ECE523: Engineering Applications of Machine Learning and Data Analytics

I acknowledge that this exam is solely my effort. I have done this work by myself. I have not consulted with others about this exam in any way. I have not received outside aid (outside of my own brain) on this exam. I understand that violation of these rules contradicts the class policy on academic integrity.

Name: ______________________________

Signature: ______________________________

Date: ______________________________

Instructions: There are five problems. Partial credit is given for answers that are partially correct. No credit is given for answers that are wrong or illegible. Write neatly.

Problem 1: ___________

Problem 2: ___________

Problem 3: ___________

Problem 4: ___________

Problem 5: ___________

Total: ___________

Problem #1 – Neural Networks (10 Points)

(a) One method for preventing the neural networks' weights from overfitting is to add regularization terms. You will now derive the update rules for the regularized neural network. Recall that the non-regularized gradient descent update rule for w_1^{t+1} is:

$$w_{ji}^{t+1} = w_{ji}^{t} + \eta \sum_{n=1}^{N} e_j(n)\phi'\left(v_j(n)\right) y_i(n)$$

Derive the update rule for w_{ji}^{t+1} in the regularized neural net loss function which penalizes based on the square of each weight. Use $\lambda \geq 0$ to denote the regularization parameter. Use the following regularizer:

$$R(w) = \lambda \sum_{i} w_i^2$$

Re-express the regularized update rule so that the only difference between the regularized setting and the unregularized setting above is that the old weight w_{ji}^{t} is scaled by some constant. Explain how this scaling prevents overfitting.

(b) The definition of a sigmoid function is given by

$$\phi(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

Show that $\phi'(x) = \phi(x)(1 - \phi(x))$

Problem #2 – AdaBoost (10 Points)

In class, we discussed the Adaboost algorithm in four general steps and each step was relatively general. Recall that you were told that the weight for a hypothesis h_t in the Adaboost algorithm was given by

$$\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$$

Some discussions about Adaboost may simply say "*Choose* α_t" without giving a closed form expression for α_t. For this problem, prove that the expression shown above is the ideal value for α_t and justify your response.

Problem #3 – True/False: A Gamblers Ruin (10 Points)

[True/False] (1 point): The theory behind AdaBoost proves that the error on the testing data is upper bounded by

$$\widehat{\text{err}}(H) \leq 2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t(1-\varepsilon_t)}$$

[True/False] (1 point): The gradients cannot explode in an RNN, which is a desirable property of the backpropagation through time.

[True/False] (1 point): The multi-armed bandit addresses problems that require exploration of new arms and exploitation of the ones we know perform well.

[True/False] (1 point): Classification error is typically the best way to measure the performance of an RNN language model.

[True/False] (1 point): One of the disadvantages of deep learning with auto-encoders is that we need a large volume of labeled data to train each layer.

[True/False] (1 point): In the context of a adversarial MAB, the term $\gamma \in [0, 1]$ controls the trade-off between the estimated reward of the arm and pure exploration.

$$\widehat{p}_i(t) = \gamma \frac{w_i(t)}{\sum_j w_j(t)} + (1-\gamma)\frac{1}{K}$$

where K is the number of arms and $w_i(t)$ is the weight of the ith arm at time t.

[True/False] (1 point): A neural network will (likely) find a local minimum for its optimization problem and the same is true for a support vector machine.

[True/False] (1 point): Using a sigmoid activation function in a neural network trained with backpropagation is one way to avoid the vanishing gradient problem.

[True/False] (1 point): A discriminator network, D, after enough training in a GAN will always be able to identify if a sample came from the data set or the generator network, G.

[True/False] (1 point): In backpropagation, the only difference between updating a hidden node versus an output node is how the local gradient is calculated.

