Getting started with Data Science and Machine Learning

39 Shares 23

4 min read

As you might have noticed Data Science, Big Data, Statistics, Machine Learning, Text Mining and Natural Language Processing are popular buzzwords at this point of time. This buzz is well justified and far more than hype. Search engines, Word Processors, recommendation engines. social media analytics, news aggregators, Kinect and self driving cars: all of them directly or indirectly depend on Data Science and/or Machine Learning.

Important Topics in Data Science:

Data Acquisition, Mining, Scraping and Parsing

This is the stage where one actually acquires data. This might involve crawling web pages, or simulating GET and POST requests to collect target data. Mechanize in Ruby or Python might be a reasonably simple starting point in this direction. They help you navigate to different web pages, enter data in forms, simulate clicks of Submit buttons, and collect the data. You might also need to understand how to parse webpages and HTML, for which you might need to understand something like XPath or something simple, friendly and intuitive likeBeautiful Soup

Text Processing and Regular Expressions

A lot of the time, in fact most of the time, the data you collect from the web, will be in a format different from what you'd like. You might need to extract something specific, like Phone Numbers. Or, you might need to identify address fragments on a business listing page. You will almost certainly need to understand parsing libraries for XML, JSON, CSV and HTML. One of the most important and efficient tools while working with data, is the Regular Expression. We have a track to get you started with regular expressions. The power of regular expressions is not to be underestimated. From cleaning and normalizing text data which you have collected, to categorizing text data based on popular words or patterns: this is often the foundation stage of several important data science experiments, specially those which involve data scraped from the web.

Machine Learning and Natural Language Processing

At this point, we move to the more Math intensive part of this exciting space, much of which is heavily dependent on Statistics, Probability and Linear Algebra. Machine Learning tries to learn from historical data, or to effectively cluster or segment data. Natural Language Processing is the science – or perhaps art – of trying to analyze, understand and generate language. There's a fair bit of overlap in the areas and Natural Language Processing uses quite a few of Machine Learning based techniques. Learning from data is powerful, and large volumes of data reveal a lot of information, which could effectively be used in solving or predicting unseen cases. If you'd like to appreciate the power of data, take a look at this simple 21 line Spelling Corrector written by Peter Norvig.
And while you're at it, try out some of the problems on our Machine Learning track. In the process, you'll learn how something as simple as a histogram can be used to build powerful real world features such as a T9 text prediction engine, or a spell checker. There are also some challenges such as the Quora Answer Classifier which might give you the experience of real world data. There's a variety of problems on this track: some which help you brush up text processing , statistics and regression; and others where you learn how to apply the fundamentals to real problems.

Online Resources for Getting Started with Data Science and Machine Learning

For someone trying to get started with ML, here is a resource where the complexity is *just* right. It introduces you to a lot of the essential Mathematics, but doesn't go too deep into it. It is an equivalent of the Applied ML course at Stanford.

Very briefly, here are the ML algorithms which are very useful and basic, and will help you solve a lot of problems.

- Regression- Single, Multiple Variables, Logistic Regression
- Overfitting and Underfitting issues- 'Bias' and 'Variance'
- Simple clustering algorithms- K-Means
- Applying basic linear algebra: Principal Component Analysis
- Recommendation Systems and Large Scale Systems.

Many people have gone on to become top Kaggle contestants (a popular data science contest portal) after doing this course. These introductory algorithms can be extremely useful.

Apart from this I'd also recommend learning a bit about text processing such as regular expressions, string functions and language models. You might find them in the first few lectures and tutorials onthis Natural Language Processing Course.

I'd like to emphasize that

- A lot of the Mathematics involved doesn’t require much more than an introductory statistics code.
- If you go through our problem statements, many of them like Spell Check, are structured like a tutorial, and actually guide you through the steps required to get an interesting and powerful feature working. So it is possible to build-solve/as well as learn at the same time. For instance, all that Spell Check requires you to know, is what a Histogram is!
- For a quick and general introduction to Data Science, the course material from this Coursera course is great, and introduces R, Python, Map-Reduce and Data Visualization techniques.
- At a later stage, and for those looking for more academically challenging and abstract courses, you might be interested in theLearning from Data course from Caltech and the Probabilistic Graphical Models course from Stanford. The first course gives more theoretical insights into the foundations of Machine Learning and Statistical Leaning Theory; the second is about mixing Data Structures with Statistics to evolve Bayesian Networks and Hidden Markov Models: powerful tools, which are used in medical diagnostics, speech recognition engines, Kinect – and have been found to be significant improvements on traditional Machine Learning techniques.