Final Project: Data Analysis

Leilani Leon

# Data Tools

Data is collected through many aspects of the digital world, but this data, regardless of its start point, eventually needs to in a more persistent storage format that is accessible after the program has stopped running and, ideally, is easy to analyze. Any related pieces stored within the file is called “tabular” data and this data is often stored in text files know as comma-separated values (CSV). CSV files are common, so many tools can read and write said files—like spreadsheets and libraries in programming languages to read and write CSV files from their applications. The drawbacks to these files, however, include a limitation in size by the computer’s hard drive space and the time it takes to open a large file. It also doesn’t come with any built-in tools for querying the data so the files must be opened in either a spreadsheet application or a program needs to be written to iterate through the data. This data is stored in a database, which is a system that allows for the data to be easily accessed, updated, queried, and deleted. This data is also sorted into appropriately sized files while remembering what files the data is in. Query languages, such as SQL (structured query language), are used to interact with the databases. Computation of statistics meant to describe the data sets can be done with the use of the data and many tools are used for this analysis depending on the needs and skills. Filters can be used to find specific data samples and, upon collection, patterns, such as trends and correlations, can be found. Trends where the numbers steadily decrease is known as a downward trend, whereas trends where the numbers are steadily increasing is known as an upward trend. For both of the exercises, I scored a proficient, though I had taken the first exercise a few times due to getting one question wrong.

# Big Data

Data sets are increasingly growing larger in size. One report estimated that, by the year 2025, the total size of data will be 175 zettabytes. A single zettabyte contains a trillion gigabytes. These big data sets have grown so large in size that the traditional methods of storing and processing them are no longer applicable. Despite this, larger data sets offer new opportunities for analysis that weren’t possible with a small data set. There are two sources of these big data sets—some comes from a very large source while most is a collection of data from many little sources. When a computing system needs to store the massive amounts of data that comes with these sets, two things need to be considered. How will the data be stored? And how can the data be processed efficiently? Through the use of disk arrays or disk enclosures, this data can be stored within the dozens of interconnected hard drives, which are then stored and managed within a data center. These data centers contain multiple aisles of computer equipment and utilize infrastructure that provides electricity to the device and prevents it from overheating. They are often highly networked so that data and computations can be shared across multiple machines. A large data set can take a long time to process, regardless of whether the data set can fit on a single hard drive or not. For the exercise, I scored a proficient on the first attempt.

# Bias in Machine Learning

A machine learning algorithm is one that automatically improves itself with experience by processing more and more data, then modifying itself based on the properties of the data. There are three general approaches: reinforcement, where the algorithm performs actions that will be rewarded the most; unsupervised, which looks for and finds patterns in unlabeled data by clustering and identifying similarities; supervised, which analyzes labeled data and learns how to map input data to an output label. These aren’t the only ones, though, as neural networks have been growing in popularity. A neural network is a type of supervised machine learning algorithm that operates similarly to how were believe brains do. Input flows through many layers of “neurons” and eventually leads to an output. Instead of programming each and every neuron, computer programmers train a neural network using a massive amount of labeled data, which depends entirely on the goal of the network. The training phase is meant to determine weights for the connections between neurons that will then classify the training data. All the weights are set to random values at the starts, so the initial classifications are way off. From there, the algorithm eventually comes up with a set of weights that do the best job at classifying all the training data by learning from its mistakes. The accuracy of these algorithms relies heavily on both the amount and diversity of the training data and its often referred to as AI, though that’s not entirely true. It is, however, susceptible to biases. The algorithm cannot identify biases and, as such, will only replicate them. The CJS uses these algorithms for risk assessment and companies use this for screening of possible hires. Facial recognition services use machine learning algorithms to scan a face and detect a person’s gender, race, emotions, or even identity. For the exercise, I scored a proficient on the first attempt.

# Unit test

The unit test repeated questions seen in the modules. Overall, the test itself was easy to understand, though I received an 8/9 due to selecting a single wrong answer on a multiple-choice question. Computing basic statistics was the module I had most difficulty on with the exercises and the questions that I had problems with were the one that were repeated on the test.