RISK LATTE

Lecture 1

Data, Algorithms and Functions

Rahul Bhattacharya

Machine Learning: The New Science

Part I

What is this New Science of Machine Learning?

Putting Machine Learning in the context of Artificial Intelligence (AI)

💻 Introduction, Lecture 0

Part II

Foundational Concepts of Machine Learning

💻 Lectures 1 – 12.

Part III

Solving Real Life Problems with Machine Learning

💻 Lectures 13 – 25.

Part IV

Machine Learning: The Changing Landscape of Knowledge

(Near and the Not So Distant Future)

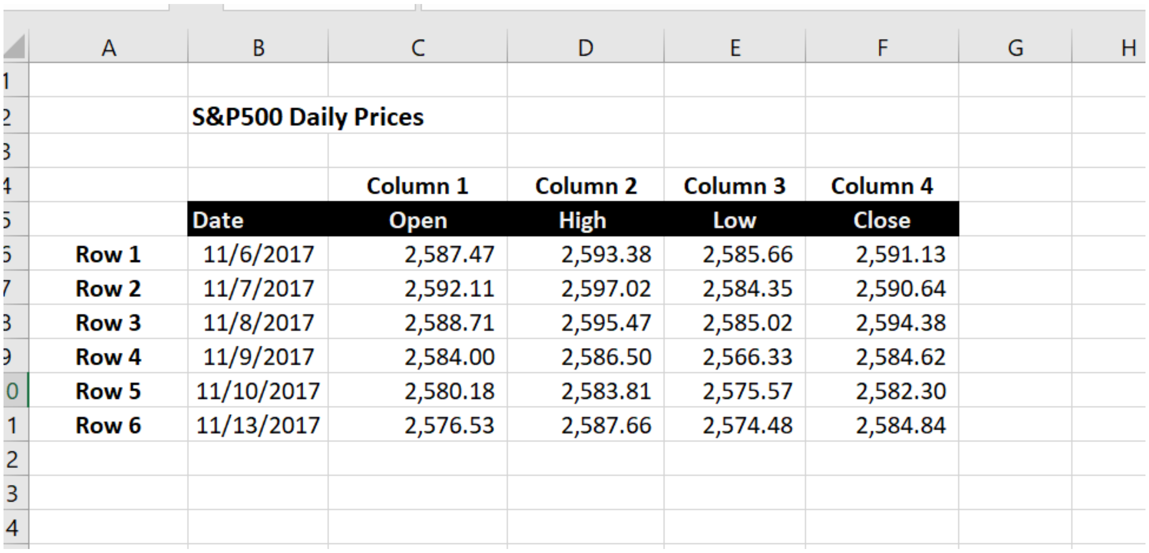
💻 Lectures 26, 27.

Algorithms and functions are key concepts in machine learning. However, before we can fully immerse ourselves into understanding what the new science of machine learning is and how it is changing the way we live and work, we need to get some preliminary, but essential, understanding about concepts like algorithms, functions and data (and, later on data structures).

Data is extremely important in machine learning and the entire field of machine learning is predicated on science of manipulating, dissecting and analyzing data – learning from the data, as it has come to be known – for the purpose of extracting a world view or knowledge – an output – that the data represents. And the way that learning happens is via a very special kind of algorithm, known as a **learning algorithm**. But before we talk about learning algorithms, or, the other kinds of algorithms that are generated from this learning algorithm, let's quickly take a look at how to represent data. To get started in machine learning, we need to properly understand and use the right terminology when talking about or representing data. In this lecture we'll take about data, functions and algorithms.

Visualizing Data through Spreadsheets

The best way to visualize data in machine learning is via an Excel™ spreadsheet. In a spreadsheet we have columns, rows, and cells as shown below in Figure 1.



