CS 66: Final Project Proposal

Census Data to Predict Education Level

Kenneth Gwon, Raymond Liu, and Richard Muniu

Dataset

We will be using the 1996 Census Income Data Set from the UCL Machine Learning Repository. This dataset, also called the "Adult" dataset, contains data on whether an individual's income exceeds $50K/yr based on demographic data from the 1994 census. The census data contains 14 features for each person(datapoint), which we highlight in the goals section below. There are a total of 48842 data points, but as we would like to use an assortment of learning methods and also potentially K-fold cross-validation, we will trim this size down to 2000 examples.

Goal

While the classical approach would be to model a binary classifier that can predict a person's income based on their demographic data, we seek to instead investigate how features such age, work class, marital status, occupation, relationship, race, sex, hours worked per week, native country, and income (either greater than 50k or less than 50k) are related to level of education.

Set of algorithms we will apply to this dataset

The algorithms we will use on our data include SVM, Naive Bayes, and Decision Trees.

A scientific question(s) we are trying to answer

- Which classification algorithm will produce the highest test accuracy?
- Which features are most important in predicting income?
    - Do our algorithms agree on the most important features?

Evaluation/Interpretation of results

In order to evaluate how effective our algorithms are in classifying the level of education, we will graph both test and validation accuracies, and display confusion matrices. Furthermore we will use K-fold cross validation and calculate the accuracies across disjoint subsets of the data.

References

Census Income Data Set , https://archive.ics.uci.edu/ml/datasets/Census+Income, Ronny Kohavi and Barry Becker (1996)

Lemon, Zelazo, Mulakaluri, "Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques", http://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf,

`