

CS 7641 Machine Learning

Assignment 4

Philip Bale
pbale3

Due Sunday April 22nd, 2018 11:59pm

Introduction

This assignment explores unsupervised learning and dimensionality reduction. It begins by examining clustering algorithms, specifically k-means and expectation maximization. It then proceeds to cover four dimensionality reduction algorithms: principal components analysis, individual components analysis, randomized projections, and random forests. After running these six algorithms on the original datasets and observing the results, the results are then piped into a neural network learner for further examination.

Datasets

The datasets chosen were the same datasets chosen for assignment 1. The first dataset is the US permanent visa dataset. This dataset is interesting due to its potential to aid in the visa application process from a cost and time savings potential. It could also enable confidence in those interested in applying for a US permanent visa but doubting their chances of acceptance. At the end of the day, the goal is it to try to determine the application result before time, money, and other resources are spent. As before, 6 features are used.

The second dataset is a home sale price prediction dataset taken from an ongoing Kaggle competition. 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 successfully 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 one's algorithms over time. As before, 11 features are used.

Part 1: Clustering Algorithms

Introduction

K-means clustering is the first algorithm applied to the datasets and expectation maximization is the second. Both algorithms work by clustering: gathering groups of instances together based upon their features. The rationale is that similar instances will likely be labeled the same way—such as identical visa applications obtaining the same outcome.