#h1 Chapter 6 - Data Types

#h2 Words, Numbers, Categories, and Scales

#h3 6.1 – Introduction

#pg In this first module, we will spend some time defining the various terms and discussing where they overlap and where they do not. On more than one occasion, people working with machine learning engineers and data scientists expressed that they were unsure what the difference is between data science, machine learning, and artificial intelligence. So, it is worth spending some time delineating these concepts. All machine learning relies on some type of data or a mixture of different data types. Understanding the characteristics and limitations of your data type is key to building well-behaving machine learning algorithms. Too often, failure to align data with the algorithms yields models that underperform and show high variance and bias. Good practice in any machine learning setting is scrutinizing your data, noting outliers, imbalances, and interdependence between measurements. In addition, specific data types call for a particular way of normalizing them to mitigate any effects of different scaling between inputs. This chapter deals with common and sometimes uncommon data types. Data types differ, but vastly different fields of scientific inquiry might still produce a similar, if not identical, data type as measurements are recorded and stored as data. Someone might argue that gathered data from peering out into space using optical telescopes is quite different than the minute details of cancer cells captured with an optical microscope. I agree, but that is not what I mean by data type. An optical telescope is a data source. I used the word optical because the visual detail from the stars and the cells is still captured using imagery. If I compared radio telescopes with electron microscopes instead of optical ones, I would be more hesitant to equate them in terms of type. Both methods use a lot of postprocessing to turn their signal into actual images. Postprocessing may create artifacts that our machine learning algorithm we designed to analyze photographic images fails to understand or incorrectly understand. But optical telescopes produce images, and so do optical microscopes. Images are a unique data type, requiring a different approach than we would use for numerical or textual data.

#h3 6.2 - Numerical Data

#pg Numerical data is probably still the most ubiquitous data source used by machine learning algorithms. We can easily manipulate and transform numerical, and we created many computer algorithms specifically to apply complex mathematical operations to numerical data (e.g., the fast Fourier transform). It also gets rid of the problem explained above. A patient's heart rate is a complex function of their physiology, but since the data only vary along one dimension, we can use that as a numerical indicator. The number of penguins accidentally falling over on their belly is also a complex function of the current situation at their present location. But again, the best that signal has to offer is an integer value (assuming here, for the sake of simplicity, that penguins fall or do not fall). But not all numerical data is created equal. They differ in their range, domain, skew, and other moments of the distribution across data points. Most importantly, the mathematical operations we can apply to numerical data will vary. In general, the rules of mathematics apply to all measured physical quantities (like velocity, mass, and volume). However, derived measures and numbers, like IQ, must be treated with caution. We will discuss this in greater detail later. Another critical aspect of numerical data is the nature of the generative model that gave rise to it. Most algorithms assume that inputs have some known or approximate distribution, such as the normal, binomial, or multinomial distribution. Suppose our data deviates too much from these assumptions. In that case, some of our machine learning might be difficult or result in a biased model. Our models optimize their parameters based on statistical assumptions, like the observed error distributed uniformly and symmetrically around a stationary mean. Numerical data often (but not always) benefit from being normalized before training an ML algorithm. Normalization aims to scale the different input signals to have comparable means and variance and mitigate the effect of possible outliers. And a good starting point is typically the z-scoring approach which we discuss later.

#im ../assets/figures/006/006-01.png 50 256 Figure 6.1 - The first 10000 digits of Pi. Pi is a mathematical constant. It is defined in Euclidean geometry as the ratio of a circle's circumference to its diameter. As an irrational number, π cannot be expressed as a common fraction, although fractions such as 22/7 are commonly used to approximate it. Equivalently, its decimal representation never ends and never settles into a permanently repeating pattern. It's decimal (or other base) digits appear randomly distributed and are conjectured to satisfy a specific kind of statistical randomness.

#h3 6.3 - Categorical Data

#pg Proper categorical data does not allow us to apply mathematical transformations to them. The label 'male' simply is not a quantity. It is not more or less than the label 'female'. You can only derive equivalence from it (male is male, but male is not female). Other categorical values might look like numbers, but close inspection reveals that they are not numerical. The label' income > $100.000' can be higher than the label 'income > $50.000 and income < $100.000', but we cannot tell from the label alone how much higher, as the labels specify a range. Similarly, we can recode 'October' to 10 (the 10th month of the year). September would be 9, so is October one unit more than September? We cannot say from the month along, as the September in question could have been in the year prior or the year after.

#im ../assets/figures/006/006-02.png 50 256 Figure 6.2 - Categories of the animal kingdom. All organisms are classified into increasingly smaller subcategories in biology, ranked by kingdom, phylum, class, order, family, genus, and species.

