CS451
Final Project Proposal
11/20/17

Team:

Alex Berry and William Frazier: “Sommelier Neural Network”

Summary:

We are basing our project on a data set of 150,000 wine reviews from Kaggle. The data set includes ratings from 0 to 100, the type of grapes used to make the wine, a short description of the taste, where the wine was produced (down to the specific vineyard within each winery), and the cost of the bottle. Our initial plan is to implement a linear regression model using a neural network. We will divide the data set into a training set, a cross validation set, and a test set along a 60%, 20%, 20% split, respectively. We will convert the text features to numerical values and then perform mean regularization. The network will use the training set and examine the given features to make a prediction about the wine’s rating. Once we have a functioning network, we will explore ways to improve our feature vector (perhaps analyzing the short taste descriptions) and to tune the hyperparameters of the model (the number of hidden layers, the type of activation function, the learning rate, the lambda regularization value, and the weight value initialization).

Resources:

1. Our data set comes from Kaggle. It has 150,000 data points so it should be large enough to obtain a low-bias model without too much regularization.
2. We may use the neural network we built in Homework 3 as a base to build our network. Because that network was a model for logistic regression, we will need to modify it to implement linear regression. This conversion should not be too difficult; we will use this guide (which suggests removing the activation function in the output layer, keeping the hidden layers with nonlinear activation functions, and using a squared error function).
3. We will attempt to use the Natural Language Toolkit to turn the data points' taste descriptions into an objective numeric value which we can then add to the feature vector. We need to do more research on exactly how to convert the descriptions into useable features but the NLTK library is both the most powerful and easy-to-use natural language processing package for Python.