جامعة الطائف
TAIF UNIVERSITY

PRINCIPLES OF DATA SCIENCE

2021 - 2022

Assoc. Prof. Dr. Salha Alzahrani

Department of Computer Science

جامعة الطائف
TAIF UNIVERSITY

CHAPTER 2: TYPES OF DATA

- FLAVORS OF DATA
- STRUCTURED VERSUS UNSTRUCTURED DATA
- DATA PREPROCESSING
- QUANTITATIVE VERSUS QUALITATIVE DATA
- THE FOUR LEVELS OF DATA
- DATA IS IN THE EYE OF THE BEHOLDER

FLAVORS OF DATA

Flavors of data

- It is important to understand the different flavors of data for several reasons. Not only will the type of data dictate the methods used to analyze and extract results, knowing whether the data is unstructured or perhaps quantitative can also tell you a lot about the real-world phenomenon being measured.

- We will look at the three basic classifications of data:
  - Structured vs unstructured (sometimes called organized vs unorganized)
  - Quantitative vs qualitative
  - The four levels of data

data

STRUCTURED VERSUS UNSTRUCTURED

Structured versus unstructured data

- Structured (organized) data:
  - data that can be thought of as observations and characteristics. It is usually organized using a table method (rows and columns).

- Unstructured (unorganized) data:
  - data that exists as a free entity and does not follow any standard organization hierarchy.

Structured versus unstructured data

Here are a few examples that could help you differentiate between the two:

- Most data that exists in text form, including server logs and Facebook posts, is unstructured.
- Scientific observations, as recorded by careful scientists, are kept in a very neat and organized (structured) format.
- Tweets, e-mails, literature, and server logs are generally unstructured forms of data.

Structured versus unstructured data

- Structured data is generally thought of as being much easier to work with and analyze. Most statistical and machine learning models were built with structured data in mind and cannot work on the loose interpretation of unstructured data.

- So why even talk about unstructured data? Because it is so common! Most estimates place unstructured data as 80-90% of the world's data. This data exists in many forms and for the most part, goes unnoticed by humans as a potential source of data.

Structured versus unstructured data

- While a data scientist likely prefers structured data, they must be able to deal with the world's massive amounts of unstructured data.
- So, with most of our data existing in this free-form format, we must turn to pre-analysis techniques, called preprocessing, in order to apply structure to at least a part of the data for further analysis.
- The next chapter will deal with preprocessing; for now, we will consider the part of preprocessing wherein we attempt to apply transformations to convert unstructured data into a structured counterpart.

data PREPROCESSING

Data preprocessing

- When looking at text data (which is almost always considered unstructured), we have many options to transform the set into a structured format. We may do this by applying new characteristics that describe the data.

- A few such characteristics are as follows:
  - Word/phrase count
  - The existence of certain special characters
  - The relative length of text
  - Picking out topics.

Data preprocessing : Example

- I will use the following tweet as a quick example of unstructured data, but you may use any unstructured free-form text that you like, including tweets and Facebook posts.

" This Wednesday morn, are you early to rise? Then look East. The Crescent Moon joins Venus & Saturn. Afloat in the dawn skies".

"صباح يوم الأربعاء ، هل أنت مبكّر للاستيقاظ؟ ثم انظر إلى الشرق. الهلال القمر ينضم إلى كوكب الزهرة وزحل. طافية في سماء الفجر ".

- Pre-processing is necessary for this tweet because a vast majority of learning algorithms require numerical features.
- Pre-processing allows us to explore features that have been created from the existing features. For example, we can extract features such as word count and special characters from the mentioned tweet. Then, we look at a few features that we can extract from the existing features.

Data preprocessing: Word/phrase counts

- We may break down a tweet into its word/phrase count.
- We can represent this tweet in a structured format, as follows, thereby converting the unstructured set of words into a row/column format:

	this	wednesday	morn	are	this wednesday
Word Count	1	1	1	1	1

- Note that to obtain this format we can utilize `scikit-learn's CountVectorizer`.

