CIS 472/572, Winter 2018

Written Homework 3: SVMs and Neural Networks

DUE DATE: Friday, February 23rd at 11:00pm

1. Let $D = \{(x_i, y_i)\}_{i=1}^{5} = \{(-2,-1), (-1,-1), (0.5,-1), (1,+1), (2,+1)\}$.

   (a) Describe the parameter space of w and b for which the line $wx + b$ separates the positive and negative classes. (Hint: There are an infinite number of solutions — mathematically describe the entire set of solutions.)

   (b) Write down the parameters (w and b) for the maximum-margin separator.

   (c) Consider the soft-margin SVM, where margin violations are penalized but not prohibited. Compute the optimal parameters (w,b) for the soft-margin SVM using the following slack penalties (C).

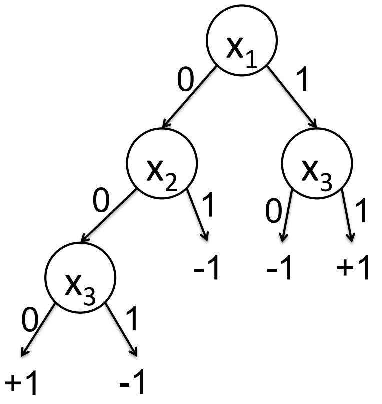
   HINT: Choose which points will be allowed to violate the margin constraint and find the max-margin separator among the remaining points. Repeat with different sets of points and choose the separator that has the lowest regularized hinge loss.

   i. $C = 5$

   ii. $C = 0.5$

   iii. $C = 0.05$

2. (a) Construct a neural network that represents the same classification function as the decision tree below. Specify the overall network structure, the activation function, and the parameters of each node.



   (b) Describe a general method for converting a decision tree to a 2-layer neural network that represents the same classification function. You may assume that the label and all attributes are binary.

(c) Consider a set of k decision trees, each of which can predict a label (+1 or -1) for any instance x. Instead of using the prediction of any one tree, we could instead use all of them and predict the label that was most common. This is often referred to as an *ensemble*. For example, if there are 21 trees, 13 of them predict +1, and 8 of them predict -1, then the overall prediction for the ensemble is +1.

Describe a method for converting an ensemble of decision trees into a single neural network that makes the same predictions. The resulting neural network should be linear in the total size of all decision trees in the ensemble.

3. **Grads only.** In order to minimize some cost functions in machine learning, a first step is to find the gradient of the cost function given the data and the current parameters. Please derive the closed-form gradient of multi-class logistic regression for the observed data point/label pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbf{R}^d$ and $y_i \in \{1, 2, \ldots, C\}$ (Hint: for multi-class logsitic regression we have: $p(Y = i | x, w_1, \ldots, w_C, b_1, \ldots, b_C) = \frac{e^{w_i^T x + b_i}}{\sum_{k=1}^{C} e^{w_k^T x + b_k}}$)