S&P500 Daily Prices

		Column 1	Column 2	Column 3	Column 4
	Date	Open	High	Low	Close
Row 1	11/6/2017	2,587.47	2,593.38	2,585.66	2,591.13
Row 2	11/7/2017	2,592.11	2,597.02	2,584.35	2,590.64
Row 3	11/8/2017	2,588.71	2,595.47	2,585.02	2,594.38
Row 4	11/9/2017	2,584.00	2,586.50	2,566.33	2,584.62
Row 5	11/10/2017	2,580.18	2,583.81	2,575.57	2,582.30
Row 6	11/13/2017	2,576.53	2,587.66	2,574.48	2,584.84

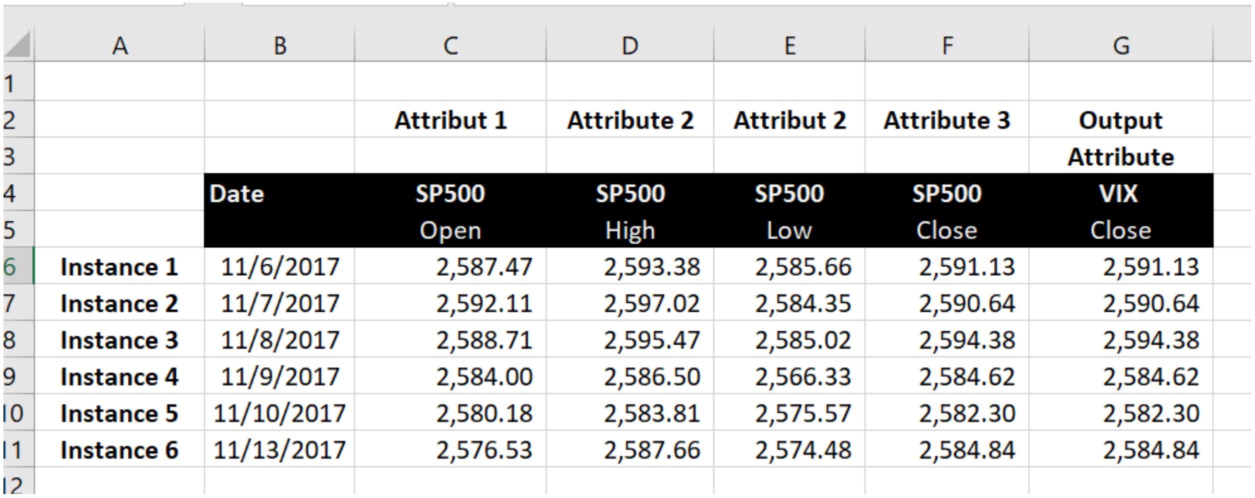
Figure 1.1: Visualizing Data in Machine Learning

- **Column (Attribute or Feature)**: A column in an Excel spreadsheet usually describes data of a single type. For example, you could have a column of daily or weekly stock prices. All the data in one column will have the same scale and dimension and will also have the same meaning relative to each other. In machine learning, a column represents the "attributes" of data. The term "attribute" is used in Computer Science and by programmers and an equivalent word in statistics is "properties". Another term for "attribute" used commonly by computer scientists and programmers is "Feature". A feature describes a property of an observation.

- **Row (Instance)**: A row describes a single observation and the columns describe the attributes about that observation. For example, if we are talking about daily stock prices then each row will signify the observation made about the stock price on each trading day and the column will represent the actual stock price on that day. The larger the number of rows the more examples from the problem domain we have. In computer science the row – or, the observation – is termed as an “Instance”. A row signifies a single instance of the observed data.

- **Cell**: A cell is an intersection between a row and a column and shows a single value – numeric, alpha-numeric, integers, real or complex numbers, etc. – that is displayed (or, stored) in a particular row and a column.

This is how we should visualize data via columns, rows and cells of an Excel spreadsheet.



	A	B	C	D	E	F	G
1							
2			Attribut 1	Attribute 2	Attribut 2	Attribute 3	Output
3							Attribute
4		Date	SP500	SP500	SP500	SP500	VIX
5			Open	High	Low	Close	Close
6	Instance 1	11/6/2017	2,587.47	2,593.38	2,585.66	2,591.13	2,591.13
7	Instance 2	11/7/2017	2,592.11	2,597.02	2,584.35	2,590.64	2,590.64
8	Instance 3	11/8/2017	2,588.71	2,595.47	2,585.02	2,594.38	2,594.38
9	Instance 4	11/9/2017	2,584.00	2,586.50	2,566.33	2,584.62	2,584.62
10	Instance 5	11/10/2017	2,580.18	2,583.81	2,575.57	2,582.30	2,582.30
11	Instance 6	11/13/2017	2,576.53	2,587.66	2,574.48	2,584.84	2,584.84
12							

