CSE 417T: Homework 5

Due: November 18 (Sunday), 2018

Notes:

- You may work in groups of up to two persons on this homework. Each group only needs to submit one copy of thee homework. Check the following link on how to submit group assignments on Gradescope:
  `https://www.youtube.com/watch?v=a6DERS94qPY&feature=youtu.be&t=32s`
- Please check the submission instructions for Gradescope provided on the course website. You must follow those instructions exactly.
- Please download the following stub Matlab files for this homework.
  `http://classes.cec.wustl.edu/~cse417t/hw5/hw5_files.html`
- Homework is due **by 11:59 PM on the due date.** Remember that you may not use more than 2 late days on any one homework, and you only have a budget of 5 in total.
- Please keep in mind the collaboration policy as specified in the course syllabus. If you discuss questions with others you **must** write their names on your submission, and if you use any outside resources you **must** reference them. **Do not look at each others' writeups, including code.**
- Please do not directly post your answers on Piazza even if you think they might be wrong. Please try to frame the question such that you dont give the answers away. If there is specific information you want to ask about your answers, try the office hours or private posts on Piazza.
- Please comment your code properly.
- There are 2 problems on 2 pages in this homework.

Problems:

1. (50 points) The purpose of this problem is to write code for bagging decision trees and computing the out-of-bag error. You may use matlab's inbuilt `fitctree` function, which learns decision trees using the CART algorithm (read the documentation carefully), but do not use the inbuilt functions for producing bagged ensembles. In order to do this, you should complete the stub `BaggedTrees` function. Note that it only returns the out-of-bag error. You may want to use other functions that actually construct and maintain the ensemble. You may assume that all the **x** vectors in the input are vectors of real numbers, and there are no categorical variables/features. You will compare the performance of the bagging method with plain decision trees on the handwritten digit recognition problem (the dataset is in `zip.train` and `zip.test`, available from `http://amlbook.com/support.html`.[1])

[1] Check the links to "training set" and "test set".

We will focus on two specific problems distinguishing between the digit one and the digit three, and distinguishing between the digit three and the digit five. Here are the steps for this problem:

a Complete the implementation of `BaggedTrees`. You may choose any reasonable representation that you wish; the two strict requirements are that you plot the out-of-bag error as a function of the number of bags from 1 to the number specified as input (`numBags`), and that you return the out-of-bag error for the whole ensemble of `numBags` trees. Include the plots (with clearly labeled axes) in your writeup, and, of course, submit your code.

b Run the provided `OneThreeFive` script, which creates training datasets based on the one-vs-three and three-vs-five cases we are interested in, and calls both the in-built decision tree routine and your bagging code, printing out the cross-validation error for decision trees and the OOB error for your bagging implementation. Report the results in your writeup.

c Now, learn a single decision tree model for each of the two specified problems (one-vs-three and three-vs-five) on the training data, and test their performance on `zip.test` what is the test error? Similarly, learn a single ensemble of 200 trees on the training data for each of the two specified problems and test the performance of the ensembles on the test data. Report your results.

d Summarize and interpret your results in one or two concise paragraphs as part of your writeup.

2. (50 points) Implement AdaBoost using decision stumps learned using information gain as the weak learners (you may use the `fitctree` function to implement the weak learner. Look at the "deviance" split criterion), and apply this to one-vs-three and three-vs-five problems (as described in Question 1) on the `zip.train` and `zip.test` data. In order to do this, you should complete the stub `AdaBoost` function. Graphically report the training set error and the test set error as a function of the number of weak hypotheses, and summarize and interpret your results.