CSM148 Homework 3

Due date: Friday, March 10 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?

Solution: The overall entropy is $-\frac{5}{8}\log_2\left(\frac{5}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) = 0.954434$.

(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?

Solution:
InfoGain(Chest Pain) $= 0.954434 - 0 - \frac{1}{2}(-\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4})) = 0.5488$
InfoGain(Male) $= 0.954434 - \frac{1}{2} \times 1 - \frac{1}{2}(-\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4})) = 0.0488$
InfoGain(Smoke) $= 0.954434 - \frac{5}{8}(-\frac{4}{5}\log_2(\frac{4}{5}) - \frac{1}{5}\log_2(\frac{1}{5})) - \frac{3}{8}(-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})) = 0.1589$
InfoGain(Exercise) $= 0.954434 - \frac{5}{8}(-\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5})) - 0 = 0.3476$
Choose "Chest Pain" to split at the first step because it gives the highest information gain.

(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?

Solution: Possible solutions are:

- all instances in the region belong to the same class
- if number of intances in the subregion fall below pre-defined threshold
- total number of leaves exceeded the predefined threshold
- no features remain for further split

- reach the largest depth
- Information Gain is smaller than the predefined threshold.

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

Solution: There is no need to standardize/normalize our features because decision trees are not trying to fit the data to a particular line. For each node in our tree, a given feature is being evaluated for making a decision independent of the other features.

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

Solution: Decision Trees are generally robust to outliers because it's merely drawing separation lines on a per-feature basis. It's okay if some values are extremely far from others, as we are effectively trying to figure out how to separate our data, not fit a line that represents all of the data.

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? Write $\mathbf{w}$ 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.

Solution: $\mathbf{w} = y_1\mathbf{x}_1 + y_3\mathbf{x}_3 + y_4\mathbf{x}_4$

(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	3	-1	+1

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

Solution: $\mathbf{w} = [1, 3, 2]$. It will not make mistake again and will predict +1.

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

Solution: The activation function is different, where the Perceptron model here uses a sign function for activation and the logistic regression model we learned uses a sigmoid function.

3 Neural Networks [15 points]

(a) **(4 points)** Refer to Lecture 13 for the activation functions of neural networks. Considering a binary classification problem, what are possible activations choices for the hidden and output layers repectively? Explain why.

Solution: Hidden layers: Any activations are ok. ReLU, Leaky ReLU, ELU and variants are popular choices. Activation functions in the hidden layers are meant to make our model sparse and address the gradient vanish or exploding issues.

Output layers: sigmoid, softmax, or tanh activations to generate probabilities of the input being in each class. We need specific activation functions to map the output of our neural network to the desired format.

(b) **(3 points)** 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.)

Solution:
$z_1 = W_1^T X = W_{11}X_1 + W_{12}X_2 + W_{10} = 0.9 \times 2 + 0.4 \times (-3) + 0 = 1.8 - 1.2 = 0.6$
$h_1 = \text{ReLU}(z_1) = 0.6$
$z_2 = W_2^T X = W_{21}X_2 + W_{22}X_2 + W_{20} = (-1.5) \times 2 + (-0.7) \times (-3) + 0 = -3 + 2.1 = -0.9$
$h_2 = \text{Sigmoid}(z_2) = 0.28905$
$\hat{y} = W_{31}h_1 + W_{32}h_2 + W_{30} = (-0.2) \times 0.6 + 1.6 \times 0.28905 \approx 0.34.$

(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 $\frac{\partial \mathcal{L}}{\partial \hat{y}}$? (**Hint.** Refer to the lecture slides for the definition of a binary cross entropy loss function)

Solution:
$\mathcal{L} = -y \ln(\hat{y}) - (1-y)\ln(1-\hat{y}) = -1 \times ln(0.34) - 0 = 1.0788.$
$\frac{\partial \mathcal{L}}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}}(-y \ln(\hat{y}) - (1-y)\ln(1-\hat{y})) = -\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}} = -\frac{1}{0.34} - 0 = -2.9412.$

(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. $\frac{\partial \mathcal{L}}{\partial W_{12}}$.

Solution:
$\frac{\partial \mathcal{L}}{\partial W_{12}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial W_{12}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_1} \frac{\partial h_1}{\partial z_1} \frac{\partial z_1}{\partial W_{12}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} W_{31} z_1 X_2 = -2.9412 \times (-0.2) \times 1 \times (-3) = -1.7647.$

(e) **(4 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.)

Solution: $3 \times 2 + 3 = 9$ parameters

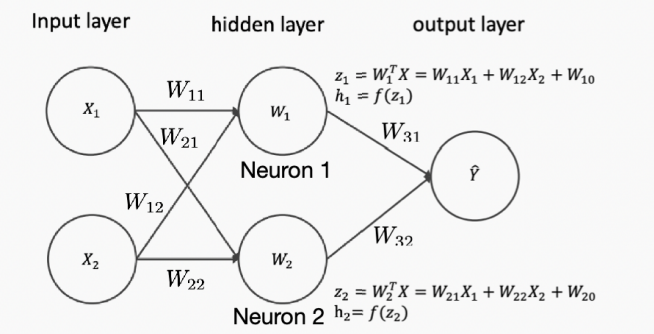

Figure 1: Neural Network

4 Multi-class Classification [10 points]

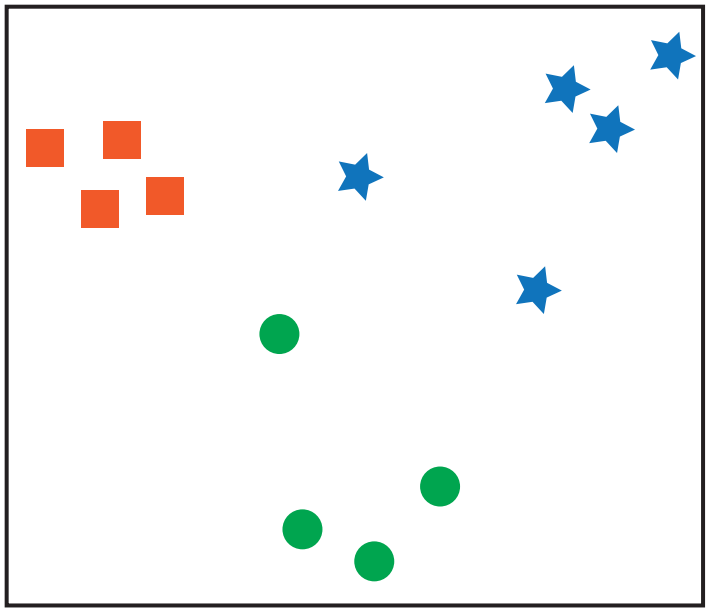

Figure 2: Multiclass Logistic Regression

(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.)

Solution: $21 * 5 = 105$

(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.

Solution: 3 lines for OvR where each line separates a class with all others. 2 lines for Multinomial, where each line separates a class with the reference class; any reference class is fine.

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) Decision Tree

Solution: (c) Decision boundary should be straight line

(2) Random Forest

Solution: (d) Random Forest looks at the average of deep decision trees (smaller straight lines)

(3) Neural Network (1 hidden layer with 10 ReLU)

Solution: (b) It should be piecewise linear due to the ReLU activiation function

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

Solution: (a) It should be curved due to the tanh activiation function

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.