**Data**

Feat\_data.csv is the feature dataset. The first column is the image ID. The second column is the image label. Columns 3-40 are the features of the images.

Normfeat\_data.csv has the same data except standardized and normalized (i.e. mean 0 and variance 1).

**Code Considerations**

Currently, the code is written to load the entire dataset and then split into training and testing datasets within the algorithm. Later, we can write a line or two to simply import pre-specified training and testing data. Also, there is nowhere to put CV data, currently.

The bulk of the Random Forest Classifier is in lines 33-34 and lines 41-43. For some strange reason, when I wrote the original piece of code (used on corporate data), I didn’t specify the parameters of the RandomForestClassifier in line 33. They are specified, however, in lines 41-42. Concretely:

* N\_estimators: number of decision trees in the random forest
* Max\_depth: maximum height of any decision tree in the random forest
* Min\_samples\_leaf: the minimum number of samples in a given leaf node (this is important so that you don’t have leaves that are just made of 1 example - overfitting)’
* Max\_features: number of features to consider for best split
* N\_jobs: should be set to 1, don’t worry about it
* Random\_state: seed used by random number generator

Line 62: dropFeatures(headers, data, 0.01)

What this does is it trains a random forest on the data that is given and then finds the feature importances. It then removes features that had feature importances < 0.01. This is basically like PCA’s job.

I don’t quite remember the function of lines 87-93, so there might be some errors there, but I have updated everything else, so I think it should work fine.