10-601: Homework 3 Solution

Due: 9 October 2014 11:59pm (Autolab)
TAs: Henry Gifford, Jin Sun

Name: ______________________________
Andrew ID: ______________________________

Please answer to the point, and do not spend time/space giving irrelevant details. You should not require more space than is provided for each question. If you do, please think whether you can make your argument more pithy, an exercise that can often lead to more insight into the problem. Please state any additional assumptions you make while answering the questions. You need to submit a single PDF file on autolab. Please make sure you write legibly for grading.

You can work in groups. However, no written notes can be shared, or taken during group discussions. You may ask clarifying questions on Piazza. However, under no circumstances should you reveal any part of the answer publicly on Piazza or any other public website. The intention of this policy is to facilitate learning, not circumvent it. Any incidents of plagiarism will be handled in accordance with CMU's Policy on Academic Integrity.

⋆: Code of Conduct Declaration

- Did you receive any help whatsoever from anyone in solving this assignment? Yes / No.
- If you answered *yes*, give full details: ______________________ (e.g. *Jane explained to me what is asked in Question 3.4*)
- Did you give any help whatsoever to anyone in solving this assignment? Yes / No.
- If you answered *yes*, give full details: ______________________ (e.g. *I pointed Joe to section 2.3 to help him with Question 2*).

⋆: Notifications

This is the handout for theoretical questions in homework 3, you need to download the handout for programming part as well. If you have any questions, please post it on Piazza or email:

Henry Gifford: hgifford@andrew.cmu.edu

Jin Sun: jins@andrew.cmu.edu

1: Decision Boundaries and Complexity (TA:- Jin Sun)

(a) Figure 1 in appendix shows three decision boundaries. Please list **all possible** decision boundaries for the following classifiers. Please write down the labels. No explanations required.

[2pts for each. For each classifier, deduct all points unless the answer is perfectly correct.]

Decision Tree: (c)

Logistic Regression for binary classification: (a)

Perceptrons (Single-layer Neural Networks): (a)

Multi-layer Neural Networks (Single Hidden Layer): (a) (b) or (a) (b) (c) [Accept both answers]

[*8 points*]

Explanation:
LR and Perceptrons should have linear decision boundaries.
The decision boundaries for a decision tree must be perpendicular to one of the axes.
ANN with one hidden layer (with arbitrary number of hidden units) is capable to represent any Borel measurable functions. Borel measurable functions include all continuous functions and most of the discontinuous functions. It is okay if students fail to get (c), since this is beyond their scope.

(b) For the four classifiers mentioned in part(a), analyse the separability and complexity on several datasets. For separability, you need to state whether the classifier is able to perfectly separate the data points. For complexity, you only need to state whether the decision tree need to be a full tree (at each leaf node there is no attribute to split) to achieve best performance. Please refer to the appendix for detailed explanation on these datasets.

- Logic OR
- Logic XOR
- Majority
- Parity

[*12 points*]

[3pts for each dataset. For each dataset, deduct 1pt for each mistake, deduct all points if there are three mistakes or more. If students with incorrect answers provide some good insights, they can get partial credit (0.5pt)]

	DT	LR	Perceptron	ANN
OR	Perfect, Not full	Perfect	Perfect	Perfect
XOR	Perfect, Full	Not Perfect	Not Perfect	Perfect
Majority	Perfect, Not full	Perfect	Perfect	Perfect
Parity	Perfect, Full	Not Perfect	Not Perfect	Perfect

Explanation:
For separability, perceptrons and LR will fail if the data is not linear separable. OR dataset is linearly separable since there is only one point with label 0 positioning at a corner. Majority dataset is linearly separable since all label 0 strings will be close to the string with all 0s, and all label 1 strings will be close to the string with all 1s.
For complexity, the key question is that if all the bits are always required before classification. For example, in OR case, if the first bit in the string is 1, the final label must be 1 no matter what the values are for the rest of the bits. Thus the $X_1 = 1$ branch from root node does not need to split anymore. In majority, if more than half of the bits are 1, the rest of the bits will not change the label. For XOR and Parity cases, you need to know all the bits before knowing the label.

2: Activation Function (TA:- Jin Sun)

In lectures we use the logistic sigmoid function as the activation function for logistic regression and neural networks. However, there are many other activation functions such as linear function, hyperbolic tangent function and Gaussian function. In this homework, you need to derive the gradient on **one sample** for logistic regression using hyperbolic tangent function as activation function.

