5 Data representations

Most artificial intelligence systems operate with some type of data. For example, systems might be designed to understand, act upon, or generate sound, images, or text. Exactly how the data is represented and processed can have a large influence on how the system can interact with the data, so it is important to understand and think carefully about data representation.

5.1 Data representations in the computer

In a computer all data is represented in a binary form, as a sequence of zeros and ones. When an analogue signal from the real world (such as a sound or an image) is captured by a sensor, it is transformed into a binary representation before it goes into the computer.

5.1.1 Numbers

A *bit* is a single value, zero or one, in the computer. Usually multiple bits are put together to represent a number. Most computers can work with two fundamentally different types of numbers:

Integer An integer is a whole number. If we use 8 bit to represent an integer, we can make $2^8 = 256$ different numbers. The bit sequence 0000 0000 represents the number 0, 0000 0001 represents 1, and 0000 0010 represents 2 and so on. An integer can either be defined as signed or unsigned: An unsigned 8-bit integer can represent any value between 0 and 255, whereas a signed 8-bit integer can represent values between -128 and 127. If we need to represent larger values, we can use a 16-bit integer (or 32 bits or 64 bits etc.)

Floating point A floating point number is a way to approximately represent real numbers. The numbers that can be represented are on the form (significand $\times$ 2^{exponent}) where the significand and exponent are signed integers. In a typical 64-bit floating point number, the significand is represented by 53 bits and the exponent has 11 bits. This means that numbers have around 16 decimal digits of precision (because $2^{53} \approx 10^{16}$) and that the largest number that can be represented is around 10^{308} (because $2^{2^{10}} \approx 10^{308}$).

5.1.2 *Characters*

Characters (letters and symbols) are represented as integers in the computer. Each possible character is given a unique number, and the list of which number corresponds to which character is called a *code page*. One of the classical code pages is ASCII, which is a 7-bit code designed to represent the 26 letters a–z, the capital letters A–Z, the numbers 0–9, a number of special graphical symbols, as well as some control characters such as a line shift character etc. For example, in the ASCII code page the letter 'a' has the integer value 97.

The ASCII code page cannot represent many of the special characters and symbols used in different languages, so many extensions have been developed and standardized. Many of the modern code pages remain compatible with the ASCII table. *Unicode* is a modern standard for character encoding that can represent most of the characters needed in all of the World's writing systems. The Unicode standard contains different encoding formats, one of which is UTF-8. This is a variable width character encoding, where most common characters are encoded by 8 bits, whereas more rarely used characters require up to 32 bits.

5.2 Measurement

In order to operate computationally or mathematically on data, we need to carefully consider how we measure. Roughly speaking, *measurement* is the process of assigning numbers to observations. To determine how to do this best, requires both knowledge about the data (what it means, how it is recorded, etc.) and how it is to be processed subsequently.

5.2.1 *Quantitative and qualitative data*

One way to characterize data is to consider whether it is *qualitative* or *quantitative*.

Qualitative Data that is non-numerical and recorded in free form is called *qualitative*. It is often gathered to gain insight in some phenomenon and answers questions such as "why?", "how?", and "what?".

Quantitative Data that is numerical data and recorded in a standardized way is called *quantitative*. It is often gathered to explain, predict, or control some phenomenon and answers questions such as "how many?", "how much?", and "how often?".

Example 5.1 A picture of some food items

Let us consider the photo in Fig. 5.1. An example of *qualitative* data describing the picture could be:

> A brown cupcake with the American flag on a plate, and a half-full glass bottle of Coke.

An example of *quantitative* data describing the picture could be:

Number of cupcakes	1
Color of cupcakes	brown
Number of bottles	1
Amount of Coke	17 [cl]

Figure 5.1: A picture of some food items.

Figure 5.2: A picture of some bottles.

5.2.2 *Types of quantitative data*

Quantitative data can be further characterized as either *numeric* or *categorical* data.

Numeric Data that is represented by numbers is called *numeric*. Numeric data can further be characterized as discrete or continuous.

Discrete data can only take certain values, and includes things we can count or assign to a category.

Continuous data can take any numerical value (within a range), and includes things we can measure in a continuum.

Categorical Data that puts each observation into one of a set of mutually exclusive groups is called *categorical*. Categorical data can be further characterized as either nominal or ordinal.

Ordinal categories can be logically arranged in a sequence.

Nominal categories serve only to identify or label each group, and there is no further structure in the set of categories.

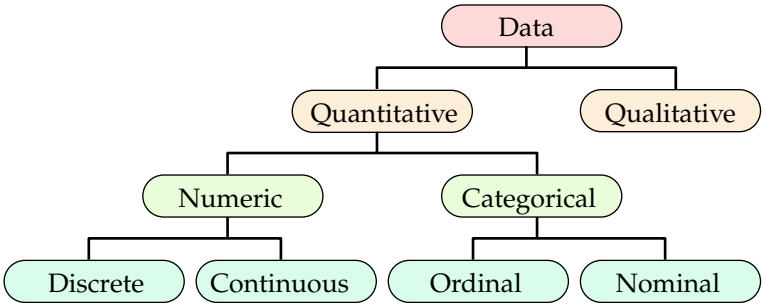

Figure 5.3: Categorization of different data types.

