Assignment 4

Applied Machine Learning

Generally, a parameter selection procedure might be necessary to evaluate Probability of Detection versus Probability of False Alarm (i.e., Pd versus Pf) in order to make a decision about a classifier model selection and/or hyper parameter tuning for a classifier.

In this assignment we will produce an ROC plot presenting operating points of various classifiers and their varying parameters so that we can make a justifiable operating classifier/parameter selection for the following problem.

The classification of fake news or misinformation is a very important task today. Download the fake news dataset (https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset), `Fake.csv` and `True.csv` files. Load the data sets in your model development framework, examine the features to see they are text in `title` and `text` columns. Set `fake` as 1 and `true` as 0. Concatenate to have around 44880 rows. Apply necessary pre-processing to extract `title` column in `Tf-Idf`. These are basically words and their term frequency – inverse document frequency. Use around 50 features. Make sure sanity check the pipeline and perhaps run your favorite baseline classifier first.

```
df_true['class'] = 0; df_fake['class'] = 1
df = pd.concat([df_fake, df_true])
X = TfidfVectorizer(stop_words='english', max_features=40
).fit_transform(df['title'])
```

1. [70 pts] By using three classifiers, decision tree, random forest and neural network and at least 2 different hyper-parameter settings for each, generate operating points (via cross validation, mean FPR and mean TPR) and plot them on a ROC. Do not hesitate to use/modify the ROC plot code in the module notebook. In case you do not see enough variety in Pd-Pf you might need to work on the classifiers set and/or hyper parameters. And do not hesitate to try hundreds, if necessary, since the ROC is just a natural scatter plot. (Recommended parameters and ranges: depth [3-12], number of features [3-20], number of estimators [20-100], layer size [1-10], learning rate; and total of 10-20 OPs)

2. [10 pts] What kind of behavior would you expect to see in Pd – Pf interaction of an ROC plot? Do you see it in yours? (Hint: Pd and Pf corresponds to TPR and FPR)

3. [10 pts] From the ROC plot that you created make a selection of the classifier and hyper-parameter setting for this problem. Note that we are classifying fake news so your conclusion might be subjective but has to be supported by your findings.

4. [10 pts] Try adding `text` column to the features. Report the performance of a new classifier model of your choosing. Why do you think the performance is much higher than the previous one which only uses `title` column?