Proofpoint Data Scientist Interview Assignment

1 Overview

In this assignment, you must create a multi-class classifier using the provided training set and produce class predictions for the test set. Please use either Python, Matlab, or R and any libraries you would like. Python is strongly preferred.

2 Data Description

- Classes: 3
- Features: 903
  - The first three features are categorical, containing string values
  - The remaining features are numeric (real-valued)
- Training instances: 140
- Testing instances: 560

3 Evaluation

Your submission will be evaluated in two ways:

- Quantitative: your class predictions on the test set will be scored using standard accuracy scoring:

  $$ACC = \#correct/\#instances$$

  For example, if this is your confusion matrix:

	Predicted 0	Predicted 1	Predicted 2
True Class 0	40	24	19
True Class 1	1	313	4
True Class 2	1	4	154

then the accuracy is: $(40 + 313 + 154)/560 = 0.905$

- Qualitative: you should provide any source code you used and a brief description of your approach.

4 Data Files

- `train_features.csv`: A comma-delimited csv file with one training example per line
- `train_labels.csv`: The labels for each training example, one per line. The i-th label in this file corresponds to the i-th instance in the training features file.
- `test_features.csv`: A comma-delimited csv file with one testing example per line

5 Results Submission

Please send your submission in a single archive (.zip or .tar.gz) with the following contents:

- `test_predictions.csv`: This file contains your class predictions on the test set using the same format as `train_labels.csv`. Each line in this file must contain a single number (0, 1, or 2) which is the class prediction for the corresponding instance in `test_features.csv`. Please double-check that this file has 560 lines, one for each test instance.
- `summary.(txt/doc/pdf)`: A brief description of your approach in a text file, Word doc, or pdf
- Source code files