Intro to Data Analytics

Analyzing Data

Why Analyze?

Data is a big business these days. Organizations meticulously plan out their data strategies to maximize the potential of the data they store. They plan out methods of capturing and storing the data, implement security schemes and protocols, and scrutinize vendors to find the best fit for their needs. This process requires a lot of effort and expense so it's logical to assume they expect a significant return on that investment. The data isn't worth much on its own, but all the effort pays off when the data is analyzed and turned into actionable plans. In this lesson we will review the different techniques we use to analyze data in order to maximize the information and value we are able to extract from it.

4 Types of Analysis

There are many techniques that have been developed over centuries to help us analyze data. In more recent history, technology has been incorporated to enhance the types and quantity of analysis that can be performed. These advances are necessary to keep up with the types and volume of data we capture. The various analysis techniques can be broken down into different categories based on the level of understanding we can gain from them.

Descriptive analysis is one of the most rudimentary forms of analysis and therefore one of the most common. This analysis is often described as understanding "what happened". For instance, this type of analysis can tell you what time of day is hottest.

Diagnostic analysis is associated with understanding why something happened. This analysis may take into account the relationships between datapoints to explain the measurements. For instance, you may explain an unusual temperature reading by confirming that it was raining at the time of that reading.

Predictive analysis is concerned with using data from prior occurrences to anticipate possible outcomes. In other words, what is likely to happen given certain conditions. For example, if it is 76 degrees at 4 AM and 82 degrees at 8 AM, what is the temperature likely to be at noon?

Prescriptive analysis adds logic to suggest what actions to take based on analysis of the dataset. This might incorporate some predictive analysis to help you determine whether you should take an umbrella or sunglasses each day.

Regardless of what types of data you have or the type of analysis you are trying to perform, the most common of these techniques will involve quantitative data and rely on some mathematical underpinnings – namely statistics.

Samples and Populations

Intro to Data Analytics

Whether you're calculating standard deviations and z-scores for numeric data, finding word frequencies in a document, or converting color values in an image, math plays an important part in data analysis. For now, we will focus on some important, high-level concepts and save the math details for another time.

Most analysis starts with a dataset which is a sample of a larger population. For example, let's say we want to understand something about dogs. The population in this case is every dog that exists in the world. In our example, the time, expense, and effort required to measure every dog that exists makes the task impossible. This is true in many situations as the size and spread of populations can make an accurate measurement of an entire population unfeasible. As a result, we may limit the scope of our population to something like "all dogs in the United States".

Measuring a sample, or subset, of the population is a much more realistic goal. A sample of dogs may include some from your town, or even from multiple cities throughout the country. We can then use the data from our sample and "extrapolate" the results using mathematics. This is just a fancy term meaning we can assume what we find applies to all dogs, not just the ones in our sample. Obviously, this isn't always a safe assumption. We have a lot of variables at play… for instance what breed of dog we measured, their environment, their age, etc. Getting a large enough sample helps to ensure that it is "representative" of the whole population, which is important when trying to draw conclusions for a population based on a sample.

Statistics gives us a way to understand the distribution of values in a population. You can think of a distribution as the pattern that represents how frequently values occur. This is dependent on something called the Central Limit Theorem, but all you need to know for now is that if we assume that a population fits into a pattern and if our sample is large enough, the means from each sample will form a normal distribution. This well-known distribution (sometimes called a Gaussian distribution) that takes the shape of a bell curve. In normal distributions, the most common values will appear in the middle, forming a peak, and as you move farther away from the middle the values become less frequent, forming the tails. There are other types of distributions as well, like Gamma and Poisson, which take different forms.

Outliers

As values are collected, there will likely to be some values that don't fit the distribution perfectly. These values are known as outliers and can distort your analysis if you don't identify and investigate them. For example, imagine you are recording the time it takes to swim one lap at a local pool. You are recording times of local swimmers when an Olympic athlete shows up and swims a much faster lap than anyone else. It is not common to swim that fast so including that time in your analysis would shift your expectations for normal swimmers.

Outliers are often associated with anomalies, niche cases, or recording errors. There are different definitions for what counts as an outlier, and it is often up for interpretation. Some common rules of thumb are entries that are more than 2 standard deviations from the mean, or fall

outside of 1.5 times the inner quartile range. You may choose to remove them from the dataset, modify the values to control their effect, or even leave them in if they reflect valid data entries.

Organizing Data

Messy Data

Data in the real world is messy. It's very common to have datasets that are missing values, contain outliers, or don't contain the exact information you need. Sometimes you'll need to combine datasets, convert units, calculate additional values, etc. "Data wrangling" is a term used to describe combining and cleansing data for further analysis. You may also hear the term "data munging" used in the same way.