Figure 1.2: An Alternative Nomenclature for Data

Statistical Learning Framework

In a statistical learning framework, data is observed and understood within the context of a hypothetical function f that a particular learning algorithm is trying to learn. Input and Output can be understood in terms of that function, f as:

$$Output = f(Input)$$

In mathematical terms, we can say that a fixed set of inputs or a vector of inputs is being mapped – or, associated – with a single output.

$$Output = f(Vector\ of\ Inputs)$$

Using math notation, we can write the above equation as

$$y = f(\boldsymbol{x})$$

Where, we have bold faced x to denote that it is a vector. In the most general sense, goal of machine learning is to extract the nature of the function f from the data. Our knowledge of the universe, and everything contained in it and how they interact each other, is encapsulated in a function f (we can call it a general-purpose function). This function, f, is much more general and all-encompassing than what we have learnt in our high school or college mathematics. All human knowledge, all our skills, reside inside such a general-purpose function. Human progress can be aptly summarized as the story of how, over several millennia, we have been able to successfully receive precepts from our environment and perform the necessary actions that have resulted in progress and development in the fields of science, technology, philosophy, arts and everything else that we see around us. And, has been possible because we have been able to successfully map the percept sequences to our actions.

Many of you must have heard of artificial intelligence (we touched very briefly on this subject in our introductory lectures). AI is nothing by a study of how computers can, like human beings, successfully map the percepts into actions via a general-purpose function. Machine learning, which is a sub-field of AI, aids in that process.

Generally speaking, in machine learning, the function, f, can also be viewed – somewhat simplistically, though – as an algorithm. The output in the above equation represents the state of the world, as is embedded in and learnt from the data and the algorithm – that very special learning algorithm – is the process by which the computer has learnt that.

Another way to represent the above equation is to not think in terms of inputs and outputs or functions. If instead, as Jason Brownlee suggest, we think in terms of “Model” and “Data”, the two building blocks of all science then we can express the relationship as:

$$Model = Algorithm(Data)$$

Jason Brownlee considers “a model as the specific representation learned from data and the algorithm as the process for learning it”. Many a times, people confuse models with algorithms. But we need to be aware of the difference between the two. A model is a way to represent what we have learnt from the data – a particular world view – whereas an algorithm is process by which we have undergone that learning.

Now we are ready to ask the quintessential question in machine learning. How precisely does “learning” happen in machine learning? It happens via a leaning algorithm. Let’s get started with a more detailed understanding of what algorithms are in general and in the context of conventional computer science and how a learning algorithm is different from all other algorithms and, finally, how such a learning algorithm acts upon data to make the computer learn.

References:

- Abu-Mostafa, Yaser S., Malik Magdon-Ismail, Hsuan-Tien Lin, *Learning from Data*, *A Short Course*, AMLBooks.com, 2017
- Brownlee, Jason, *Machine Learning Algorithms*, Edition v1.13, 2018
- Alpaydin, Ethem, *Introduction to Machine Learning*, 3rd edition, MIT Press, 2014

Machine Learning: The New Science is an **online course** developed by Risk Latte AI to increase awareness and a very basic understanding of the field of Machine Learning, the new science that is taking our world by storm. In the next ten years, the world that we see around us would have changed completely, thanks to machine learning. In the next 20 to 30 years at most we, the human beings, would be in a totally uncharted territory, once again thanks to machine learning and the much bigger and related discipline of artificial intelligence (AI). In more ways than one, machine learning is the stepping stone to understanding artificial intelligence. Machine learning is a very important sub-field of artificial intelligence.

This online course prepares a person to start off with a more formal **Machine Learning 101** course. This course contains very easy to understand lectures and simple tutorials implemented in Microsoft Excel™ to illustrate real life problem solving using machine learning algorithms.

- This course is primarily targeted towards high school, college and university students and middle to senior level working professionals and executives around the world who have a very basic knowledge of mathematics and a very limited or even **no knowledge** of computer programming.
- For certain segments of our target audience, such as certain specific groups of high school, college and university students in India and other developing countries this online course costs only US$8.00 (USD Eight only).
- This course comes with on-demand and customized onsite, classroom lectures and workshops at a minimal cost.
- A lot of the lectures and many Excel spreadsheets containing worked examples are **FREE** and can be easily downloaded from our website. These convey the general flavor of the course.

Risk Latte AI is a unit of **Risk Latte Americas Inc.**, a Montreal, Canada incorporated company, that is in the business of developing machine learning algorithms and software for banks, financial institutions, healthcare and the education industry as well as developing gamification and social learning applications for the general public.