COSC 311 - Project 1

Dr. Joe Anderson

Due: 24 October

1 Objectives

This lab will be the first of a three-phase project for the semester. Now that we have developed some tools to import and organize data, you will begin the process of studying it, cleaning it, and starting to "engineer" the data. In the second and third phase, you will think about building models from the data, designing a predictive analysis study, and communicating the results through prose and visualization.

2 Requirements

1. You may work in groups of one or two members to complete this project. However, if you work with a partner, you will expected to keep the same partner for Projects 2 and 3; however, exceptions can be made if communicated in advance with the instructor.
2. Begin searching for some freely available, large-enough-scale (thousands or hundreds-of-thousands data-points) datasets online. Some good places to start are:
   (a) UCI Machine Learning Repository
   (b) Kaggle
   (c) Google Dataset Search
3. Select at least three data sets and answer the following questions about them:
   (a) What type of population is being sampled? What are the "things" getting measured – usually one per row of data.
   (b) What features does each sample have, i.e. what is being measured?
   (c) Are the features quantitative or qualitative? Ordinal or nominal? Continuous or discrete?
   (d) Is the data "complete" or do some of the samples have null or absent values for certain features? Why are these samples still useful? Why are they incomplete?
   (e) Why are these features chosen to be part of the dataset?
   (f) What are some other features that are not included but that you think might make sense to include for this dataset?
   (g) Give at least one way that you can pivot the dataset to get a slightly different representation of some values. Explain what this is and how you would use it for a visualization.
   (h) Identify any possible relationships between features included in the data: which ones are likely to affect others?
      i. Show at least one plot or visualization to illustrate this (possible) relationship.
      ii. What numerical or statistical techniques might you consider using to determine whether the relationship is reliable?

iii. Are there external inferences you think might be possible? For instance, can you hypothesize a relationship with data not included in the dataset? Why or why not?

(i) What "extra" features can you perhaps compute from the data? For example, if you have data that includes product dates of purchase, you can "engineer" the data to construct the most popular products over various lengths of time (e.g. a particular holiday season). How might you use this information? Using the holiday example, you might try to correlate holiday sales of a product to some mainstream event that popularized it.

3 Submission

Use a Jupyter notebook for each one of your datasets, answering the questions with Markdown or Text cells.

Upload all responses, source files, and documentation to the course MyClasses assignment in a `.zip` or `.gz` archive.