Capstone 2 Project Proposal

James Olmstead

The Problem

For my second capstone project, I'd like to utilize a different field of machine learning from my first. To this end, I'd like to utilize the natural language toolkit (nltk) library in python to create a fake news classifier. The problem of fake news is quite relevant with the proliferation of social media and with the onslaught of news from a wide range of reputable or not reputable sources. Having a fake news classifier can help social media platforms or other entities flag potentially questionable news stories and either prevent its distribution or warn viewers of its falsity.

The Data

The data comes from the Kaggle competition: https://www.kaggle.com/c/fakenewskdd2020. This dataset contains the text of news articles (mostly related to celebrity gossip) and has labels of 'real' or 'fake'. It also has a test set with text of other articles but without the labels. Once I make predictions on the test set, I can submit them to the website to check the accuracy. I can also compare my performance to others in the competition on the leaderboard.

Solving the Problem

I imagine solving the problem first with cleaning the data to feed into my algorithm. Ideally this won't be too bad since it is already split into training and test sets and isn't missing data. Next, I'll utilize text preprocessing techniques to convert the news articles into a format that can be read by a naïve bayes classifier. Finally, I'll run the training set into the classifier, make predictions on the test set, and see how they stack up to the leaderboard and iterate accordingly by changing parameters to improve the performance.