Data preprocessing : Presence of certain special characters

- We may also look at the presence of special characters, such as the question mark and exclamation mark.
- The appearance of these characters might imply certain ideas about the data that are otherwise difficult to know.

- For example, the fact that this tweet contains a question mark might strongly imply that this tweet contains a question for the reader. We might append the preceding table with a new column, as shown:

	this	wednesday	morn	are	this wednesday	?
Word Count	1	1	1	1	1	1

Data preprocessing : Relative length of text

- This tweet is 121 characters long.

```
>> len("This Wednesday morn, are you early to rise? Then look East. The Crescent Moon joins Venus & Saturn. Afloat in the dawn skies.")
```

- The average tweet, as discovered by analysts, is about 30 characters in length. So, we might impose a new characteristic, called relative length, (which is the length of the tweet divided by the average length), telling us the length of this tweet as compared to the average tweet. This tweet is actually 4.03 times longer than the average tweet, as shown: $\frac{121}{30} = 4.03$

- We can add yet another column to our table using this method:

	this	wednesday	morn	are	this wednesday	?	Relative length
Word Count	1	1	1	1	1	1	4.03

Data preprocessing : Picking out topics

- We can pick out some topics of the tweet to add as columns. This tweet is about astronomy, so we can add another column, as illustrated:

	this	wednesday	morn	are	this wednesday	?	Relative length	Topic
Word Count	1	1	1	1	1	1	4.03	astronomy

- And just like that, we can convert a piece of text into structured/organized data ready for use in our models and exploratory analysis.
- Topic is the only extracted feature we looked at that is not automatically derivable from the tweet. Looking at word count and tweet length in Python is easy; however, more advanced models (called topic models) are used to derive and predict topics of natural text as well.

QUANTITATIVE VERSUS QUALITATIVE DATA

Quantitative versus qualitative data

- When you ask a data scientist, "what type of data is this?", they will usually assume that you are asking them whether or not it is mostly quantitative or qualitative. It is likely the most common way of describing the specific characteristics of a dataset.

- These two data types can be defined as follows:
  - Quantitative data: This data can be described using numbers, and basic mathematical procedures, including addition, are possible on the set.

  - Qualitative data: This data cannot be described using numbers and basic mathematics. This data is generally thought of as being described using "natural" categories and language.

Example – coffee shop data

- Say that we were processing observations of coffee shops in a major city using the following five descriptors (characteristics):

- **Data: Coffee Shop**
  - Name of coffee shop
  - Revenue (in thousands of dollars)
  - Zip code
  - Average monthly customers
  - Country of coffee origin

- Which of the above characteristics can be classified as either quantitative or qualitative?

THE FOUR LEVELS OF DATA

The four levels of data

- It is generally understood that a specific characteristic (feature/column) of structured data can be broken down into one of four levels of data:
  - The nominal level
  - The ordinal level
  - The interval level
  - The ratio level

The nominal level

- The first level of data, the nominal level, (which also sounds like the word name) consists of data that is described purely by name or category.
- Basic examples include gender, nationality, species, etc. They are not described by numbers and are therefore qualitative.

- The following are some examples:
  - A type of animal is on the nominal level of data.
  - A part of speech is also considered on the nominal level of data.
- Of course, being qualitative, we cannot perform any quantitative mathematical operations, such as addition or division. These would not make any sense..

The nominal level

Mathematical operations allowed

- We cannot perform mathematics on the nominal level of data except the basic equality and set membership functions, as shown in the following two examples:
  - Being a tech entrepreneur is the same as being in the tech industry, but not vice versa
  - A figure described as a square falls under the description of being a rectangle, but not vice versa

Measures of center

- A measure of center is a number that describes what the data tends to. It is sometimes referred to as the balance point of the data. Common examples include the mean, median, and mode.
- In order to find the center of nominal data, we generally turn to the mode (the most common element) of the dataset.
- Measures of center such as the mean and median do not make sense at this level as we cannot order the observations or even add them together.