Problem #4 – Adaboost (10 Points)

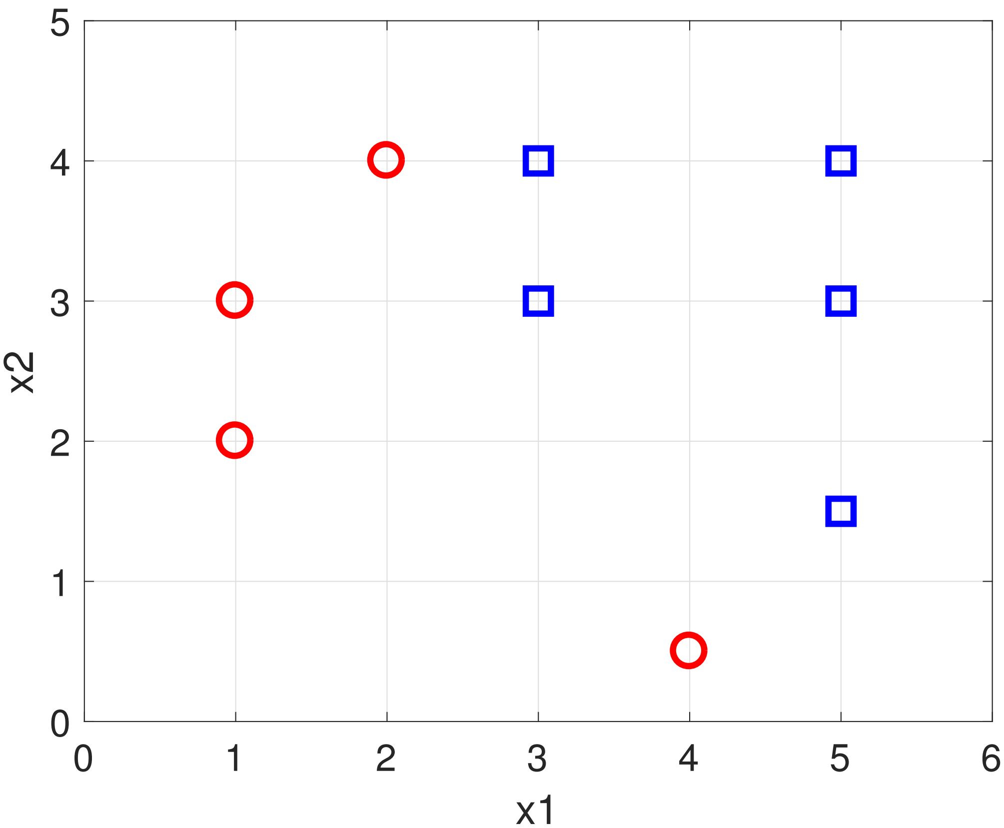

Figure 1: Labeled training points for Problem 2.

Consider the labeled training points in Figure 1, where ∘ and □ denote positive and negative labels, respectively. We wish to apply AdaBoost with a threshold classifier (i.e., pick an axis then pick a threshold to label the data). In each boosting iteration, we select the threshold that minimizes the weighted training error, breaking ties arbitrarily. Use the AdaBoost pseudo-code to help with this question

1. In Figure 1, draw a decision boundary on x_1-axis (i.e., vertical line) corresponding to the first threshold that the boosting algorithm could choose. Label this boundary (1), and also indicate +/- side of the decision boundary.

2. In the same figure also circle the point(s) that have the highest weight after the first boosting iteration.

3. What is the weighted error of the first threshold after the first boosting iteration, i.e., after the points have been re-weighted?

4. Draw a decision boundary corresponding to the second threshold using the weights, again in Figure , and label it with (2), also indicating the +/- side of the boundary. For clarity grading exams draw a decision boundary on x_1 (i.e., vertical line).

Problem #5 – Random Short Answer (10 Points)

(SA:1) Describe the process of learning and testing a random forest on a data set with n samples and p features.

(SA:2) What is an appropriate way to train a deep neural network? The key word in that sentence is "appropriate".

(SA:3) Explain the differences between an adversarial bandit and stochastic bandit. Also, describe the concept of regret.

(SA:4) Explain the difference between backpropagation and backpropagation through time.