

Homework 2, Part II and clarifications.

**Bagging and Boosting Decision Trees:** Try bagging and boosting decision trees to recognize the digit “8”. Experiment with a few different “bag sizes” (i.e., size of training sample used in each round of bagging or boosting), such as 10% of all training data, 20%, 50%, etc. to see which works the best.

Note: When you compile C4.5 you may get several compiler warnings. These should be safe to ignore.

**Margins:** After boosting for  $T$  iterations and evaluating the ensemble classifier on the test set, calculate the margin of each example in the test set, and plot the cumulative margin distribution; that is, for each margin value between 0 and 1 (on the  $x$ -axis), plot (on the  $y$ -axis) the number of examples in the test set that have margin less than or equal to this value. Do this for each boosting run ( $T = 1, 5, 10, 50, 100$ ).

**Optional:** Repeat the boosting runs and margins analysis for two harder concepts: “digit 8 or digit 4 = true, everything else = false” over the entire optdigits training set; and “digit 4 = true, digit 9 = false”, for the data limited to examples of 4 and 9.

**ROC curves:** We’ll do something on this in a different assignment.

**What to turn in:**

Writeup (see below)

Well-commented code for Bagging, Boosting, and creating margin distributions

**Send to karan@pdx.edu by 5pm, Feb. 16.**

**Results to give and questions to answer in writeup:**

- Give the classification error on the test data for the SVM classifier (i.e., the ensemble 10 SVMs ) using (a) linear kernels; (b) polynomial kernels; (c) radial basis function kernels; (d) sigmoid kernels. If one of the kernel types works better than another kernel type, any ideas why? (Or if not, why not?)

- Give confusion matrices for the SVM classifier performance on the test data for each of the kernel types. Summarize the results in these matrices.
- For the decision-tree digit classification task(s) with bagging and boosting, give the classification error and confusion matrix of the final ensemble on the test data for  $T=1,5,10,50,100$  iterations of Bagging and Boosting.
- Give the margin-distribution plots as described above for your boosting runs. Summarize how the value of  $T$  affects the margin distribution.

$$\text{margin}(\mathbf{x}, y) = \frac{y \sum_t \alpha_t h_t(\mathbf{x})}{\sum_t \alpha_t}$$

- Add any additional observations you found interesting in your results.