Midterm Project Report

CS 6350 Machine Learning

Ryan Dalby

26th October, 2021

Progress

Overview

First I attempted to use SGD logistic regression, but found more consistent convergence results using lbfgs solver. (occasionally did not converge to good optima) Then plotted learning curve and saw that the model suffers from high bias meaning that we would benefit from a more powerful model. Exploring nonlienar etc.

Normalizing helps a lot: test AUC from .6 to .75 compare logstic regression plots

SVC with a linear kernel performed the same or slightly worse than logistic regression, SVC with an rbf kernel performed the same as logistic regression in terms of AUC, did show overfitting, but increasing regularization did not help much. test AUC of .77 (although more overfit so performed slightly worse in kaggle competition)

Normalizing: Logistic Regression Classification model: train accuracy = 0.85615, train AUC = 0.7693313780111599 test accuracy = 0.8434, test AUC = 0.75450991533708

Support Vector Machine Classification model: train accuracy = 0.8877, train AUC = 0.8209118412629072 test accuracy = 0.8486, test AUC = 0.7703933696259709

Data Preprocessing

Logistic Regression

Support Vector Machine Classification

Next Steps

Try optimizing hyperparameters, MLP, ensemble methods (specifically gradient boosting classifier).