	-1	+1
5	1	1	-1	+1

Following the formulation of your answer in (a), what is $\mathbf{w}$ given the values of the data points? Express $\mathbf{w}$ as vector of numbers this time. Furthermore, if we iterate through the dataset again, will the model make a mistake on $\mathbf{x}_1$ again? Write it's prediction on $\mathbf{x}_1$.

(c) **(5 points)** State one difference between the given Perceptron model and the logistic regression model taught in class.

3 Neural Networks [20 points]

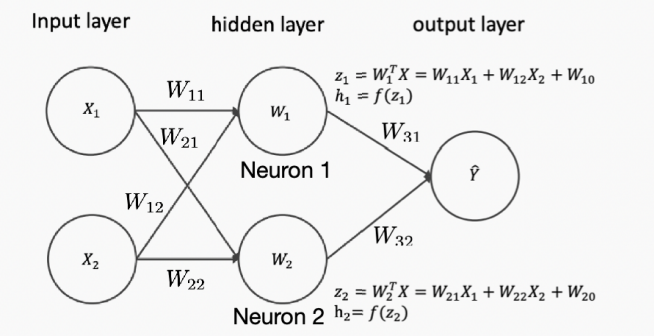


Figure 1: Neural Network

(a) **(4 points)** Refer to Lecture 13 for the activation functions of neural networks. Considering a binary classification problem where $y \in \{0, 1\}$, what are possible activations choices for the hidden and output layers repectively? Briefly explain why.

(b) **(3 points)** Consider a binary classification problem where $y \in \{0, 1\}$. And we consider the neural network in Figure 2 with 2 inputs, 2 hidden neurons, and 1 output. We let neuron 1 use **ReLU** activiation and neuron 2 use **sigmoid** activiation, respectively. And the other layers use the linear activiation function. Suppose we have a input $X_1 = 2$ and $X_2 = -3$, with label $y = 1$. And the weights are initialized as $W_{11} = 0.9, W_{12} = 0.4, W_{21} = -1.5, W_{22} = -0.7, W_{31} = -0.2, W_{32} = 1.6$, and the bias term W_{10}, W_{20}, W_{30} are all initialized to be 0. Compute the output of the network. (Round to the 2nd decimal in your final answer.)

(c) **(4 points)** We consider the binary cross entropy loss function. What is the loss of the network on the given data point in (b)? What is $\frac{\partial \mathcal{L}}{\partial \hat{y}}$ where $\hat{y}$ is the output of the neural network? (**Hint.** Refer to the lecture slides for the definition of a binary cross entropy loss function)

(d) **(6 points)** We now consider the backward pass. Given the same initialized weights and input as in (b), write the formula and calculate the derivative of the loss w.r.t W_{12}, i.e. $\frac{\partial \mathcal{L}}{\partial W_{12}}$.

(e) **(3 points)** Given the neural network as in (b), how many parameters does the network have? (**Hint**. Each weight unit counts as a parameter, and we also consider the bias terms $(W_{10}, \ldots)$ as parameters.)