These activities are often mentioned alongside ETL, which stands for extract transform load. Extract refers to pulling data from a given location like a database or storage location. Transform describes the data wrangling actives before loading into another system or location. A good example might be extracting data from various databases used by a company, then combining and cleaning those datasets into a larger dataset, which is loaded into a data warehouse for further analysis.

Let's take a closer look at some of the most common activities when performing data wrangling.

Unioning

You can combine datasets vertically which is called "unioning". This method of combining datasets allows for combining by column position, but it's more common to match by column name. One important detail is ensuring all datasets are using the same units before combining them. When combining datasets that include columns that are not universal, the datasets that did not have this column will often populate it with null values, which indicate the data was not present.

Joining

Joining data refers to combining multiple datasets based on rows. After a successful join, the dataset will have more columns than before, expanding the dataset horizontally. This can also be used to "match" values from different datasets.

You can join a dataset by position, meaning that the values will continue to exist on the same row as they did in their original form. This is a good option if all of the datasets are arranged in the same way, but if values differ, you may end up with a mess. Alternatively, you can combine datasets by matching values from corresponding columns, like a unique identifier or key. For instance, we may want to combine two datasets by matching the names to ensure the additional values are on the correct row. The matched values that result are sometimes called an "inner join".

Intro to Data Analytics

Sorting

A common step in analysis is sorting values within a column in ascending or descending order. Sorting will rearrange the values in the dataset, whether sorting on a string datatype or numeric field. You can apply multiple sorts to a dataset and they will be processed in order, meaning the first sort will take precedence, then the next until complete.

Filtering

Filtering is a method of dividing the rows of the dataset based on some criteria you specify. You can think of this as a test with rows that pass, or meet the criteria, move on to one side and those that fail move to the other. In filtering, you will often see true and false categories that correspond to the criteria listed. For example, a filter on age = 12 will separate all entries with an age of 12 from the rest of the dataset. It's also possible to have a compound filter that includes additional criteria by using the terms AND & OR. "and" filters require that both/all conditions be satisfied, while "or" statements will accept either option as sufficient.

Pivoting

No data format is perfect. Sometimes you will need to change the format of a dataset to meet your needs. For example, it may need to be rearranged to enhance human readability, better understand a particular relationship within the dataset, or to satisfy the requirements of a secondary system (e.g. creating a visualization or satisfying a database schema). Rearranging the data is known as pivoting the dataset and it often rearranges the columns and rows into different orientations, including making rows into columns. Depending on the environment you are using, you may hear some additional terms for specific actions like transpose, cross tab, and pivot tables.

Altering Data

Up to this point, the data wrangling methods we have discussed do not alter the values in the dataset, only rearrange them in different ways. Sometimes it's necessary to create new values, calculate additional values, or modify the existing values for a variety of reasons (e.g. standardizing values, formatting, cleansing).

Functions

Functions are critical to data analysis because they allow you to perform a wide range of tasks on the dataset. In a programmatic context, a function is a key phrase (sometimes called a command) that will trigger an action (based on code). The commands often include parentheses to contain the necessary parameters, aka arguments. The parameters provide relevant details, like which column contains the data or specifying units of measure. If a function includes multiple parameters, they are often separated by commas and presented in a specific order.

Intro to Data Analytics

Functions are often sensitive to datatypes which is why it's common for them to require the data they manipulate to conform to a particular schema. For example, if you were to perform a calculation on the number of days that have passed since a given date, the function you use will likely require that the incoming date be formatted in a date-specific datatype. The result of the function will usually require a column with a specific datatype, which will vary from function to function.

Functions are often grouped into families based on the actions they perform. In programming languages, functions may be part of a library which is downloaded separately (e.g. numpy & scikit-learn in python) or they can be part of an application's base code. Libraries are collections of assets that are useful for particular tasks and enhance overall functionality. In applications, functions are usually integrated into the application's code, so there is no need to download a separate library.

Formulas

Formulas use operators and arguments to perform calculations without requiring a key phrase like functions. You can think of formulas just the same as you do in math. Operators are things like + - * and /. Formulas are often used to perform mathematical operations on numeric data but can also be used with string data to perform concatenation, which is joining multiple data entries into a single entry.

Summarizing

One thing to note about working with tabular data is that columns have different properties than rows. A column (aka field) of values is expected to have the same units and share a column name. Values that share a row are related to the same event and are sometimes called records. Some data wrangling techniques separate functionality between columns and rows, even though the dataset may be pivoted or arranged in a manner inconsistent with the expected row and column format.

The term summarizing is used to describe calculations on a column of data, which is often associated with a sum function. This may also be called aggregation. Calculations across a row are often the domain of formulas, rather than a summary operation.

