CS 578 Project Proposal:
Loan Default Prediction with Support Vector Machines

Priyank Jain
Phil Van Every
September 29, 2016

1 Problem Statement and Data Set

Can we predict whether or not it is safe to make a loan to a person based on his or her transaction history? In 2014, researchers at Imperial College London sponsored a competition challenging participants to predict the quantity of loss for a certain set of loans. In this project, we propose to simplify things by looking only at whether or not a loan will default. Using the same data given by the researchers, this simplification will allow us to focus on inherently understanding the performance characteristics of several prolific machine learning algorithms. The training and test data files consist of 105471 and 210944 independent samples respectively, each with 779 anonymous features including the label. It has already been de-trended, randomly sorted, and standardized. The data set was suggested on the class website and can be found at:
https://www.kaggle.com/c/loan-default-prediction

2 Experimental Setup

We will split the training data file into separate training and cross-validation sets.

Algorithms we can use to attempt to answer the above question include: binary svm, one-class svm, and support vector data decision.

Parameter tuning can be done as follows:

1. Pick ranges of regularization and kernel parameters and loop over both.
2. Use n-fold cross validation to generate an average ROC curve for each parameter pair.
3. Integrate the ROC curves and take the parameters that give the largest integral as optimal.

Performance will be evaluated over different training set sizes, different values of n in our n-fold cross validation, different kernels, and different feature subsets. The first two experiments tell us about a minimum criteria threshold for the algorithm to perform well. This allows us to compare algorithms based on their lowest minimum thresholds. The kernel and feature subset analyses tell us about the nature of the data itself - namely: which particular characteristics of a loan make it unsafe and how these characteristics relate between faulty and non-faulty loans so that a particular kernel makes the data separable in its hyper-space.

In each case, performance metrics can include the ROC curve integral value, the false positive rate and true positive rate pair at the 'elbow' on the ROC curve, accuracy, precision, recall, etc... A visual inspection of each of these will lead us to select the one that best illustrates the performance of each algorithm.

We will implement the above algorithms and plot results using scipy and matplotlib libraries of the Python programming language.