Assignment: Design and Application of a Machine Learning System for a Practical Problem

Set by: Prof. Luca Citi (lciti@essex.ac.uk)

Word Count: 720

Pilot-Study (Task 1)

- There are different types of problems in the field of Machine Learning. Sometimes based on the expected outcome, you can decide which kind of learning method (supervised/unsupervised) should be utilized to tackle the problem. According to the pilot-study proposal, the nature of the mentioned problem could be interpreted as a binary classification problem. [1] In detail, the proposal asked for a learning algorithm to answer the following question:
  Whether opening a hotel in a specific location is profitable or not? (without specifying the amount of profit and based on some relevant historical data).
  Obviously, this is a yes/no question model which is regarded as a binary classification problem that can be addressed by a supervised learning algorithm to make a classifier to distinguish between a profitable case and a non-profitable case. As a result, we should use a classification algorithm to build the mentioned classifier and predict the desirable output.

- In my perspective, this question can be interpreted in two levels: the conceptual level which discusses what are the desired attributes for this problem, and the implemental level which indicates the proper formatting of the attributes. In this case, the location of the hotel, the number of rivals in the location, considering occupancy rate of the area, and local demand as well. [3] Besides, the distance between this hotel to the closest bus/tube station,hospital, store, and a mall. In addition, assessing how much an area has potential in terms of tourist attraction could be a helpful attribute. On the other hand, When it comes to implementation, we need to make sure the data format we use is compatible with the algorithms we deploy in some cases. In this example, locations can be described as latitude and longitude and it is possible to use miles as the unit to measure distance. In conclusion, there must be other contributing factors in the prosperity of a hotel at a certain location. However, I realized that those attributes could be pivotal to the success of a hotel.

- To answer this question, having a roadmap comes in handy. Suppose we have a problem and a dataset. By answering the following questions, we have a better picture of a suitable algorithm for our problem. [2]
  1. What kind of machine learning problem are we dealing with? (Supervised/Unsupervised/Reinforcement Learning)
  2. In the case of a Supervised learning task, Do we need to predict a real-valued attribute? or Do we need to specify the class of an observation? (Regression/Classification)

3. How large is the dataset and how sophisticated are the relationships between the attributes in the dataset? (For instance, is the problem linear separable in nature?)
4. What kind of balance do we seek in terms of run-time (time complexity), model accuracy, and simplicity in terms of model interpretation?

In this example, the first two questions have already been answered (eg. Binary Classification which is a supervised machine learning problem), Which shrinks down the number of options that we can have to solve a binary classification problem. We have different options such as Decision Tree, K-NN, Naive Bayes, Linear SVM, Logistic Regression, Random Forest for Classification, Non-Linear SVM, and even Neural Networks. By answering question 3, and considering the trade-off in question 4, we can reduce the number of introduced algorithms, and come up with a few algorithms to solve this problem. In short, in the case of the simpler dataset in terms of size and relationships between the attributes, using a simpler model like decision tree, K-NN, or Linear SVM could be a better choice while in a more complex dataset some sophisticated models like Random forest or Non-Linear SVM reflect better performance.

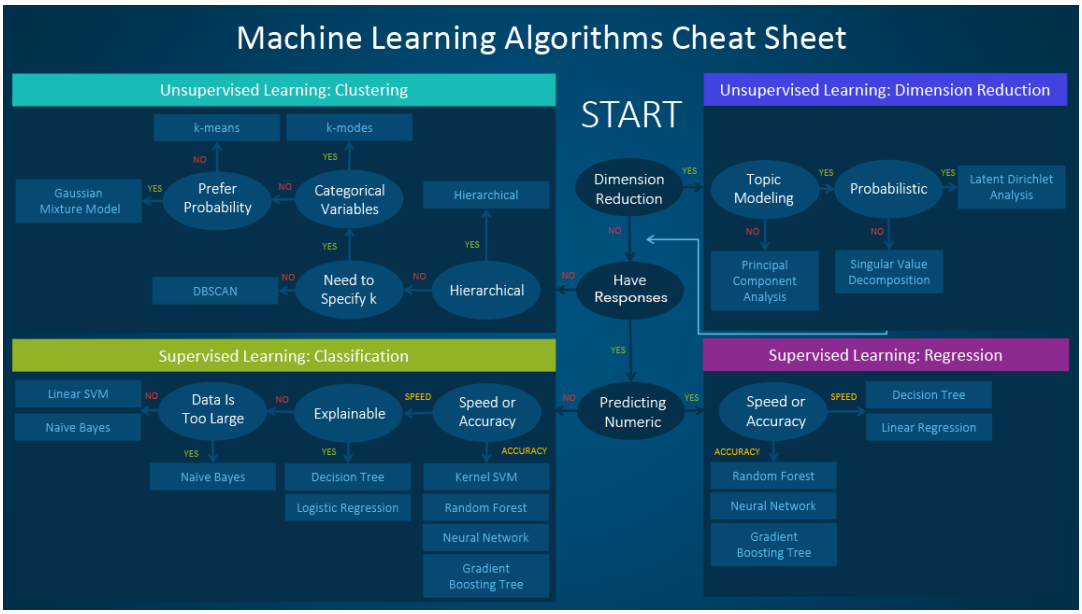

Fig1. Depicts a helpful chart to select the suitable algorithm for a machine learning problem.

- When it comes to evaluation of a model, cross validation methods play an important role. Cross validation could separate the data into two sets: train and test set. [4] Alongside Cross Validation, we can use some technique to train a model with different values for each of its parameters. [5] Besides, we can try different algorithms doing so and then test their performance on unseen data (test data) using cross validation methods. After that, we can choose the best model which shows better performance with test data. In summary, after completing the above steps we can have confidence that the selected model can provide an accepted performance on solving the problem.

References

[1]. Wikimedia Foundation. (2022, January 2). Machine learning. Wikipedia. Retrieved January 4, 2022, from https://en.wikipedia.org/wiki/Machine_learning

[2]. Li, H. (2020, December 9). Which machine learning algorithm should I use? The SAS Data Science Blog. Retrieved January 4, 2022, from https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

[3]. Forbes Biz Council Expert Panel. (2019, March 28). Council post: Investing in a hotel property? keep these eight factors in mind. Forbes. Retrieved January 4, 2022, from https://www.forbes.com/sites/forbesrealestatecouncil/2019/03/28/investing-in-a-hotel-property-keep-these-eight-factors-in-mind/?sh=771b07ff1f2b

[4]. Mitchell, T. M. (2017). Machine learning. Amazon. Retrieved January 4, 2022, from https://docs.aws.amazon.com/machine-learning/latest/dg/cross-validation.html

[5]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.