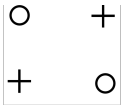
CSE 417T (Machine Learning): Exam 2 Practice Questions

1. Consider the following dataset:



   Which of the following classifiers will achieve zero training error on this data set. (There might be more than one.)

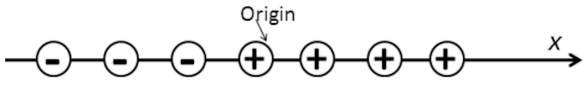
   (a) Logistic regression
   (b) SVM (polynomial kernel)
   (c) Depth-2 ID3 decision trees
   (d) 3-NN classifier
   (e) Neural networks with a single hidden layer

2. Assume we have a decision tree to classify binary vectors with 100 features. Can you specify a 1-Nearest Neighbor model that would result in exactly the same classification of all examples as the decision tree? Explain how, or provide a counter-example.

3. Suppose you are using decision stumps as the weak learners in AdaBoost, but you know that all your features $x_i \in \{+1, -1\}$ (that is, you're solving a Boolean problem, and all features, as well as the class, are either true or false). Show that the final hypothesis output by AdaBoost after T rounds of boosting is a linear classifier.

4. Using a training data set, you compute a (1-)nearest neighbor classifier and apply it to a held out test data set. You find that this hypothesis achieves a test error of E_{test}. You know that the nearest neighbor model suffers from high variance so you decide to apply bagging. However, you observe that when the number of bags is very large, the test error of the aggregated hypothesis is also exactly E_{test}. Explain why this occurs.

5. Given the following dataset in 1-d space, which consists of 4 positive data points $\{0, 1, 2, 3\}$ and 3 negative data points $\{-3, -2, -1\}$:

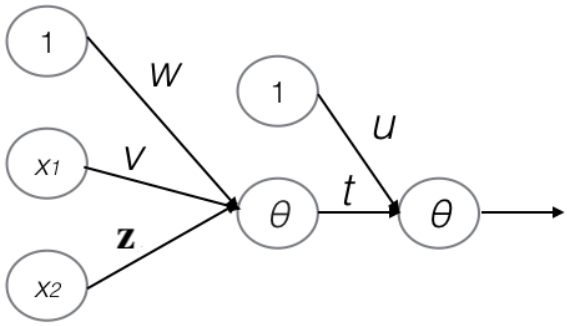


   Suppose that we want to learn a soft-margin linear SVM for this dataset. Consider the choice of the slack penalty term (i.e., C). Answer the following two questions:

(a) If $C \to \infty$, how many support vectors do we have for the final hypothesis? Explain why.

(b) If $C = 0$, how many support vectors do we have for the final hypothesis? Explain why.

6. Consider the following neural network with one hidden layer.



(a) Assume we use the linear function $\theta(s) = s$ as the activation function for both the hidden layer and the output layer. Draw a neural network with no hidden layers (but with the same input and output layers) that is equivalent to the given one. You need to clearly specify the network structure and the weights.

(b) Assume we use the sigmoid activation function $\theta(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$. Can we still construct a neural network with no hidden layers that is equivalent to the given network? If yes, draw it. If no, please provide brief justifications.

Can we construct a neural network with one hidden layer and with *sign* function as the active function to represent the above classification function? If yes, draw it. If not, explain why not.

7. Suppose you apply bagging and boosting to linear classifiers. For which of the two will the output ensemble method also be a linear classifier?

◯ Boosting only.

◯ Bagging only.

◯ Both boosting and bagging.

◯ Neither boosting nor bagging.

8. Since the VC dimension of SVM with Gaussian RBF Kernel is infinite, such an SVM will have worse out-of-sample error than an SVM with polynomial kernel with finite VC dimension.

◯ True

◯ False

9. Consider a data point that is correctly classified and distant from the decision boundary. For logistic regression and SVMs, which ones would be affected by a small perturbation to this data point, assuming it still remains far from the boundary?

◯ Logistic regression only.

◯ SVM only.

◯ Both logistic regression and SVM.

◯ Neither logistic regression nor SVM.

10. What is one reason that you might prefer a single decision tree over a random forest for a particular problem?

    ◯ Decision trees are more interpretable than random forests

    ◯ Decision trees are less likely to overfit than random forests

    ◯ Decision trees have lower variance than random forests

    ◯ None of the above

11. Which of the following techniques help prevent Neural Networks from overfitting?

    ◯ Early stopping

    ◯ Dropout

    ◯ Data augmentation (adding noises to existing data points)

    ◯ All of the above