Fig. 5.3 outlines the different types of data discussed above. [1]

[1] In the research literature, there are also more advanced and detailed ways to characterize different data types with "Stevens's typology" perhaps the best known classification scheme (also known as a *level of measurement* or a *scale of measure*).

Example 5.2 Examples of types of quantitative data

Numeric-Discrete

- Number of students in a class.
- The outcome of rolling a die.

Numeric-Continuous

- Height of a person.
- Time of a 100 meters race.

Categorical-Ordinal

- Ratings on a scale (e.g. Good, Average, Bad).
- Letter grades (e.g. A, B, C, D, F)

Categorical-Nominal

- Brand of car (e.g. Seat, VW, Toyota, etc.)
- Marital status (e.g. Married, Single, Widowed).

5.3 Text representation

In the computer, a text string is simply a sequence of characters. In many applications, the natural level of analysis is the level of *words*, so often we think of text as a sequence of words. Just like characters can be represented using a code page where each character is assigned a unique number, *words* can also be represented in a similar manner. The total number of different words that exist is large but finite, so in theory one could simply use a huge dictionary of possible words to encode each word as a number. However, one would quickly run into problems with rare words, such as uncommon names, names of businesses, and made-up words that might be used in some literature. Nevertheless, this approach is often used in practise.

5.3.1 Bag-of-words

The *bag-of-words* representation is a very simple way to represent a text document. We simply count how many times each possible word occurs in the document, and store only the counts, ignoring the order in which the words occur.

Some words carry more information than others: For example, common words such as "the" and "and" might occur in most documents, but are not so important. In many practical representations, such common words are simply ignored.

Another issue is that many words essentially carry the same information: For example, different spellings or grammatical tenses of the same word are often semantically interchangable. To handle this issue, such words can simply be combined so that for example all words starting with *reali* are combined which would cover the words *realise*, *realize*, *realised*, *realization*, etc. This methods is known as *stemming* as it reduces each word to its stem.

Example 5.3 Document retrieval

One of the important tasks in natural language processing is information retrieval, where one wants to find a docu-

ments that match some search query. For example, this is the task that Internet search engines solve when we search for web pages that macth a search string.

With the bag-of-words representation we could find the documents that are relevant for the query simply by counting how many words the document and the query have in common. However, a simple count of the number of co-occuring words might not be optimal if the documents have very different lengths, because long documents would always tend to have more words and thus a higher chance of matching the query. This issue could be handled by normalizing the counts by the document length.

5.3.2 *n-grams*

While the bag-of-words representation is attractive because of its simplicity, it might be too crude to completely ignore the order in which the words occur in the text. One way to keep the representation simple while not completely ignoring the word order is to use an *n-gram* representation: Here we count the number of occurences of each possible length n sequence. For example, in a a bi-gram (2-gram) representation we would count how many times each possible pair of words occur in sequence.

Example 5.4

Bag-of-words and bi-gram representations Let us consider the following small piece of text:

> To be or not to be—that is the question.

In a bag-of-words representation, this would be represented as follows (in no particular order, but here alphabetized)

> be, is, not, or, question, that, the, to

Whereas in a bigram representation, we have the following

> be or, be that, is the, not to, or not, that is, the question, to be

Problems

1. Descibe the picture in Fig, 5.2 qualitatively and quantitatively.

2. Consider the following bag of words

   am, and, can, check, commencing, control, earth, eight, far, five, for, four, god, ground, here, i, liftoff, major, moon, nine, now, one, planet, seven, six, take, tell, ten, this, though, three, tom, two, you, your, a, above, am, and, be, blue, can, capsule, circuit, countdown, dare, dead, different, do, door, engines, feeling, floating, go, grade, hear, helmet, her, here, hundred, if, ignition, in, is, it, know, knows, leave, look, love, made, may, me, miles, most, much, my, nothing, on, one, papers, past, peculiar, pills, protein, put, really, round, she, shirts, sitting, something, spaceship, stars, stepping, still, the, there, think, thousand, through, time, tin, to, today, very, want, way, wear, which, whose, wife, with, world, wrong, you, your

   What do you think the original text is about?

Solutions

1. Example of qualitative description:

   > Four full glass bottles of Coke.

   Example of quantitative description:

Number of cupcakes	0
Color of cupcakes	*not applicable*
Number of bottles	4
Amount of Coke	132 [cl]

2. It is the lyrics from the David Bowie song *Space oddity* which is about an astronaut who is cut off communication with earth and floats into space.