#h3 6.4 – Scales

#pg To interpret your data correctly, spend some time understanding the scale of your data. The scale of your data dictates what you can do with it and how to interpret the results of your model that uses that same data as inputs (and outputs). Experimental science distinguishes between 4 scales, from purely categorical to purely numerical: nominal, ordinal, interval, and ratio. The nominal scale represents purely categorical data. There is no quantitative comparison possible between members of the category. Red is not more or less than green. And a list of names can be sorted alphabetically, but this ordering is arbitrary. One level above we nominal we find the ordinal scale. We can compare data points at this scale, but only in relative terms. We can use ordering, but the difference between any two classes is not meaningful from a mathematical point of view. For example, education level has a natural order. For example, 'some high school', 'high school diploma', 'college degree', or 'graduate degree'. While we all agree that this is a logical ordering, there is no number we can put on the differences between the classes that allow them to be compared mathematically. Next up is the interval scale. We can compare differences between pairs of points numerically on an interval scale. Consider a credit card score. The difference between 500 and 600 is 100, the same as between 600 and 700. The units of axes are fixed and the same across the entire domain of the scale. However, it is still not a purely numerical value because no meaningful zero exists. I am sure somebody can give you a score of 0, but it does not convey the absence of something. A credit card score of 200 isn't twice as high as a credit score of 100. Finally, ratio scales incorporate all the necessary numerical properties to allow a broad range of mathematical operations. The values are real, although they do not necessarily need to be floating point. They have a natural ordering, meaningful differences, and a true zero. Think of a white blood cell count or a cholesterol level. One thousand blood cells are twice as many blood cells as 500. And a cholesterol level of 100 is four times higher than a level of 25.

#im ../assets/figures/006/006-03.png 50 256 Figure 6.3 - Different scales or levels of measurement of data. Understanding your data in terms of the four different scales (nominal, ordinal, interval, and ratio) will guide how you use your data.

#h3 6.5 - Mixed Data

#pg Often, our data will be a mixture of the two above types. For example, demographic data typically has both numerical (age, height, weight) and categorical (gender, race, ethnicity). In this case, our choice of machine learning algorithm can guide how we align the different data types. A standard approach is to turn purely categorical labels into numerical-like data points. Should we decide on a more rule-based ML algorithm like decision trees, continuous data can be incorporated at the expense of the algorithm's efficiency.

#h3 6.6 - Event data

#pg Event data typically comes in logs that record certain events of interest. Event data like logs are often high in volume (think of Facebook posts, logins, tweets) and typically contain a mixture of data types (tweet geolocation, tweet text, gender of the person tweeting). Practically, what is unique about event data is that it usually requires heavy joins, as we have distinct types of data points for particular types of events. In addition, we can expect the data to come in a more loosely defined schema, like JSON or XML objects, encoding only values of interest for that event, again prompting the necessity of joining logs of different event types. However, from the event logs, we can reconstruct fascinating timelines of how certain large-scale phenomena occur and spread, how information flows through a system, and detect unusual events to flag and investigate.

#im ../assets/figures/006/006-04.png 50 256 Figure 6.4 - Online services like Twitter, Facebook, and Amazon have massive amounts of event data. Every click, like, comment, post, view, and even letter, word, and keystroke (including those deleted before they were even posted or sent) are stored for later analysis and the development of predictive models.

#h3 6.7 - Timeseries Data

#pg Timeseries data are numerical and always have a natural ordering in time. Examples include stock prices, the rise in population numbers, and brain activity recorded by electroencephalography (EEG). This natural ordering is essential because it means that time points from a single time series are not independent. Another way of thinking about this is to assume that specific processes generate periodicity, seasonality, and other predictable fluctuations. Treating the data points within a single time series as independent will obscure these interesting fluctuations and create significant statistical issues in our machine learning algorithm.

#im ../assets/figures/006/006-05.png 50 256 Figure 6.5 - Given that it contains some regularities and periodicities, rather than just representing random fluctuations, time-series data, like stock prices, can potentially predict from past observations, a forecasting method.

