CS 6220 Project Proposal

Chuong Tran

Tim Holland

Immigrants living in the United States are vitally important to its economy, culture, and history. In addition, immigrants are a significant factor in population growth, with 12.9 percent of the US population being foreign-born as of 2010. The topic of immigration has become an increasingly divisive topic in US politics, with more time being used to express opinions and less time being used to investigate the data.

In order to get a better understanding of US immigrants and how their backgrounds factor in to the type of contributions and influences they have, we will investigate datasets provided by the US Census as well as data provided by FiveThirtyEight.

To that end, we wish to answer the following questions, as well as any ancillary questions that may arise from our investigation of them:

* Among foreign-born groups, who's likely earn more than 50k per year?
* Does job/marital status matter in terms of income earned?
* What is the most important factor/feature that contributes to that?

We propose using the following approaches to answer these questions:

* Decision tree
* K-nearest neighbors
* Logistic regression

Decision trees are a type of supervised learning method and will be good to use for classification as it can help to answer investigated questions better. After we have classified the data, we will know which group, which race and ethnicity earn more than 50k per year. The dataset comes with different types of data and it indeed includes the numeric and nominal features which are a great use for decision tree. In another way, it can handle the continuous and categorical data. Plus, decision trees are a great way to visualize data and help people understand how classes are classified. Another method that can be applied is K-nearest neighbors (kNN). It is also a supervised learning method and can be good to use for classification (or regression) tasks. We can also do cross-validation to compare the efficiency of two approaches and see which one has higher efficiency or lower error rates. Furthermore, we might also use a logistic regression method as well for the classification task as we can calculate the odds of the outcome for some of the investigated questions. Since the decision boundaries are constructed differently for each of these approaches, a question then can be answered by one technique over the others. Again, our datasets come with different features and types we will have to do some preprocessing work first before we can apply these approaches directly as each of them takes on a different form of data.

Since we don’t know the answer to these questions or if they can be answered using the data to which we have access, some flexibility may be required in the final product. Perhaps interesting patterns can be discovered that may provide additional insight into our original inquiries once they’re discovered.