The ordinal level

- The nominal level did not provide us with much flexibility in terms of mathematical operations due to one seemingly unimportant fact—we could not order the observations in any natural way.
- Data in the ordinal level provides us with a rank order, or the means to place one observation before the other; however, it does not provide us with relative differences between observations, meaning that while we may order the observations from first to last, we cannot add or subtract them to get any real meaning.
- Examples
  - The Likert is among the most common ordinal level scales. Whenever you are given a survey asking you to rate your satisfaction on a scale from 1 to 10, you are providing data at the ordinal level. Your answer, which must fall between 1 and 10, can be ordered: eight is better than seven while three is worse than nine. However, differences between the numbers do not make much sense. The difference between a seven and a six might be different than the difference between a two and a one.

The ordinal level

Mathematical operations allowed

- We are allowed much more freedom on this level in mathematical operations. We inherit all mathematics from the ordinal level (equality and set membership) and we can also add the following to the list of operations allowed in the nominal level:
  - Ordering
  - Comparison

Measures of center

- At the ordinal level, the median is usually an appropriate way of defining the center of the data. The mean, however, would be impossible because division is not allowed at this level.
- We can also use the mode like we could at the nominal level.

The interval level

- Now we are getting somewhere interesting. At the interval level, we are beginning to look at data that can be expressed through very quantifiable means, and where much more complicated mathematical formulas are allowed.
- The basic difference between the ordinal level and the interval level is, well, just that—difference.
- Data at the interval level allows meaningful subtraction between data points.

- Example
  - Temperature is a great example of data at the interval level. If it is 100 degrees Fahrenheit in Texas and 80 degrees Fahrenheit in Istanbul, Turkey, then Texas is 20 degrees warmer than Istanbul.
  - This simple example allows for so much more manipulation at this level than previous examples.

The interval level

Mathematical operations allowed

- We can use all the operations allowed on the lower levels (ordering, comparisons, and so on), along with two other notable operations:
  - Addition
  - Subtraction

Measures of center

- At this level, we can use the median and mode to describe this data; however, usually the most accurate description of the center of data would be the arithmetic mean.
- Recall that the definition of the mean requires us to add together all the measurements. At the previous levels, addition was meaningless; therefore, the mean would have lost extreme value. It is only at the interval level and above that the arithmetic mean makes sense.

The interval level

Measures of variation

- A measure of variation (like the standard deviation) is a number that attempts to describe how spread out the data is. Along with a measure of center, a measure of variation can almost entirely describe a dataset with only two numbers.
- Standard deviation is the most common measure of variation of data at the interval level and beyond. The standard deviation can be thought of as the "average distance a data point is at from the mean". While this description is technically and mathematically incorrect, it is a good way to think about it.
- The formula for standard deviation can be broken down into the following steps:
  1. Find the mean of the data.
  2. For each number in the dataset, subtract it from the mean and then square it.
  3. Find the average of each square difference.
  4. Take the square root of the number obtained in step three. This is the standard deviation.

The ratio level

- Finally, we will take a look at the ratio level. After moving through three different levels with differing levels of allowed mathematical operations, the ratio level proves to be the strongest of the four.
- Not only can we define order and difference, the ratio level allows us to multiply and divide as well. This might seem like not much to make a fuss over but it changes almost everything about the way we view data at this level.

- Examples
  - While Fahrenheit and Celsius are stuck in the interval level, the Kelvin scale of temperature boasts a natural zero. A measurement zero Kelvin literally means the absence of heat. It is a non-arbitrary starting zero. We can actually scientifically say that 200 Kelvin is twice as much heat as 100 Kelvin.
  - Money in the bank is at the ratio level. You can have "no money in the bank" and it makes sense that $200,000 is "twice as much as" $100,000.