The hyperbolic function is defined as follows:

$$\tanh(z) = \frac{\sinh(z)}{\cosh(z)} = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{1}$$

and you should calculate the following term:

$$\frac{\partial Loss(\mathbf{w})}{\partial(\mathbf{w})} \tag{2}$$

Let's start with writing down the loss function on one sample for logistic regression:

$$Loss(\mathbf{w}) = -\ln P(Y = y|X = \mathbf{x}, \mathbf{w}) = -y \ln p - (1 - y)\ln(1 - p) \tag{3}$$

where $p = \tanh(z)$ and $z = \mathbf{w}^T\mathbf{x}$

And then you should derive the derivative and use the chain rule to get the final answer.

[*20 points*]

Total: 40

Solution: Chain rule: [3 pts]

$$\frac{\partial Loss(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial Loss(p)}{\partial p}\frac{dp}{dz}\frac{\partial z}{\partial \mathbf{w}}$$

Derive $\frac{\partial(Loss)}{\partial p}$: [3 pts]

$$\frac{\partial Loss(p)}{\partial p} = -\frac{y}{p} + \frac{1-y}{1-p} = \frac{p-y}{p(1-p)}$$

[$\frac{\partial Loss(p)}{\partial p}$ is the negative of $\frac{\partial L(p)}{\partial p}$, which is the derivative of log-likelihood we used in gradient ascent]

Derive $\frac{\partial \tanh(z)}{\partial z}$: [10 pts]

$$\frac{dp}{dz} = \frac{d\tanh(z)}{dz} = \frac{d}{dz}\frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{d}{dz}\frac{e^{2z} - 1}{e^{2z} + 1} = \frac{2e^{2z}(e^{2z}+1) - 2e^{2z}(e^{2z}-1)}{(e^{2z}+1)^2} = \frac{4e^{2z}}{(e^{2z}+1)^2} = (\frac{2e^z}{e^{2z}+1})^2$$

$$\frac{dp}{dz} = sech^2(z)$$

[Accept all forms that are equivalent, e.g. $1 - \tanh^2(z)$. Deduct 10pts if students fail to show the derivative of tanh function. They are supposed to derive the derivative themselves rather than copy it from the web.]

Derive $\frac{\partial z}{\partial \mathbf{w}}$: [3 pts]

$$\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \mathbf{w}^T \mathbf{x} = \mathbf{x}$$

Put everything together: [1 pt]

$$\frac{\partial Loss(\mathbf{w})}{\partial \mathbf{w}} = \frac{(p - y)\mathbf{x}}{p(1 - p)} sech^2(\mathbf{w}^T \mathbf{x})$$

3: Appendix

You do not need to include this page and the programming part into the pdf file for submission.

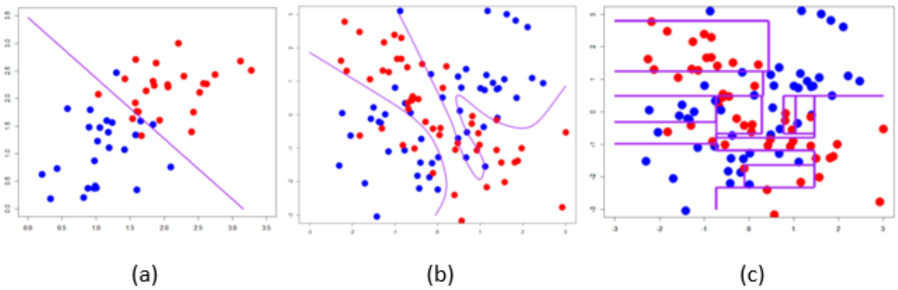

Figure 1: Decision Boundaries

In the datasets for problem 1(b), each sample is a binary string contains zeros and ones, and each bit is a feature. The length of the strings are at least 2 and same among all samples.

Logic OR

The label of each string is the logic OR value among all the bits.

Logic XOR

The label of each string is the logic exclusive OR value among all the bits. For example, if X_i is the ith digit in string X, the label is calculated by: X_1 XOR X_2 XOR $X_3 \ldots$

Majority

The label of each string is the digit with the most occurrence (either 0 or 1). The length of the strings is odd.

Parity

If the string has odd number of zeros, the label will be 1; if the string has even number of zeros, the label will be 0.