Homework 6: Logistic Regression

Fundamentals of Data Science 2016/2017

Your goal is to build a script that performs a logistic regression on a training dataset, and uses it to predict labels on a test dataset. The script must be named `ID.py` where as usual `ID` is your student ID. The script will accept as command-line options the names of two files:

```
$ python ID.py train.csv test.csv
```

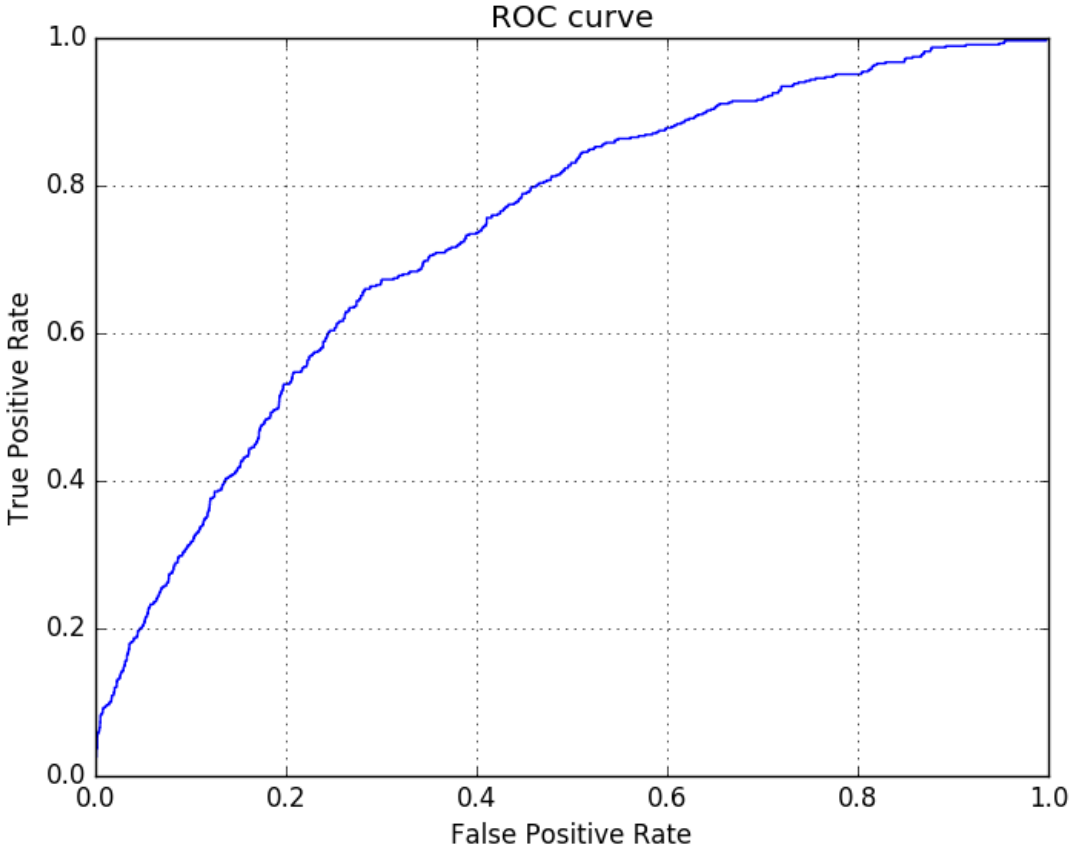
Each file will contain a dataset in CSV format. The datasets may have any number of rows. The first row contains the name of the columns. The first dataset has one column more than the second; this is the last column, which contains the binary labels. All other variables are, in both datasets, in the same column order. The datasets can contain NaNs, and the script must drop all rows containing a NaN. You can assume that all variables are already normalized. The script should build a logistic regression model on the training dataset, and use it to predict probabilities on the test dataset. Do not forget to include a constant variable for the intercept term.

The script should output:

1. the vector θ of parameter estimates obtained through the logistic regression
2. the ROC curve obtained on the training dataset, on a file called `ID-roc.png`, in PNG format
3. the AUC value obtained on the training dataset
4. the scores predicted for the observations in the test dataset

Small deviations are not an issue (e.g. if your AUC is 0.68 but I obtained 0.70). As an example, using all but 20 rows of the framingham dataset as training set, and the remaining 20 rows as test set, I obtained:

```
$ ipython 00000.py fram-train.csv fram-test.csv
[-2.    0.27  0.54 -0.05  0.03  0.21  0.03  0.05  0.1   0.01  0.11  0.34
 -0.04  0.03 -0.04  0.16]
0.735813707018
[ 0.12  0.09  0.03  0.15  0.04  0.12  0.08  0.31  0.22  0.07  0.26  0.03
  0.29  0.51  0.04  0.08  0.73  0.49  0.3   0.13]
```


ROC curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0