CS 7641 Machine Learning
Assignment 1

Philip Bale
pbale3

Due February 4th, 2018 11:59pm

2 most important plots: learning curve and model complexity curve

Classification Problems

1) US Permanent Visa Applications

The first classification problem revolves around classifying whether or not a person's US permanent visa application will be accepted or denied based on the parameters of their application.

2) Home Sale Price Predictions

Overview

The second classification problem revolves around classifying a home's price bracket based upon the various characteristics of the home. Among the features used in the classification are :

- Subjective measurements: Exterior condition, house style, overall quality rating, and overall condition
- Objective measurements: Type of dwelling, building type, lot size, neighborhood, year built, and year sold

After an initial review of the dataset, the classes were defined as pricing brackets divided into 100k groups. I.e: 0-100k, 100k-200k, 200k-300k, etc. The dataset contains 1451 samples. An additional dataset containing another 1400 testing samples exists but was not used as it contains unclassified sale prices. It will, however, prove useful for unsupervised learning.

Why is the dataset interesting?

This dataset is interesting for two primary reasons: real-world applicability and participating in a Kaggle challenge. First, modeling home prices is both a difficult and lucrative task. If one can succesfully model home sale prices on large sets of data, he/she can make large amounts of money investing in real estate when he/she detects outliers in listed price vs. what it is expected to sell for. This applies to flipping, investing, and remodeling. Second, the dataset is part of an ongoing Kaggle competition that does not have a winning solution yet. By taking part of the competition, the dataset presents the opportunity to work towards a winning solution and advance ones algorithms over time.

Houses can have a very large amount of features–with a large amount of variety in the individual features. Similarly, housing is prone to personal taste and frequent need for upgrades/modernization. In such, I believe price estimation is an excellent problem, full of depth and complexity, that is suitable for a machine learning approach.