The E Corp Challenge

The Intro

E Corp, one of the largest multi-nationals in the world, has provided us with two data sets [luckily they use Brightspace as well]. For various reasons, but particularly due to the sensitivity of the application involved, the company is unwilling to reveal anything else about the actual application to us than that it is "part of a global research effort that aims at the eradication of attacks by hackers and similarly subversive elements."

The Data

The problem is a two-class classification problem in 204 dimensions. The small *labeled* training set [train.csv; *note that the first column contains the labels*] has a size of 50 samples per class. The *unlabeled* data [test.csv] provides 20,000 additional samples. These samples have, however, undergone a corruption of some sort. The only additional information we have regarding this corruption, is that it is additive uniform noise and that the noise comes from a fixed p-norm ball, $\|x\|_p \leq r$, where both p and the radius r are unknown.

The Challenge

We are looking for the best performing classifier on the unlabeled data. The evaluation measure on this test set is the *error rate* [$= 1 -$ accuracy].

The "What to Submit to Brightspace?"

1. You are expected to provide a brief report on your work. The work should be done individually; you are the sole responsible for your own end report. The report should contain, at the least, the following :

- Your name, netid, and student number;
- Your estimate of the performance on the 20,000 unlabeled data set;
- A clear description of the complete and best-performing method you implemented;
- Clear arguments for the different choices you made in building your classifier;
- Sufficient experimental results, learning curves, error bars, or whatever else you need E Corp that little improvement will be possible beyond the system that you present;
- Any references that have been used.

Note : *your report should not be more than 2000 words!* Please include [an estimate of] the number of words you used in your report.

2. You should also provide a csv file with the exact filename label.csv in which you provide the 20,000 labels on the test data your procedure gives. Every label in row n of your csv file is assumed to be your estimated label associated to the sample in row n from test.csv. An example label.csv file is provided.

The Grading

You grade for this final assignment will be based on your report, your error estimate, and the actual error rate that you achieve on the test set provided.

Don't call us; we call you…