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?

2. What is a classifier?

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.

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?

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

3. How is generalization different from other optimization problems?

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?

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

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

4. How is learning like farming? ☺

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?

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

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

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

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?

8. Feature engineering

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

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

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

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?

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

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?

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

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.

11. Simplicity does not imply accuracy

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

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.