10-701/15-781 Machine Learning, Fall 2003

Homework 3
Out: October 14, 2003 **Due: start of class October 28, 2003**

If you have questions, please contact Jiayong Zhang `<zhangjy@cs.cmu.edu>`.

This assignment gives you an opportunity to formulate and solve a machine learning problem using a real-world data set. You may work with either a text data set or with a data set containing images of former Machine Learning students.

What you will do.

- Find a partner to work on this project (two person teams are encouraged, though you may work alone if you prefer. No three person teams please.)
- Select your data set: text or images. The text data and a NaiveBayes classifier package is available at

  `www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/mlc/hw3/text/`

  The image data and a neural network package is available at

  `www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/mlc/hw3/face/`

  The `README` file under each directory provides the necessary information on the data and associated codes to get you started.
- Define your learning task. The text data consists of 20,000 online newsgroup postings taken from twenty newsgroups, with each text document labeled by the newsgroup it came from. Therefore, classifying documents by newsgroup is one possible learning task. The image data contains approximately 30 different images from each of twenty people, with each image labeled by several properties including the person ID, whether they are wearing sunglasses, etc. Therefore, classifying images to determine whether the person is wearing sunglasses is one possible choice.
- Perform the work. As a guideline, we expect each student to spend 7–12 hours on this homework over the course of two weeks (remember you are working in pairs, so you can do a fairly substantial project).
- Turn in a 2–4 page write-up. Your write-up should describe *precisely* your learning task(s), your learning method(s) including how you represented the data for input to the learner, your experiments, results, and any conclusions your draw from this. The clarity and content of your write-up will have a primary impact on your grade.

The reports must not be more than 4 pages, 11-point font, including figures. Each two-person team must hand in a single write-up.

Grading and determining when you have done enough.

A good strategy for this homework is to divide your effort into two one-week blocks. During week one you might complete training a classifier for text or image classification using the provided code, and experimenting with some with the parameters to determine what works best. Then during the second week you could make up your own follow-on question to study (perhaps trying a competing method, or studying some issue you noticed during the first week). A project that does a solid job applying the given code and carefully evaluating and describing it might get 75–80% credit. A project that in addition pursues an interesting second approach or alternative problem might get 90–100% credit.

Be creative! Exploring your own interesting ideas and comparing them with the baseline approaches will receive credit whether they beat the baseline or not.