#h3 6.8 - Image Data

#pg Images (stills or videos) are one of the more interesting data types. Images are typically encoded as 3D matrices, specifying the width and height of the image across several channels. Historically, these channels are red, green, and blue. Therefore, we specify a pixel's color as an RGB triplet: (100%, 100%, 0%), for example, will be a bright yellow (full green and red, no blue). We can transform the RGB triplets to yield other quantifications, such as hue, saturation, and luminance. Remember, there is much more to color vision than just these three values. Mapping these values into our perception of color and brightness is extremely complicated. A standard 4K image at a 24bit depth (8bits per channel) is one out 3840 x 2160 x 3 x 256 images possible at the resolution. In other words, we can create 6,370,099,200 numerically distinct images (## 6.4 billion). For several reasons, comparing a camera with the human eye and the human visual system is not straightforward. Still, a 4K image per frame is likely to be similar by some order of magnitude to the data load on our eyes. What is key to working with images is understanding that natural images (an image that occurs or can occur in the world) occupy a much smaller subspace in this larger space of all images.Nonetheless, even this smaller subspace is still too large to consider capturing completely. In addition, another vital point is that images do not consist of independent observations (pixels in this case). Look at a photograph you took lately. You can quickly tell that the value of a pixel correlates highly with the pixel next to it but correlates less and less with pixels further away. This spatial dependence is a blessing in disguise. The problem with applying ML to images is their high dimensionality: each pixel is an input dimension, so a 4K image already consists of 12 million input dimensions. However, the spatial dependence across pixels allows us to reduce this high dimensionality, typically using a dimensionality reduction technique like PCA.

#im ../assets/figures/006/006-06.png 50 256 Figure 6.6 - For now, images still consist of pixels. That might change soon. We have long known the human visual system does something very similar. Images can be compressed quite a bit in the digital realm without losing much detail and quality. Compression relies on correlations between pixels, meaning we can represent pixels as a mixture of more complex representations. Recent advances in image classification neural network models have shown that these complex, lower dimensional representations arise naturally as the neural networks learn and improve their performance.

#h3 6.9 - Textual Data

#pg Textual data is a unique source of data, and as such, most of the methodologies and algorithms used have been specifically designed around it. Substantial effort must go into preprocessing of textual data to realize its full potential. Such preprocessing ensures we have properly formed words and sentences, removing any surplus characters, ambiguities, and spelling errors. It also includes normalizing the data by mapping words with the same linguistic meaning to a single example, removing things like tense and collapsing plural/singular forms of words. Natural Language Processing (NLP) algorithms can use a 'Bag of Words' perspective: a document, phrase, tweet, or conversation is a collection of words, and the exact order they occur within a document is ignored. We typically use 'Bag of Words' techniques when categorizing and grouping articles based on their content; a field referred to as topic modeling. In contrast, probabilistic models focus on the probabilities of certain words or sequences of words, given what has already been said or written. These models allow for more generative algorithms, like chatbots and text completion. Using graph theory, we can approach the relationship between documents and the information and knowledge in them, creating networks that show clusters of communities of certain words that together define a concept or semantic structure. Finally, vectorization transforms our text data into convenient vectors that allow us to apply numerical methods and metrics.

#im ../assets/figures/006/006-07.png 50 256 Figure 6.7 - One of the remaining Gutenberg Bibles. The Gutenberg Bible was the earliest major book printed using mass-produced movable metal type in Europe. It marked the dawn of the age of printed books in the West. The book is valued and revered for its high aesthetic and artistic qualities and historical significance.

#h3 6.10 - Connected Data

#pg Connected or networked data captures the relationships between entities, like people, places, and spatial proximity. We can easily conceptualize such data as graphs where the entities are nodes, and the links specify the relationship between the nodes ('is friends with,' 'is a subset of,' etc.). Connected data typically needs to be derived from collected data. The metrics and algorithms we use to describe and predict from it are quite different than the methods used for more traditional data sources. In addition, graph data requires rethinking how we store and operate on such data effectively.

#im ../assets/figures/006/006-08.png 50 256 Figure 6.8 - Zachary's Karate club Graph. The network captures 34 karate club members, documenting links between pairs of members who interacted outside the club. A conflict arose between the administrator "John A" and instructor "Mr. Hi" (pseudonyms) during the collection of the data. This conflict led to the split of the club into two. About half of the members aligned with Mr. Hi; the remaining members found a new instructor or gave up karate. Based on collected data Zachary correctly assigned all but one member of the club to the groups they joined after the split.

#h3 6.11 - Speech Data

#pg Typically, we think of electromagnetic radiation and pressure waves as light and sound, respectively. Light gives us vision, which we like to think of as images. In contrast, sound gives us audition. But our viewpoint is limited. Some animals produce sounds outside our audible range, and some can detect electromagnetic energy outside our visual range, like ultraviolet and infrared. We have spoken on images previously, but it is worth noting that sound carries an essential signal for us: the speech of others. In that sense, human speech (and, by that token, animal vocalization) is unique. Although still recorded as soundwaves, we traditionally combine human speech into larger, more meaningful, distinct units. For example, by combining short sections of speech waveform into phonemes the smallest perceptually distinct units of sound, like 'pet,' 'bar,' 'bat'), phonemes into morphemes (the smallest meaningful unit of a language, like 'come 'or 'ing' in coming) and morphemes into distinct words. From a content perspective, nothing is lost in this more succinct representation, although we lose other useful information that conveys emotion through tone of voice.