Parsing

Sometimes, individual values need to be split into multiple pieces. For example, a dataset may contain a full name and we want to split the names into a column for first names and another for last names. Splitting values into multiple pieces is known as parsing data and there are multiple tactics used to achieve the desired results. It is easiest to accomplish parsing when there is a delimiter, which is a term for features that can be used to separate the values. One of the most common delimiters is a comma, but you can use other characters like a dash, pipe, or

even a blank space. If there is no delimiter or requires more complex logic, there are more advanced techniques like regular expressions, or REGEX, that can be used to split the values.

Cleansing

Some values require additional cleansing to ensure uniformity in the dataset. Since computers encode each character as a unique string of binary, each character is treated as unique. This means some of the details that humans tend to ignore can cause issues for analysis. For example, capitalized letters are not equivalent to lowercase letters so you may need to standardize the capitalization before joining datasets or performing analysis. Other complications may arise from typos or abbreviations in the dataset.

Another tricky aspect that may require cleansing is whitespace, which are non-visible characters such as spaces and line breaks. Leading or training whitespace can impact a computer's ability to recognize matching values as well.

Depending on your environment, missing values can take different forms and might be interpreted as empty, null, or NaN. Those terms all mean something different and it's up to you to determine how you want to handle them. To better understand the differences, imagine filling out a form. If the form included a blank for middle name but you chose not to fill it in, this would be a blank or "empty" value. Empty refers to a value that did not receive an input.

If you combined the results of this form with the results of a previous version that did not have a field for middle name, those previous results would display nulls instead of empty values. A null signifies that there is no data available. While empty and null values amount to the same thing (i.e. you don't know a person's middle name), they are different and can be treated differently when performing calculations.

NaN stands for "not a number" and populates in numeric data fields when they receive an incompatible value. For instance, someone typing out a number in words rather than using numeric characters.

Ultimately, how you handle missing values is a judgement call and will often depend on the type of analysis you plan to perform. Imputation is one method of replacing missing values by calculating a replacement, for example the column's average or mode. Another method would be to find the most similar record in the dataset based on the other columns, then use its value, which is called using the "nearest neighbor". You can also remove the entry altogether or replace with a zero. Other times, null values can be left in the dataset if they won't impact the results.

Working Smart

Automation

Intro to Data Analytics

All of these data wrangling steps are used regularly when performing analysis. As more and more data is captured and stored, the need for data cleansing grows as well. Instead of performing the same actions over and over on new datasets, we use automated, repeatable processes to apply the same cleansing techniques to new datasets without requiring a human to perform each step again. The benefits of automating data cleansing are easy to understand, but automating solutions requires programmatic problem solving.

Computers are much faster than humans at certain tasks, but they require explicit instructions in the form of programming. Some of that programming is built into the operating system and application levels, but in data analysis we also have to provide step-by-step instructions for the analysis that can be carried out by the computer. Think of these as the steps in a recipe. These step-by-step instructions are sometimes called algorithms. Alternatively, they may be called workflows or scripts depending on the environment.

In order to automate solutions, the steps in the algorithm must accommodate the entire dataset (and any future states that may differ from the current state). As such, we may need to separate and pivot the data in order to apply some steps to only select pieces of data, then recombine the other data before we are done. This type of problem solving can be confusing at first, but it comes with significant benefits and new possibilities. Solutions that are built to handle new data in the future without requiring human interaction are known as "dynamic solutions".

Advanced Analytics

Artificial Intelligence (or AI) is a popular subject of science fiction and advertisements because the term is so impactful, but what does it mean for data analysis? AI is a branch of research that uses machines to simulate the learning and thinking processes of humans. Some of the most advanced implementations of AI have many layers of logic to help them improve over time (i.e. learn). As available datasets continue to grow, we rely on AI to help us process them into actionable insights. AI is a rich and complex area of study which is often divided into specialized areas.

Machine learning (or ML) is a sub-field of Artificial Intelligence that specializes in creating the tools, methods, and techniques that find patterns in datasets to make predictions. There are many fields of study in machine learning which focus on different aspects of analysis. Some of the most commonly used areas of study fall under a category known as Data Mining. This encompasses regression and classification models which are used to predict values based on other variables (columns). Another popular area is Time Series analysis which forecasts values based on the historical values of the same column over time. Other areas of interest include Natural Language Processing (or NLP) which focuses on understanding human languages and Computer Vision which specializes in interpreting image files for tasks like object recognition.

While many portrayals of AI in pop-culture may lead you to believe that it will be the downfall of humanity, in today's reality these areas of study are often used to complement human

intelligence. Many applications rely on ML to perform the “grunt work” so a human doesn’t have to supervise mundane tasks. They also allow us to work with much larger, less organized datasets than before. As the hardware and software we use improves, the techniques we use to analyze data evolve with them!