Andrew Donelick
Alex Putman
Machine Learning – CS 158
6 April 2015

Predicting Restaurant Revenue

Project Summary:

Our machine learning final project is to compete in a Kaggle competition for predicting restaurant revenue. Tab Food Investments manages several large chain restaurants worldwide but until recently has always selected new opening locations by a subjective process that is based on nonscientific methods. They are now seeking ways to predict locations that will yield the highest revenue based on their plentiful example locations and associated revenues. This is what we intend to do: based on many data fields of a potential location, we seek to correctly predict the revenue that that location is most likely to make.

We first plan on visualizing our data to get an idea of its shape. This will help us with attribute selection when implementing several of our six planned regression techniques (KRR, SVR, Decision Tree Regression, Gradient Boosting Regressor, KNN Regressor and Neural Networks). Since we plan on using Scikit Learn and OpenCV for out-of-the-box implementations, it will be easy for us to test out these different ideas and see how useful they are for our data set. For KNN regressor, we plan on building our own semi supervised approach since we have found code online that will allow us to do this more easily than semi supervised approaches to any of the other techniques. We will evaluate and differentiate these different examples by using cross validation for mean squared error and R^2 metrics, as well as the given Kaggle evaluator.

After evaluating which of these techniques works best on our data set based on our evaluation metrics, we will continue to work on this best technique. This includes working on improving the hyper parameters or trying to include semi supervised approaches if they were not already being used. Our reason for attempting semi supervised learning is based on the fact that our data set has a size of around 100,000 but our training set only has a size of 137. We feel a semi-supervised approach might be best to overcome this discrepancy. We don't plan on supplementing all of our techniques using semi-supervised approaches due to time constraints.

Resources:

Scikit Learn Resources:

KRR - http://scikit-learn.org/stable/modules/kernel_ridge.html
SVR - http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
Decision Tree Regression
http://scikitlearn.org/stable/auto_examples/tree/plot_tree_regression.html
Gradient Boosting Regressor - http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
KNN Regression - http://scikit-learn.org/stable/auto_examples/neighbors/plot_regression.html

OpenCV Resources:
Neural Network - http://docs.opencv.org/modules/ml/doc/neural_networks.html

Relevant Research Papers:

Semi-supervised Regression with Co-training: http://ijcai.org/papers/0689.pdf
Semi-supervised Bias Reduction - http://papers.nips.cc/paper/3376-statistical-analysis-of-semi-supervised-regression.pdf