Questions from paper "A Few Useful Things to Know about Machine Learning"

Please try to answer the questions below in your own words as far as possible. If you don't understand a term or a concept, feel free to search it online. The idea of this assignment is to get an understanding of what is being done in the practice of machine learning today, not necessarily to master everything at this stage.

1. Introduction

1. What is the definition of ML?

Machine learning is study of computer algorithms that can learn to do specific task and improve using Experience and data.

2. What is a classifier?

A classifier is a system that takes a vector of discrete and/or continuous feature values and outputs a single discrete value.

2. Learning

1. What are the 3 components of a learning system, according to the author? Explain them briefly.

Note: Don't worry if you don't understand Table 1 fully yet. We will work on these throughout the semester.

2. Algorithm 1 presents a decision tree learner that determines whether to split a decision tree node and how to split it. It depends on information gain between attributes and the predicted value. Do a quick search on information gain and write down its definition and equation below.

2.1 •

To solve a machine learning problem we have wide variety of available algorithms. The task of finding the right learning algorithm consists of three components.

a. Representation

Algorithms must be represented in a formal language that a computer can understand. Representation is formal portrayal of a potential learner.

b. Evaluation.

Evaluation is needed to evaluate the performance of a learner.

c. Optimization

To find the best performing algorithm we need optimization to get the most efficient algorithm possible.

2.2.

Information gain is the reduction in entropy or surprise by transforming a dataset. Entropy is uncertainty in a group of observation.

Information Gain = entropy(parent) – [average entropy(children)]

3. Generalization

1. Why is generalization more important than just getting a good result on training data i.e. the data that was used to train the classifier?

Good performance on training Example doesn't prove that classifier will perform well on unknown test Examples. Hence, generalization is important for us to be certain about performance of classifier.

2. What is cross-validation? What are its advantages?

Cross validation is the process of dividing dataset in k parts treating (k-1) set as training and remain data for testing. Doing this for all individual k datasets and averaging the performance. This technique is used for making the most out of available data when available data is sparse.

3. How is generalization different from other optimization problems?

In generalization we don't have access to the function we want to optimize.

4. Data alone is not enough

1. Try to understand how a function involving 100 Boolean variables would lead to a total 2^{100} different possible examples (no need to write anything down, just try to understand).
If you have a scenario where the function involves 10 Boolean variables, how many possible examples (called instance space) can there be? If you see 100 examples, what percentage of the instance space have you seen?

There are 2^10 Examples in instance space. That would be 9.76% of total examples.

2. What is the "no free lunch" theorem in machine learning? You can do a Google search if the paper isn't clear enough.

The No free lunch theorem says that all optimization algorithm perform Equally well over a large set of problems. i.e. there is no single best algorithm that performs well for all problems.

3. What general assumptions allow us to carry out the machine learning process? What is the meaning of induction?

The general assumption is similarity among examples seen by the model to guess the possible output using induction. Induction is coming to conclusion based on past experiences.

4. How is learning like farming? ☺

we can draw similarity between farming and learning by thinking of seeds as data and techniques like watering the crops, space between sown seeds, what type of Fertilizer to use etc as knowledge. using which we can grow crops or in our case develop programs.

5. Overfitting

1. What is overfitting? How does it lead to a wrong idea that you have done a really good job on training dataset?

Overfitting is a result of a learner trained on small dataset which leads to good performance on training data but very poor performance on testing data. Since an overfitted model performs well on training data it might give a false impression of being a good learner, which it is not.

2. What is meant by bias and variance? You don't have to be really precise in defining them, just get the idea.

Bias means tendency to give similarly incorrect output for every input.
Variance means giving random output for any given input.

3. What are some of the things that can help combat overfitting?

we can use techniques like cross validation to improve model's accuracy by making the most of available data. Also using statistical tests like Chi-square test, f- test can help us understand the significance level of Each feature before learning. we can also use regularization that discourage learning of models that are highly complex and cause overfitting.

6. Intuition fails in high dimensions

1. Why do algorithms that work well in lower dimensions fail at higher dimensions? Think about the number of instances possible in higher dimensions and the cost of similarity calculation

Algorithms fail at higher dimensions because the features, if irrelevant add lot of noise cancelling out learning from any relevant features. Else features being relevant which can causes all Examples to look alike. Also as Dimension increase the instance space also increase exponentially.

2. What is meant by "blessing of non-uniformity"?

Blessing of non uniformity means that for instances of higher dimension, in most applications, examples tend to cluster near lower dimensions. So, we can use this in our advantage with dimensionality reduction.

7. Theoretical guarantees

* This section is a bit involved, so just read the first paragraph *

1. What has been one of the major developments in the recent decades about results of induction?

we can have probabilistic guarantees on the result of induction.

8. Feature engineering

1. What is the most important factor that determines whether a machine learning project succeeds?

Most important factor that determine whether machine learning project succeeds are the selected features.

2. In a ML project, which is more time consuming – feature engineering or the actual learning process? Explain how ML is an iterative process?

Feature engineering is more time consuming than actual learning., because raw data need to be processed before being useful for learning.
ML is an iterative process because we need to process the learner, analyze the results, and update data/learner then repeat the process again.

3. What, according to the author, is one of the holy grails of ML?

Automating the process of feature engineering is one of the holy grails of machine learning.

9. More data beats a cleverer algorithm

1. If your ML solution is not performing well, what are two things that you can do? Which one is a better option?

We can either opt for a better algorithm or get more data. choosing more data is better choice because simple algorithms with more data perform similar to clever Algorithms.

2. What are the 3 limited resources in ML computations? What is the bottleneck today? What is one of the solutions?

The three limitation are time, memory and training data. The bottleneck today is time. we can use simple algorithms that take less training time to overcome this.

3. A surprising fact mentioned by the author is that all representations (types of learners) essentially "all do the same". Can you explain? Which learners should you try first?

Author meant that all representations try to group nearest neighbours into same class only the way they do it is different. we should try simple learners to begin with and then move to more complex learner.

4. The author divides learners into two types based on their representation size. Write a brief summary.

There are two type of learners based on their representation:
- The fixed size learners with fixed amount of parameter that don't change based on amount of data available for.eg linear classifier.

- The non parametric learner which changes their representation size depending on amount of available data, this can also lead to increase in computation time eg. Decision tree.

10. Learn many models, not just one

1. Is it better to have variation of a single model or a combination of different models, known as ensemble or stacking? Explain briefly.

It is better to have stacking because instead of a single variation, we can use stacking to combine multiple variation to find best combination for our problem which would result in much better learner.

11. Simplicity does not imply accuracy

1. Read the last paragraph and explain why it makes sense to prefer simpler algorithms and hypotheses.

Preference to simpler algorithms should be given just because they are simple and not any other factors like accuracy .

12. Representable does not imply learnable

** Get an overview, no questions from this section **

13. Correlation does not imply causation

1. It has been established that correlation between independent variables and predicted variables does not imply causation, still correlation is used by many researchers. Explain briefly the reason.

Correlation can be thought of as a potential cause of action, which we can use as a guide for further investigation. And since, the goal of predictive models is to use them as a guide, many researchers still use correlation as causation.