Data Scientist

Tim Hellemans

May 21, 2023

Contents

1 Statistics for Data Science

Insert drawings to explain math in data science and the whole pipeline pointing at where you use which math/skills.
Thus far, we have mainly focused on the technical skills that allow you to work with data, this includes:

- Basic python to think programmatically (if/for/while loops, variables, errors, ...).
- We zoomed in on the python you need to work with data specifically; this includes working with files/folders on your computer but also to connect to an external database, scrape information from a website etc..
- We then focused on tabular data, this is mainly done with Pandas, a python package which allows you to quickly combine data from multiple tables and process data in a table.
- We also looked at numpy which contains functionality specifically to work with tensor data, these include 1 dimensional *vectors*, but also 2 dimensional *matrices* and we also looked at three dimensional tensors in the form of RGB images.

This chapter concludes the more technical part of the course and from here on out, in this second part of the course, we are going to focus more on the conceptual side and less on the technical. We will still have many exercises in Python and you are still expected to be able to "do"everything we discuss in Python (or any other programming language) but the technical implementation will no longer be the focal point. We start this chapter by providing an overview of the position of math in the data scientist's workflow.

In the subsequent sections, we introduce basic statistics that lie at the heart of data science. These contain fundamentals of statistics (such as the definition of a probability density function, how to visualize probability distributions etc.), but also some classical statistics results that are often (mis)used in the real world. One of the most important results in statistics is the central limit theorem. This theorem is often interpreted as *you can model any reasonable random variable as a normal distribution*, but this is not at all the case! Simply being able to spot when the normal distribution is being misused can already add a lot of value to a company (see also section ref naar sectie CLT).

1.1 Required math for data science

For many aspiring data scientists, a good foundation in math may seem daunting. However, when starting to learn data science/machine learning, it is not necessary to immediately delve into learning all aspects of mathematics. While mathematics plays a fundamental role in understanding the underlying concepts of machine learning, it can be overwhelming to tackle all math topics at once. A few reasons why you shouldn't start with learning all math for data science are:

1. **Steep learning curve**: Mathematics can be complex, and diving into advanced topics without a solid foundation can make it difficult to grasp the concepts. It's better to approach mathematics progressively, starting with the basics and gradually building up your knowledge as you gain more experience with machine learning.
2. **Practicality**: Machine learning frameworks and libraries provide high-level abstractions that allow practitioners to apply machine learning techniques without extensive mathematical knowledge. These tools offer pre-implemented algorithms and functions, making it easier to get started and achieve meaningful results without delving into the underlying mathematics.
3. **Focus on application**: Initially, it's more important to understand the practical aspects of machine learning, such as data preprocessing, model selection, and evaluation. By focusing on these areas, you can gain hands-on experience and develop intuition about how machine learning works in real-world scenarios.
4. **Learning by doing**: Machine learning is a highly practical field. Engaging in projects and implementing machine learning algorithms will help you understand the core concepts and their mathematical underpinnings more effectively. This hands-on approach will allow you to see the direct application of mathematics in machine learning.

Remark. The approach we will be taking is the learning by doing, that is, we provide the math required as we come across it. Moreover, this explanation will be on the intuition rather than arithmetics. For example:
When someone asks you to compute $\int_8^1 2\sin(x) \cdot x^2 \, dx$ and *explain* how you did it, I do not expect you to be able to take out pen and paper and be able to exactly compute this integral. You should be able to (with the help of google) use Python to compute this value through the function `np.trapz`. Ideally, you should also be able to plot the function $f(x) = \sin(x) \cdot x^2$ and explain that the area under the curve on the interval $[8, 12]$ corresponds to the integral that this person is inquiring about (see also Figure 1). Ideally, you would also be able to explain how this integral comes into existence as the limit of the surface area of histograms with an increasing number of bins.

