Machine Learning Decal, Spring 2018
Project 1: Data Cleaning, Visualization, & Mining

Machine Learning at Berkeley

1 Introduction

In this project, you will explore a new dataset and build a complete end-to-end pipeline for data visualization, from organizing and cleaning the data to visualizing its various aspects. Typically in any real-world use case, you would have the problem presented to you; you would then collect data related to answering this question and start exploring it to find answers.

You will be analyzing a Kaggle dataset from the United States Census Bureau and the 2015 American Community Survey. Your job is to discover unique and interesting patterns and insights from the American Community Survey Data as you clean the data, visualize the data, and train regression and classification machine learning models throughout the processs. For instance, you could ask questions such as: How is the average income in a census tract or a county related to the percentage of people working in different industries (professionals, service industries, manufacturing industries etc.), the racial demographics of the area, or the geographic region of the United States (Northeast vs West Coast)? How accurately can you predict the child poverty rate in a census tract or a county using linear regression and data from employment figures, racial demographics, and other factors in the 2015 American Community Survey?

You may use any python libraries you wish for this project. You will be given American Community Survey data and be asked to follow the checkpoints for cleaning/visualization/creating regressions on the data. We will not ask you to do anything specific - it is up to you to figure out what methods and avenues best fit.

There will be 2 components to this project: the jupyter notebook that contains all the code you used to manipulate the data, and a 2-page typed writeup explaining the major decisions you made and your biggest findings. The writeup will be the main component of this project, with the notebook simply to verify what you state. You don't need to put in too much effort into the writeup; We anticipate it to be a relatively simple task once you have the notebook. YOU NEED TO WORK IN GROUPS OF 3-4.

2 Setup & Understanding the Data

First download the csv file titled "acs2015countydata.csv" from the following link:

https://www.kaggle.com/muonneutrino/us-census-demographic-data/data

Then create a new jupyter notebook, and load your csv file in the jupyter notebook. Make sure to read and fully understand the Column Metadata section on the Kaggle website that describes what each of the columns of the dataset are measuring.

3 Cleaning, Organizing, Exploring, and Visualizing the Data

Now, based on what you have understood about your data, you should clean and organize it (so that it's easier to create visualizations/fit models later). Don't be afraid to throw out large parts of your dataset if you feel that it is not relevant information to the question that you are trying to answer. Make sure to drop or impute the missing values.

Once you have cleaned and organized your data, begin exploring your dataset. Create overviews, analyze

some summary statistics, and have fun visualizing the data! This is the most fun part of the project - all the other steps were leading up to this! Now you can go ahead and create your graphs, models, and other visualizations. Feel free to use any libraries to create interesting graphs. Note that each graph should be properly labeled with x and y units, proper x and y scales, a title, and a legend if necessary. Make sure you demonstrate your ability to create rich graphs by creating various different types of graphs.

Here are 2 types of graphs that you will want to create for this project.

1. Creating multiple graphs (at least 4) that plot important financial indicators (y-axis) like the median annual income, per capita income, poverty rate, childhood poverty rate, and/or unemployment rate present in a particular county AGAINST several features in the dataset (x-axis) that might affect these factors, such as the total population of the county; the racial demographics of the county; or the occupations of different workers (professionals, service workers, construction, manufacturing).

2. Using plotly as a graphing library to visualize these financial indicators in the 2015 American Community Survey data on a geographical map of the United States on a state-wide or county-based level (similar to what we did in class when we visualized global average temperatures for different countries on a global map).

Make sure to note down in your project writeup any important steps/realizations here.

4 Regression Analysis

Make sure you also demonstrate your ability to create regressions on your data. In this part of the project, you will pick 3 financial indicators of a county's economic performance in the dataset (like the median annual income, per capita income, poverty rate, childhood poverty rate, and/or unemployment rate). You will then train 3 linear regression models in scikit-learn that attempt to predict these 3 financial indicators of a county from the rest of the features collected in the dataset. You should then analyze the accuracy of your linear regression model and report important predictors or features in your dataset that had the highest effect size on these 3 financial indicators (as discovered by your regression model). Is there a way you could make the regression better (optional: try adding regularization to see if you could get better results).

Also remember to document any work you do here in your writeup! Include important graphs and plots for visualization, tables for characterizing the fitted parameters and predicted accuracy of your linear regression model you have created, as well as a thorough discussion to what you have discovered about the dataset through these visualizations and regression models.

5 Submission

Remember your write-up must be at least 2 pages in length. Please include the pictures of your visualizations/regressions in it to illustrate points you talk about! There is no upper limit on how long the report can be.

You must submit your both your writeup and jupyter notebook. Save them both as PDFs and attach them together, with the jupyter notebook at the end. Make sure to note your group member's names at the top of your writeup. Only one team member should upload the PDF to Gradescope.