The E Corp Challenge

Introduction

E Corp, one of the largest multi-nationals in the world, has provided us with two data sets. For various reasons, but particularly due to reasons of privacy, the company is not allowed to reveal much more about the actual application to us than that it is "part of a global effort that aims at making this world a better place."

Data

We are dealing with a two-class classification problem that revolves around so-called fairness. Our objects are people.

The training data [train.csv] contains two sensitive features [in columns 13 and 14]. All in all, there are 20,000 feature vectors in 14 dimensions. The last column, column 15 in the data file, provides the class labels.

Feature 14 is highly sensitive and cannot be used at test time. Therefore, the test set [test.csv that is also available] only has 13 columns, i.e., sensitive Feature 14 and the labels are left out. The test data consists of 28,000 feature vectors.

It should be noted that the first 6 features are numerical, i.e., their relative values tells us something. All other features are categorical.

The Challenge

Feature 13 is a feature through which some form of [so-called] fairness has to be enforced. The law prohibits that people with a different categorical value for Feature 13 [which equals 1 or 2] are treated differently in the following more precise sense.

Let X_{13} be the value that feature 13 takes on, let L be the possible label [which also equals 1 or 2], let c be a classifier, and denote by $c(X)$ its corresponding output [based on the 13 features available at test time]. Any such classifier should approximately achieve the following two constraints :

$$P(c(X) = 2|L = 1, X_{13} = 1) = P(c(X) = 2|L = 1, X_{13} = 2)$$
$$P(c(X) = 1|L = 2, X_{13} = 1) = P(c(X) = 1|L = 2, X_{13} = 2).$$

L 是实际标签, c(x)是预测标签
所以该限制条件即是当固定某一标签的情况下，被分类成另一类（错误分类）的概率，对不同的feature 13 是同等的

E Corp is looking for a classifier that gives a smallest overall error rate on the test set, while fulfilling the above constraints as well as possible. All in all, the management team leading this project has decided on the following loss to be minimized :

$$3\max(p_{2|11}, p_{2|12}) + \max(p_{1|21}, p_{1|22}),$$

where $p_{i|jk}$ is an estimate for $P(c(X) = i|L = j, X_{13} = k)$ based on the standard maximum likelihood estimate for $P(c(X) = i, L = j, X_{13} = k)$.

What to Submit through Brightspace

1. You should provide a csv file with the filename labels.csv [yes, exactly that name!]. In it, you give the 28,000 labels on the test data your procedure gives. Every label in a row of your csv file is assumed to be your estimated label associated to the sample in the corresponding row of test.csv. An example labels.csv file is provided.

2. You are expected to provide a brief report on your work. The work should be done individually : you are the sole responsible for your report. The first page of your report should contain the following information :

- Your name;
- Your netID;
- Your student number;
- Your final estimate for $3\max(p_{2|11}, p_{2|12}) + \max(p_{1|21}, p_{1|22})$ on the unlabeled data set.

For the rest, the report should contain

- A clear description of the complete and best-performing method you implemented;
- Clear arguments for the different choices you made in building your classifier;
- Experimental results, learning curves, error bars, or anything else you need to convince 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 the number of words you used in your report.

Grading

Your grade for this final assignment will be based for 70% on the estimate you provide and the actual loss that you achieve on the test set. The rest of the report will count, for 30%, toward your grade.