Importance of Quality Data

There are two major ingredients to develop an ML solution:

1) Learning algorithms; and

2) Data related to the problem.

If used right, these two ingredients can lead one from an average ML solution to a great one. Although a very important ingredient, learning algorithms are fairly accessible. Most of them belong to the public domain in the form of open-source projects or peer-reviewed research articles published by academia. This accessibility may mean that small startups to individual developers have access to the same learning algorithms used by giant corporations and research labs.

However, this is not the case with our second ingredient: data. Data is the key ingredient out of the two as it represents and captures the knowledge of the world, required to solve the business problem. In other words, data helps the learning algorithm to learn about a world related to solving the problem at hand. This means that, if you have quality data which represents the problem you are trying to solve with a large enough number of samples representing various conditions, already you have a head start on developing a good ML solution. Most large multinational corporations are successful in developing strong ML systems and solutions because they retrieve, extract, collect and store millions or even billions of data points while cleaning and processing this data periodically. These quality datasets allow them to develop strong ML solutions that capture the knowledge and the context of the problem the system is trying to solve. So, although a pair of companies have access to the same learning algorithms, the power of data means that the company with better data should be able to come up with better machine learning systems most of the time.

Another aspect of this is that not all data are suitable to solve all problems. The data has to capture the knowledge and the context of the particular problem it is trying to solve. For instance, say Betty is an agriculture farm owner, if she wants to predict her crop-yield for this season she could possibly use historical yield and environmental condition data from her field. Let’s say that Betty is also interested in investing in agricultural companies and she collects stock market prices for the shares of the companies she is interested in. Even though she has collected years worth of such stock market data, this data can only be used to potentially create an ML solution to predict stock prices, but it will be highly irrelevant to predict her crop-yield. This is expected as the latter dataset does not capture any knowledge with respect to the former problem.

Hence, a dataset that is of quality, large enough, and relevant to the specific problem at hand is very much essential to developing a good ML solution.

▼ Related Courses/ Tutorials

Source **:** https://www.coursera.org/learn/machine-learning-applied/supplement/Wg8L1/data-is-central-to-your-ml-problem-required