HMC CS 158, Fall 2017
Problem Set 7 Exercises: Boosting

Goals:

- To test your understanding of boosting basics.
- To investigate choices in and extensions of the boosting algorithm.

1 Boosting [10 pts]

(a) **(1 pts)** Give a one-sentence reason for why AdaBoost outperforms a single decision stump.

(b) **(3 pts)** Consider building an ensemble of decision stumps with the AdaBoost algorithm. Figure 1 displays the labeled points in two dimensions as well as the first stump we have chosen. The little arrow in the figure is the normal to the stump decision boundary indicating the positive side where the stump predicts +1. All the points start with uniform weights.

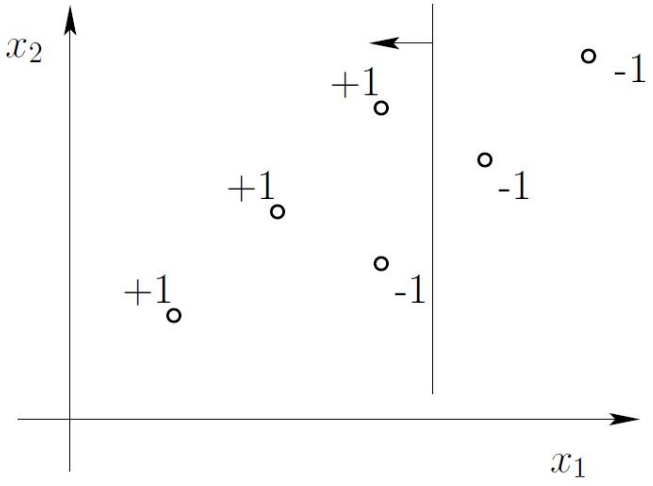

Figure 1: Boosting.

i. **(1 pts)** Circle all the point(s) in the figure whose weight will increase as a result of incorporating the first stump.

ii. **(1 pts)** Draw in the same figure a possible stump that we could select at the next boosting iteration. You need to draw both the decision boundary and its positive orientation (as in the figure for the first stump).

iii. **(1 pts)** Will the second stump receive a higher coefficient in the ensemble than the first? In other words, will $\alpha_2 > \alpha_1$? Briefly explain your answer. (No calculation should be necessary.)

Parts of this assignment are adapted from course material by Jenna Wiens (UMich) and Tommi Jaakola (MIT).

(c) **(2 pts)** Suppose that you have many boosted classifiers that correctly classify your training set. Is there an advantage to picking the classifier that minimizes exponential loss $Loss(h) = \sum_{i=1}^{n} \exp\left(-y^{(i)} h(\boldsymbol{x}^{(i)})\right)$, and if so, why?
Hint: Plot the exponential loss function. How does it vary for different values of $z = yh(\boldsymbol{x})$? What does this imply about generalization error?

(d) **(4 pts)** Let us now consider replacing decision stumps with different base classifiers.

i. **(2 pts)** Suppose the training set is linearly separable, and we use a hard-margin linear support vector machine (no slack) as a base classifier. In the first boosting iteration, what would the resulting $\hat{\alpha}_1$ be?

ii. **(2 pts)** Let us consider simple base learners but still those that can be combined into strong predictors. So, we define our base learners as radial basis functions centered at particular training points, i.e., $h(\boldsymbol{x}; \theta) = y^{(i)} \exp(-\beta ||\boldsymbol{x} - \boldsymbol{x}^{(i)}||^2)$, where the parameter θ simply identifies the training point i. These base learners return real-valued predictions $h(\boldsymbol{x}; i) \in [-1, +1]$. Provide an argument for why an ensemble with n such base learners can in principle classify a set of n points in all possible ways.