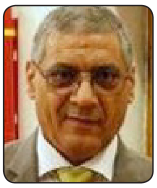
Journal of
Electrical Engineering and Electronic Technology

Abstract



Guide on Way to approach a Machine Learning problem

Mansi Priya
B.I.T. Sindri- Jharkhand, India

Abstract:

Today, algorithms are like buzz words. Everyone is going for learning different kinds of algorithms – logistic regression, random forests, decision tress, SVMs, Gradient boosting algorithms, neural networks etc.. Everyday new algorithms are being made. But Data Science is not just applying different algorithms to the data. Before applying any algorithm, you must understand your data because that will help you in improving performance of your algorithms later. For any problem one needs to iterate over the same steps- data preparation, model planning, model building and model evaluation, for improving accuracy. If we directly jump to model building, we end up directionless after one iteration. Following are few defined steps per me for approaching any machine learning problem:The first step I suggest is to understand your problem properly with a good understanding of the business market. There is no scenario like: here is the data, here is the algorithm and Bam! Proper business understanding will help you in handling the data in upcoming steps. For example, if you do not have any idea about the banking system you will not understand if a feature like income of customer, should be included or not. The next step is to collect relevant data for your problem. Other than the data you have internally in your company, you should also add external data source. For example, for sales prediction you should understand the market scenario for sales of your product. GDP may affect your sales or may be population affects. So, collect such kind of external data. Also remember the fact that any external data that you use should be available to you in the future when your model gets deployed. Like if you use population in your model, next year also you should be able to collect this data for getting predictions in the next year. I have seen many people who only use their internal data without realizing the importance of external data to their dataset. But in reality, external features have a good impact on our use case. Now when you have collected all the relevant data for your problem, you must divide it for training and testing. Many data scientists follow the 70/30 rule to divide the data into two parts: training and test set. While many follow the 60/20/20 rule to divide the data into three parts: training set, test set and validation set. I prefer the second option because in this case you use test set for improving your model and validation set for final verification of your model in actual scenario. with it. I was working on a default loan prediction problem. My accuracy was 78%. I took my problem to the person who was handling financial systems related to loans.

Biography:

I am a Data Scientist with a demonstrated history of working for Yamaha for two years. Currently I live in Japan and am actively working on different machine learning problems all over Yamaha. I am active user of LinkedIn, Kaggle and Analytics Vidhya. I have done Bachelors in Technology from BIT Sindri, Dhanbad with Computer Science branch. Some of my achievements are: I am winner of IGNITE Award all over India for Innovative ideas which is announced on the birthday of Dr. A.P.J. Abdul Kalam. I got felicitate by Dr. A.P.J. Abdul Kalam himself for the same. I have one patent till now for my prototype of User Friendly Pedal Rickshaw.   I have also been second runner up in Smart India Hackathon which is the world's largest hackathon organized by Government of India and MHRD. I have always been keen on learning on something new. This accounts for my numerous courses that I have done and tried to implement till now. Currently I am trying to learn Japanese to ease my life in Japan.

Recent Publications:

1. Cane D.E., Walsh C.T. The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. Chem. Biol. 1999; 6:R319–R325.

2. Nikolouli K., Mossialos D. Bioactive compounds synthesized by non-ribosomal peptide synthetases and type-I polyketide synthases discovered through genome-mining and metagenomics. Biotechnol. Lett. 2012; 34:1393–1403.

3. Winn M., Fyans J.K., Zhuo Y., Micklefield J. Recent advances in engineering nonribosomal peptide assembly lines. Nat. Prod. Rep. 2016; 33:317–347.

4. Cane D.E., Walsh C.T., Khosla C. Harnessing the biosynthetic code: combinations, permutations, and mutations. Science. 1998; 282:63–68.

5. Marsden A.F., Wilkinson B., Cortes J., Dunster N.J., Staunton J., Leadlay P.F. Engineering broader specificity into an antibiotic-producing polyketide synthase. Science. 1998; 279:199–202.

~~Webinar on Artificial Intelligence, Big Data Analytics & Iot; October 14, 2020; Dubai, UAE`~~

Citation: Mansi Priya; Guide on Way to approach a Machine Learning problem ; Big Data Analytics 2020; October 14, 2020; Dubai, UAE`