Lesson 3 – Data Analysis

Slide 1

Hi everyone. This week we will focus on data analysis. In our programming exercises, we will study advanced techniques for uncovering insights within a data set.

But before we begin programming, let's take a moment to define common terminology used in data analysis and statistics.

Slide 2

We hear all the time that data is all around us. But how is it generated?

The process starts when we ask questions such as how many, how often, where, and what kind? When we store the answers to these questions in a data source, then we have generated data that can be analyzed.

Nearly everything in our environment is a source of data.

We can collect data on people, such as customers, patients, or even family units.

Countries are also a source of data. For example, the World Bank and UNESCO collect country-level data on economic and educational outcomes, including GDP, unemployment rates, and degrees awarded.

Processes are another rich source of data. In business, companies track website visits, clicks, likes, and other customer activity. In healthcare, the outcomes of clinical trials are recorded to analyze the effectiveness of drugs. There are many more examples and the sources of data all around us are practically infinite.

Slide 3

All data sets are comprised of four key elements: objects, variables, values, and observations. Objects are sources of data. These can be anything upon which we can measure quantities of interest. Some examples include individuals, events, products, and even galaxies. As an example, suppose we are educational researchers that are interested in studying students. In this case, students would be our objects of interest.

Variables are characteristics of an object that can be measured. For our student example, variables of interest might include SAT scores, GPA, or state of residence.

Values represent the state or numeric level of a variable when it is measured. For example, we might obtain an SAT score of 1900 for a particular student that we survey.

And finally, observations are sets of measurements made on an instance of an object. For our research on students, an observation represents the set of variable values obtained for a particular student. For example, one student might have gotten a 2100 on their SAT exam, has a GPA of 3.8, and lives in Virginia.

Slide 4

Now that we understand the four main components of data, let's discuss the optimal way to structure it in order to facilitate data analysis and machine learning. There are many acronyms for describing data that is in optimal format. These include terms such as tidy data, data table, data matrix, and analytical form. Regardless of the term that is used, all data should be structured as follows to enable advanced analytics:

The rows of our data file should represent observations of our objects of interest.

The columns should represent the variables that were measured on our objects.

And finally, the individual cells should represent the values obtained for each combination of object and variable.

As an example, suppose we have recorded data on customer purchases. The table below shows an optimal way to structure our customer data for data analysis. Each row in this data represents a particular customer, while the columns indicate which variables were measured, such as which store the customer was at, their age, their metro area of residence, the amount they spent, and the number of products they purchased. The cells of each row contain the values that a particular customer had on the variables in our data.

Now that we understand data terminology and the optimal structure of data files, we can focus on advanced data analysis techniques in our programming material this week. See you there!