The part of Data Science for which one requires the most math is typically when developing a machine learning model. Typical projects consist of (a selection from) the following steps:

1. Collect the data from different sources (no math).
2. Merge and clean the data from all sources into a compatible format (almost no math).
3. Slice, dice and preprocess the data to be ready to be fed into some statistical tool (e.g. a machine learning algorithm). This mostly requires some linear algebra which is the mathematical field of doing computations on tensors.
4. Explore the data using standard data analysis tools (requires some probability/statistics knowledge).
5. Select and execute an appropriate method to tackle the given problem. Here the underlying probability theory can

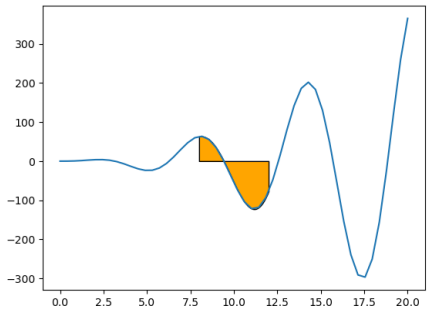

Figure 1: Integral $\int_8^1 2\sin(x) \cdot x^2 \, dx$ visualized.

be regarded as the beating heart of the method. In order to understand what you are doing and what results you are getting, you will always need a basic level of probability/statistical knowledge. In order to truly understand what is happening under the hood, you will need a lot of deep probability theory knowledge. This should not be your goal at this point.

6. Calculus and optimization come into play when we have selected a model and we now want to train our model, that is we want to tweak our conceptual model to our specific case. This means that we want to find the *optimal* parameters for our problem setting and this in turn requires some basic understanding of calculus and optimization theory in order to *kind of* know what you are doing and deep knowledge of these subjects to truly understand it (which, again, is not the goal at this point in time).

Remark. In Sections 1.6 and 1.7, we will look at our first machine learning algorithm and give a high level overview of how the above process plays out for linear regression.

We now start this chapter by first looking at the basics of (applied) statistics and look at some of the most important results in statistics. The reason why we start our quest here is three-fold:

1. Nowadays, it is a popular belief that Machine Learning is the only method that can be used to solve problems using data. This belief is however completely bogus, for many problems, simple statistics are sufficient.
2. An algorithm is often used to create predictions, these predictions are by definition uncertain. Therefore having a sense of how to talk about uncertainty is required in order to talk about predicting.
3. A basis in statistics is quite standard in many school curricula, therefore this content will be somewhat familiar to many people. However, due to this *somewhat familiarity*, the results are often used in an incorrect way leading to illogical decisions. Using a combination of data visualizations and statistical knowledge these mistakes can often be laid bare and the search for a more sensible methodology can be sought.

Remark. The field of solving problems using data is of course much broader than statistics or machine learning. There's also Operations Research, graph algorithms and simulations, just to name a few alternative methods. Which methodology should be used is highly dependent on the problem at hand and it is always an important decision to select the right framework for a given problem.

Without further ado, let us dive into an overview of the types of data one may encounter.

1.2 Types of data

Section describing all types of structured data

1.3 Estimating center locations

Section describing types of center locations and when to use which.

1.4 Variability

1.5 Exploring distributions & Bayes Theorem

In this section we talk

1.6 Linear Regression & the standard ML flow

In previous section, we learned about conditional probabilities and Bayes theorem. Conditional probabilities lie at the heart of (almost) all machine learning problems, we constantly ask ourselves questions like:

- What should the prize of a house be given its surface area, location, ...
- What is the probability of being a survivor on the titanic given your age, gender, ...
- Etc.

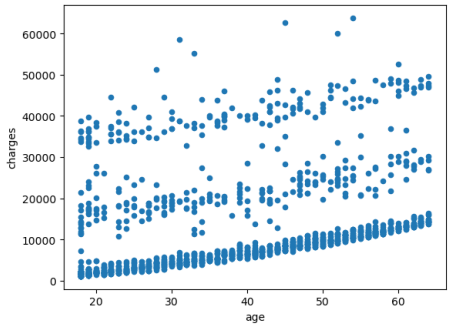

Figure 2: Scatterplot comparing the total medical cost of a person with his age.

Before we dive deeper into the *statistical* approach of solving these type of questions, let us look into the most taught machine learning method in existence: *Linear Regression*.

We are given a dataset consisting of the medical costs for a group of 1339 Americans. For these people, we are given the age, sex, bmi (body mass index), number of childen, whether they smoke, the region they live in and the total amount of medical charges they have been charged in their life.