#im ../assets/figures/006/006-09.png 50 256 Figure 6.9 - Spectrogram of a bit of human speech. Human speech. In general, the fundamental frequency of the complex speech tone – also known as the pitch or f0 – lies in the range of 100-120 Hz for men, but variations outside this range can occur. For women, the f0 for women is approximately one octave higher. For children, f0 is around 300 Hz.

#h3 6.12 - DNA and biological topology Data

#pg Our DNA and its derivatives RNA consist of a long string of 4 possible base pairs (A, C, P, T). Within the entire string, particular substrings represent genes: the smallest unit of cellular instructions, other substrings act as modifiers, allowing suppression or activation of nearby genes or compression of unused genetic material. DNA is thus like a bitstring, with the difference between that bit strings are, per definition, binary: each element in the sequence is 0 or 1, whereas DNA is quaternary.

#im ../assets/figures/006/006-10.png 50 256 Figure 6.10 - Human chromosomes. A chromosome is a long DNA molecule with part or all the genetic material of an organism. Chromosomes in humans can be divided into two types: autosomes (body chromosome(s)) and allosome (sex chromosome(s)). Certain genetic traits are linked to a person's sex and are passed on through the sex chromosomes. The autosomes contain the rest of the hereditary genetic information. All act in the same way during cell division. Human cells have 23 pairs of chromosomes (22 pairs of autosomes and one pair of sex chromosomes), giving a total of 46 per cell. In addition, human cells have hundreds of copies of the mitochondrial genome. Sequencing the human genome has provided much information about each chromosome.

#br

#pg A subtle difference, it seems, but it requires us to rethink what methods we usually use for bitstring data, like measures of similarity and correlation. Biological structures, like proteins, contain just these simple building blocks DNA has to offer. However, combined into extremely topological structures, called folding, it is hard to decipher how the structure even begins to give rise to the function and behavior of the protein. The problem is so complicated that even our most advanced machine learning models struggle, although we have made some breakthroughs recently. However, the hope is that novel promising technologies such as quantum computing, and crowdsourcing can offer more optimal solutions to this challenging but valuable problem faster. If we can predict the function of a newly created protein from its topological design, we bypass the need for costly and long-lasting research into drug development.

#h3 6.13 - Other types of data

#pg This chapter is by no means meant to be exhaustive. The internet has provided enormous amounts of textural, networked, and image data. Newer research domains in science, like DNA research and particle physics, often provide data types unique to their fields and require new methodologies within machine learning. Do not worry. We are not done analyzing and understanding them, not by a long shot.

#h3 6.14 - Conclusion

#pg Throughout this course, we will use various data sources, although the focus will primarily be numerical, textual, and image data. We are most familiar with such data sources. However, there will be a few examples of more novel data sources that have only recently been approached with machine learning methodologies. Because of that, we can focus on understanding the machine learning models we apply to the data rather than grasping a novel data type and its intricacies.

#im ../assets/figures/006/006-11.png 50 256 Figure 6.11 - Proteins fold in extraordinarily complex ways to allow for extremely diverse and complex functionality. The same set of molecules, combined and folded in slightly different ways can mean the difference between a life-saving drug, a lethal poison, or a perfectly harmless compound without any noticeable effects.

#h3 6.15 – References

#bs

#be

#bp Szeliski, R. (2010). Computer Vision: Algorithms and Applications. Germany: Springer London.

#bp Luisi, B., Travers, A., Drew, H., Calladine, C. R. (2004). Understanding DNA: The Molecule and How it Works. Netherlands: Elsevier Science.

#bp Protein Folding: Methods and Protocols. (2021). United States: Springer US.

#bp Cohen, M. X. (2014). Analyzing Neural Time Series Data: Theory and Practice. United Kingdom: MIT Press.

#bp Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. United Kingdom: MIT Press.

#bp Newman, M. (2010). Networks: An Introduction. United Kingdom: OUP Oxford.

#bp Colegrave, N., Ruxton, G. (2011). Experimental Design for the Life Sciences. United Kingdom: OUP Oxford.

#bp Eisenstein, J. (2019). Introduction to Natural Language Processing. United Kingdom: MIT Press.

#bp Kreps, J. (2014). I Heart Logs: Event Data, Stream Processing, and Data Integration. United States: O'Reilly Media.

#be