CptS 570 Machine Learning, Fall 2017
Homework #4
Due Date: Nov 16

NOTE 1: Please use a word processing software (e.g., Microsoft word or Latex) to write your answers and submit a printed copy to me at the begining of the class on Oct 6. The rationale is that it is sometimes hard to read and understand the hand-written answers. Thanks for your understanding.

NOTE 2: Please ensure that all the graphs are appropriately labeled (x-axis, y-axis, and each curve). The caption or heading of each graph should be informative and self-contained.

1. (**5 points**) Please read the following paper and briefly summarize the key ideas as you understood:

   Thomas G. Dietterich (1995) Overfitting and under-computing in machine learning. Computing Surveys, 27(3), 326-327.
   `http://www.cs.orst.edu/~tgd/publications/cs95.ps.gz`

2. (**15 points**) Please read the following paper and briefly summarize the key ideas as you understood:

   Thomas G. Dietterich (2000). Ensemble Methods in Machine Learning. J. Kittler and F. Roli (Ed.) First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science (pp. 1-15). New York: Springer Verlag.
   `http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf`

3. (**20 points**) We need to perform statistical tests to compare the performance of two learning algorithms on a given learning task. Please read the following paper and briefly summarize the key ideas as you understood:

   Thomas G. Dietterich: Approximate Statistical Test For Comparing Supervised Classification Learning Algorithms. Neural Computation 10(7): 1895-1923 (1998) `http://sci2s.ugr.es/keel/pdf/algorithm/articulo/dietterich1998.pdf`

4. (**30 points**) Implementation of Expectation Maximization (EM) algorithm and experimentation.

   You will implement the EM algorithm and Gaussian Mixture Models (GMMs) to cluster the data into k clusters. As we discussed in the class, we assume a GMM with k components for the data and we find the parameters of the model using maximum likelihood estimation.

   Implement the EM algorithm for one-dimensional GMMs (assume that each data point has only one feature). You are provided with one-dimensional dataset. Use the algorithm to cluster the data. Run the algorithm multiple times from a number of different initialized values (random) and pick the one that results in the highest log-likelihood.

   Run the algorithm for different values of k (3, 4, 5) and report the parameters you get for each value of k.

   You can use WEKA (http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html) to debug your implementation.

5. (**30 points**) Empirical analysis of ensemble methods. You will use the Weka: `http://www.cs.waikato.ac.nz/ml/weka/downloading.html` software. Please use the dataset provided for this question.

   a. Bagging (weka.classifiers.meta.Bagging). You will use decision tree as the base supervised learner. Try trees of different depth (1, 2, 3, 5, 10) and different sizes of bag or ensemble, i.e., number of trees (10, 20, 40, 60, 80, 100). Compute the training accuracy and testing accuracy for different combinations of tree depth and number of trees; and plot them. List your observations.

   b. Boosting (weka.classifiers.meta.AdaBoostM1). You will use decision tree as the base supervised learner. Try trees of different depth (1, 2, 3) and different number of boosting iterations (10, 20, 40, 60, 80, 100). Compute the training accuracy and testing accuracy for different combinations of tree depth and number of boosting iterations; and plot them. List your observations.

Instructions for Code Submission.

Please follow the below instructions. It will help us in grading your programming part of the homework. We will provide a dropbox folder link for code submission.

- Mention the programming language and version (e.g., Python 2.5) that you used.
- Submit one folder with name WSUID-LASTNAME.zip (e.g., 111222-Fern.zip) and include a README file.
- Include a script to run the code and it should be referred in the README file.
- Don't submit the data folder. Assume there is a folder "data" with all the files.
- Output of your programs should be well-formatted in order to answer the empirical analysis questions.
- Structure your code meaningfully and add comments to make it readable.