Let us start by focusing on linear regression with only one feature, we expect a person's age to be the most telling, therefore we will try to predict the a person's medical cost solely based on his age. Therefore age is the input feature and the medical cost is the target or dependent variable. We can visualize the relation between the two features using a scatterplot, see Figure 2.

Remark. From this scatterplot, we can immediately see that there clearly is some connection between age and medical cost, but that we still require some additional information in order to split the dataset further. For ease of notation we will nonetheless first focus on the case with only one input feature (age), but afterwards we will investigate how we can effectively use multiple features in order to more accurately predict the medical charges.

We have chosen for linear regression with only one feature, therefore our model looks like:

$$h(x) = \theta_0 + \theta_1 \cdot x, \tag{1}$$

that is a straight line that passes through the $y - axis$ at θ_0, we call θ_0 the intercept. It then linearly increases linearly with a slope equal to θ_1. We are now looking for a *optimal* values for θ_0 and θ_1 such that the line defined by $y = h(x)$ fits the distribution of points in Figure 2 as closely as possible.

To this end, we first need to quantify the error made by a prediction, this is done by defining a loss function $\ell(x)$. The most typical loss function used for linear regression is the Root Mean Squared Error (or RMSE in short), it is defined by:

$$\ell_\theta(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(h(x_i) - y_i)^2}{n}}. \tag{2}$$

Here, we denote by n the total number of samples, by x_i the age of sample i and by y_i the actual medical costs charged to the sample.

Remark. This loss function in (2) (referred to as the RMSE) can alternatively be written as:

$$\ell(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}, \tag{3}$$

with $\hat{y}_i = h(x_i)$, the prediction for sample i. Using this notation one can clearly see that this is a general formula which can be used to measure the performance of any model.

Remark. There is a strong resemblance between the RMSE and the standard deviation! If we would use the average of all y_i as prediction, we find that (1.6) is exactly the variance of the sample data $(y_i)_i$. This *mean estimate* is often used as a base case and any model that you use can be seen as improving this base case, that is: predicting (a part of) the variability in the distribution $(y_i)_i$.

In the loss function in we have x_i, y_i and n which are fixed given the dataset. We can however still determine the values of θ_0 and θ_1 in order to find a good fit. We can therefore visualize $\ell_\theta(x)$ as a function of θ_0 and θ_1. We

1.7 Gradient descent

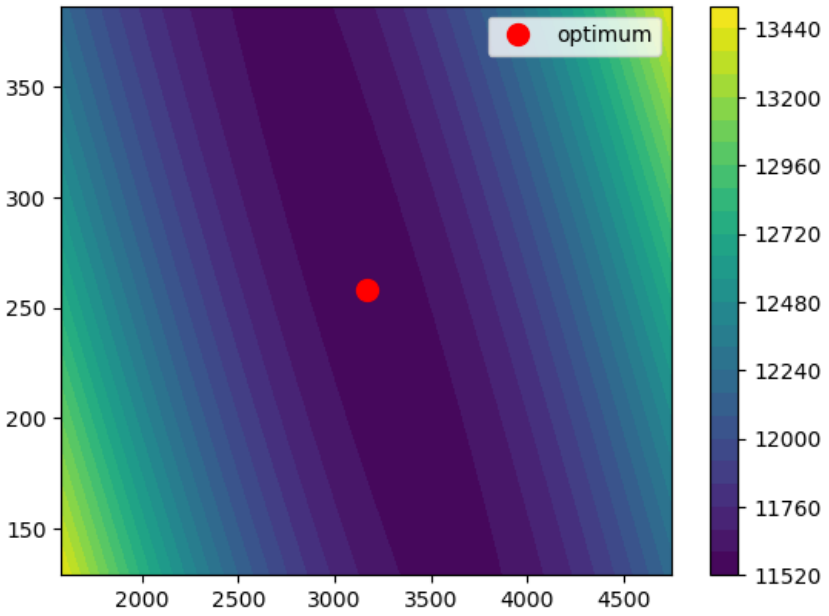

Figure 3: The loss function in (2) visualized for the medical charges example visualized.