The ratio level

Measures of center

- The arithmetic mean still holds meaning at this level, as does a new type of mean called the geometric mean. It is the square root of the product of all the values.
- For example, in fridge temperature data, we can calculate the geometric mean as shown here:

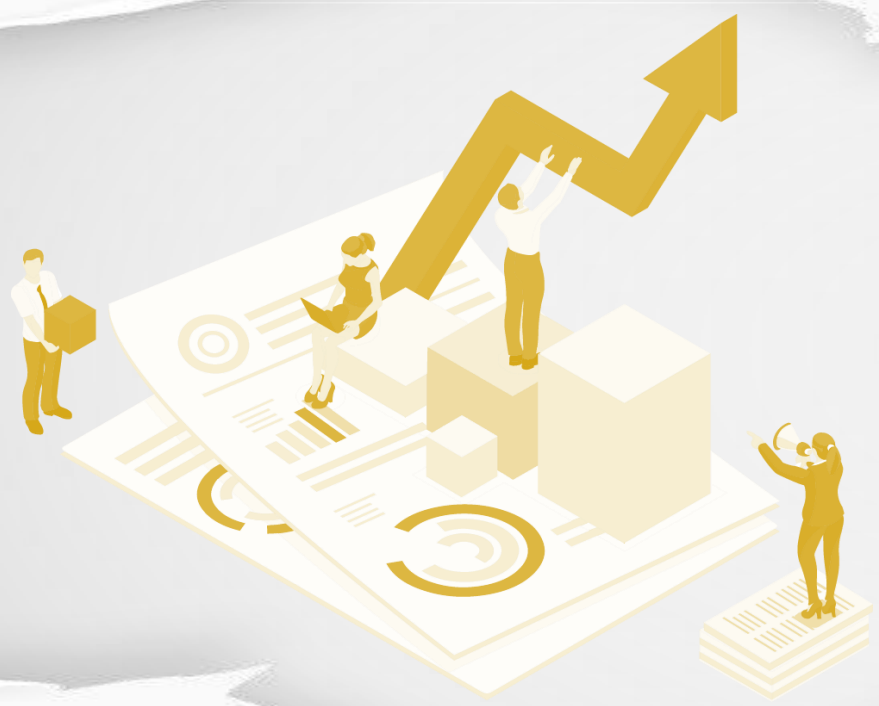
```
import numpy

temps = [31, 32, 32, 31, 28, 29, 31, 38, 32, 31, 30, 29, 30, 31, 26]

num_items = len(temps)
product = 1.

for temperature in temps:
    product *= temperature

geometric_mean = product**(1./num_items)

print geometric_mean   # == 30.634
```

The ratio level

Problems with the ratio level

- Even with all of this added functionality at this level, we must generally also make a very large assumption that actually makes the ratio level a bit restrictive.
- Data at the ratio level is usually non-negative.
- For this reason alone, many data scientists prefer the interval level to the ratio level.
- The reason for this restrictive property is because if we allowed negative values, the ratio might not always make sense.
- Consider that we allowed debt to occur in our money in the bank example. If we had a balance of $50,000, the following ratio would not really make sense at all:

$$\frac{\$50,000}{-\$50,000} = -1$$



DATA IS IN THE EYE OF THE BEHOLDER

Data is in the eye of the beholder

- It is possible to impose structure on data.
- For example, while I said that you technically cannot use a mean for the one to five data at the ordinal scale, many statisticians would not have a problem using this number as a descriptor of the dataset.
- The level at which you are interpreting data is a huge assumption that should be made at the beginning of any analysis. If you are looking at data that is generally thought of at the ordinal level and applying tools such as the arithmetic mean and standard deviation, this is something that data scientists must be aware of. This is mainly because if you continue to hold these assumptions as valid in your analysis, you may encounter problems. For example, if you also assume divisibility at the ordinal level by mistake, you are imposing structure where structure may not exist.

s.zahrani@tu.edu.sa