Machine Learning Project Proposal

Team Members

Conor McCauley (17323203)
Sean Roche (17332021)
Caolan Wall (17329660)

Motivation

We will be attempting to determine the degree to which certain philosophical viewpoints are related to and/or correlate with self-described political labels. We aren't interested in the relation between particular policy questions or contemporary political issues and political labels - we're only interested in how broader philosophical viewpoints are related to political labels.

Dataset

The data we will be using comes from test and survey results from the political philosophy test on Dichotomy Tests - a website run by one of our team members. The raw dataset is stored in a proprietary database which currently contains tens of thousands of test results. Each result consists of responses (agree/disagree/etc) to a large number of questions relating to the user's general political philosophy as well as responses to a general survey. The primary survey response that we are interested in the user's self-reported political label (right/left/moderate/etc).

Method

We will trial a number of multiclass classification techniques and analyse their performance. We have a large number of features in our dataset since there are sixty questions in the test. For this reason we will first use a logistic regression model with L1 cost function. We know from our previous assignments that the L1 cost function works well for when we have a sparse matrix of features.

We will use cross validation to assess the predictive validity of our solution. We will also train a k-nearest neighbours model and create multiple baseline classifiers to draw comparisons when assessing our solution.

Intended Experiment

We want to train the model to determine the particular questions that are conducive to one leaning or another. We will check for correctness using subsets of our data to train our model and test our model. The sixty questions will be used as the parameters to train our model and their leanings as our output - we may also need to include additional polynomial features. Cross validation will help to select the values of our hyperparameters and the number of folds we will use. We will use confusion matrices to present the correctness of our classification. Multiple baseline classifiers will be used to predict based on most-frequent and random predictions.