Assignment 1

Gayatri Venugopal
gayatri.venugopal@sicsr.ac.in
6/8/2020

Need for Machine Learning

The chapter begins with the explanation of the need for machine learning. The traditional method of solving a problem through program is to identify the requirement, build the business logic and write the program accordingly. This is essentially a rule based system wherein the logic consists of a set of rules that are used to generate the outcome. If our rule based programs work without any issue in all scenarios, with a variety of data, we do not need machine learning. 'Machine Learning' is used when the machine is required to learn from the data fed into it and adapt its program of making the prediction according to this data. For instance, I am currently working in a task of identifying complex (difficult to comprehend) words in a given Hindi sentence. In order for the machine to identify a particular word as complex, it requires a collection of examples of simple and complex words along with their charactersitics. The more the data in this case, the better approximation it can make about the complexity of a word. We cannot create a list of exhaustive rules for determining whether a word is complex or not owing to the presence of many factors. These independent variables may individually as well as in combination, have an effect on the dependent variable. Therefore in this case, we would need machine learning. This example is a hard problem because one rule will not work for all readers due to the subjectivity of the task.

Initially before the popularisation of deep models, features were manually engineered based on the domain knowledge. Today we have deep models that automatically learn the features from the dataset.

Key Concepts of a Machine Learning Problem

a. Dataset – a collection of examples containing features that are numeric representations of characteristics of a unit of the input. E.g. images are represented as matrices of pixels, text is represented by converted it to a vector. The challenges that we might encounter when dealing with data are:

i. varying length of text (vectors with non-uniform dimensions)

ii. lack of standardisation (e.g. non-uniformity in the case) of the text

iii. unrepresented group of people, i.e., the data does not represent what it should, which might lead to a biased dataset

b. Model – statistical models that can be estimated from data. Deep learning models apply transformations to the data using layers, in order to achieve the best performance.

c. Objective functions – Also known as loss functions or cost functions, they measure how good or bad a model is at a particular task. These are defined using model parameters. The objective here is to find the best parameters to minimise the loss. The training error and the test error help in learning the model parameters. Training error refers to the error on the dataset used for training and test error refers to the error on an unseen dataset. The objective is to make the model generalisable and to avoid overfitting, wherein it performs well on the training set but underperforms on the test set. This is a crucial component of machine learning.

d. Optimisation algorithms – they are used for fine tuning the models and modifying the hyperparameters. These algorithms are used to search for the best possible parameters that minimise the loss function. In deep learning, they are based on gradient descent – parameters are modified at each step and the loss is calculated. The parameter value is finally set to a value that minimises the loss.

Types of Machine Learning

a. Supervised Learning – the machine learns from labeled examples

b. Unsupervised Learning – clusters are created automatically from the data

c. Interacting with the Environment – creating intelligent agents that can perform an action rather than predicting an outcome alone
c. Reinforcement Learning – reward based method, that generates policies that map observations from the environment to actions

Machine Learning Roots and the Advent of Deep Learning
Machine learning has been around for centuries – the statistical foundations of algorithms that are prevelant today were used in ancient times as well. Today, with the growing size and variety of data, we have shifted to deep learning, which applies linear and non-linear transformations on the data along with backpropagation in order to learn the optimum parameters to produce an outcome that is close to the real world outcome. The field has expanded owing to the development in hardware as well as computational methods that make it cheaper and faster to use deep learning algorithms. AI, specifically machine learning and deep learning methods have become ubiquitous in our lives, making their presence felt in various fields such as the automotive industry, neuroscience, physics, education to name a few.

Exercises
1. Which parts of code that you are currently writing could be "learned", i.e., improved by learning and automatically determining design choices that are made in your code? Does your code include heuristic design choices?
I am currently working on supervised machine learning problem. The problem is to classify a given word as complex (difficult to comprehend) or simple. Since the complexity of a word is highly subjective and varies from person to person, it would be difficult to write rule-based code for this problem. Learning would take place if we could give the machine a large amount of data annotated by a diverse set of users.
2. Which problems that you encounter have many examples for how to solve them, yet no specific way to automate them? These may be prime candidates for using deep learning.
Identifying the author of a given content
Every author has a unique style of writing. Again, as in the previous example, we may not be able to write a specific set of hand-crafted rules to identify the author given any text. This issue might be solved to a certain extent by using deep learning techniques.

3. Viewing the development of artificial intelligence as a new industrial revolution, what is the relationship between algorithms and data? Is it similar to steam engines and coal (what is the fundamental difference)?
Yes, I believe that data does fuel algorithms. The more discoveries we make in a given dataset as we go along, the more diverse our selection of algorithm becomes. We may not choose one particular algorithm to work with the data in hand; many times we have to choose an ensemble of varied algorithms to reach a satisfactory outcome. When the size of the data increases, then algorithm has to also consider the resources available for its implementation. So algorithm is heavily dependent on data and its characteristics.

4. Where else can you apply the end-to-end training approach? Physics? Engineering? Econometrics?
The use of an end-to-end approach can be seen in the fields of education, disciplines related to the environment such as waste management, zoo, animal husbandry, surveillance systems, social media content classification and a variety of other fields. It would be hard to think of a field where end-to-end training cannot be used.