4 Multi-class Classification [10 points]

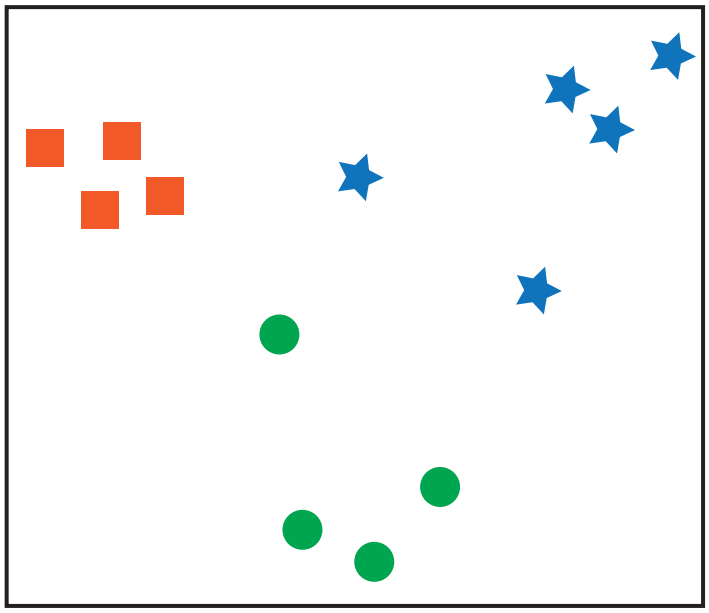


Figure 2: Multiclass Logistic Regression

(a) **(5 points)** Consider a multi-class classification problem with 5 classes and 20 features. We will use the logistic regression model to first build our binary classifier. Then, what will be the total number of parameters for using **One vs. Rest (ovr)** strategies for the multi-class classification task using logistic regression? (**Hint**. Refer to lecture 11 for multi-class logistic regression model.)

(b) **(5 points)** Consider a new multi-class classification problem with 3 classes. The distribution of the points is shown in the figure (Square - Class 1, Circle - Class 2, Star - Class 3). Draw the linear classifiers used for classifying the three classes, using (i) One vs. Rest (OvR), and (ii) Multinomial approaches. Your drawing does not have to be precise, roughly indicate the classifiers and jut make sure you can show the differences.

5 Decision Boundary [10 points]

Consider the classification problems with two classes, which are illustrated by circles and crosses in the plots below. In each of the plots, one of the following classification methods has been used, and the resulting decision boundary is shown:

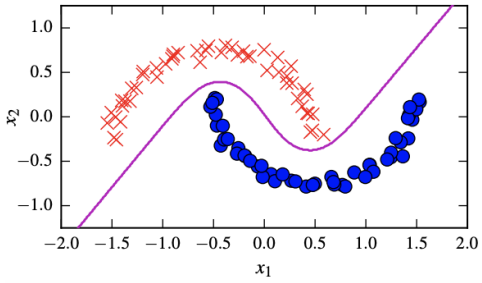
(1) **(2.5 points)** Decision Tree

(2) **(2.5 points)** Random Forest

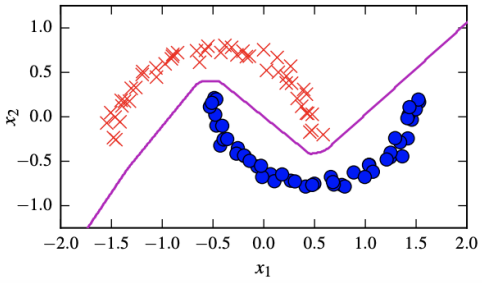
(3) **(2.5 points)** Neural Network (1 hidden layer with 10 ReLU)

(4) **(2.5 points)** Neural Network (1 hidden layer with 10 tanh units)

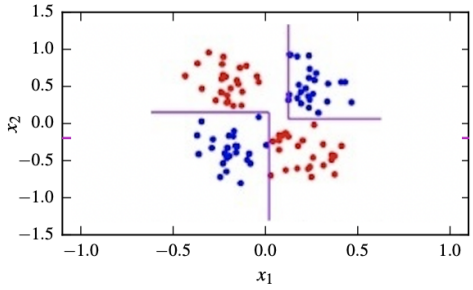
Assign each of the previous methods to exactly one of the following plots (in a one to one correspondence) by annotating the plots with the respective letters, and **explain briefly** why did you make each assignment.



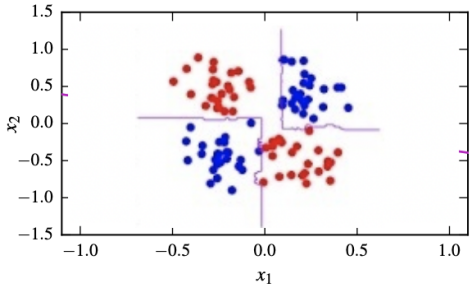
(a)



(b)



(c)



(d)