10-601: Homework 3

Due: 9 October 2014 11:59pm (Autolab)

TAs: Henry Gifford, Jin Sun

Name: Dawei Wang

Andrew ID: daweiwan@andrew.cmu.edu

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? No.
- Did you give any help whatsoever to anyone in solving this assignment? No.

⋆: 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 picture labels. No explanations required.

Decision Tree: c

Logistic Regression for binary classification: a

Perceptrons (Single-layer Neural Networks): a

Multi-layer Neural Networks (Single Hidden Layer): a, b

[*8 points*]

(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

Refer to the table below for separability.

Classifier	OR	XOR	Majority	Parity
Decision Tree	✓	✓	✓	✓
Logistic Regression	✓	✗	✓	✗
Single-Layer Neural Network	✓	✗	✓	✗
Multi-Layer Neural Network	✓	✓	✓	✓

As for complexity for decision tree - it has to be a full tree for XOR and Parity, since each bit could be decisive for the final result. In contrast, the decision tree can classify a sample once it sees a 1 among all its features (bits) in the OR case; or more than half number of 1's, in the Majority case.

[*12 points*]

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*]

Evaluation:

$$\begin{aligned}\frac{\partial Loss(\mathbf{w})}{\partial(\mathbf{w})} &= \frac{\partial Loss(\mathbf{w})}{\partial p}\frac{\partial p}{\partial z}\frac{\partial z}{\partial \mathbf{w}} = \left(-\frac{y}{p} + \frac{1-y}{1-p}\right)\left(1 - \tanh^2 z\right)\mathbf{x} \\ &= \frac{p-y}{p}\left(1 + \tanh z\right)\mathbf{x} = \left(1 - y\coth \mathbf{w}^T\mathbf{x}\right)\left(1 + \tanh \mathbf{w}^T\mathbf{x}\right)\mathbf{x}\end{aligned} \tag{4}$$

Total: 40

3: Appendix

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

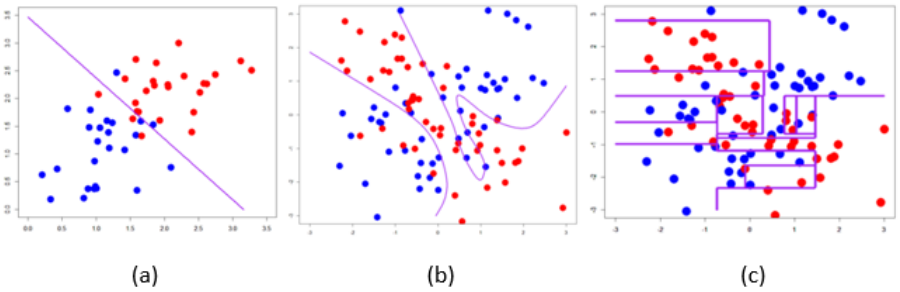

Figure 1: Decision Boundaries

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. For example, if X_i is the ith digit in string X, the label is calculated by: X_1 OR X_2 OR $X_3 \ldots$

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. For example, string "11110" has more 1s than 0s, so its label is 1.

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. For example, string "101010" has 3 zeros, its label is 1.