Nam Nguyen – npn190000

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 systems automatically learn programs from data

2. What is a classifier?

➔ A classifier is a system that inputs (typically) a vector of discrete and/or continuous feature values and outputs a single discrete value, the class.

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.

➔ Learning = Representation + Evaluation + Optimization

➔ Representation. A classifier must be represented in some formal language that the computer can handle. Conversely, choosing a representation for a learner is tantamount to

choosing the set of classifiers that it can possibly learn. This set is called the hypothesis space of the learner

- ➔ Evaluation. An evaluation function (also called objective function or scoring function) is needed to distinguish good classifiers from bad ones. The evaluation function used internally by the algorithm may differ from the external one that we want the classifier to optimize.
- ➔ Optimization. Finally, we need a method to search among the classifiers in the language for the highest-scoring one. The choice of optimization technique is key to the efficiency of the learner, and also helps determine the classifier produced if the evaluation function has more than one optimum.

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.

- ➔ Algorithm 1 (above) shows a bare-bones decision tree learner for Boolean domains, using information gain and greedy search. InfoGain(xj, y) is the mutual information between feature xj and the class y. MakeNode(x,c0,c1) returns a node that tests feature x and has c0 as the child for x = 0 and c1 as the child for x = 1.
- ➔ Decision trees test one feature at each internal node, with one branch for each feature value and have class predictions at the leaves.

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?

- ➔ Because the fundamental goal of machine learning is to generalize beyond the examples in the training set. This is because, no matter how much data we have, it is very unlikely that we will see those exact examples again at test time.

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

- ➔ Cross-validation, called rotation estimation or out-of-sample testing, is a similar model validation technique for assessing how a statistical analysis results will generalize to an independent data set. Cross-validation is a resampling method that uses different portions of the data to test and train a model on various iterations.
- ➔ Cross-validation can help combat overfitting, for example, by using it to choose the best decision tree size to learn.

3. How is generalization different from other optimization problems?

➔ Notice that generalization being the goal has an interesting consequence for machine learning. Unlike in most other optimization problems, we do not have access to the function we want to optimize! We have to use training error as a surrogate for test error, and this is fraught with danger.

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?

➔ 100 Boolean variables would lead to a total 2^{100} different possible examples

➔ 10 Boolean variables would lead to a total 2^{10} different possible examples

➔ if there are 100 Boolean features and the hypothesis space is decision trees with up to 10 levels, to guarantee $\delta = \varepsilon = 1\%$ in the bound above we need half a million examples. But in practice a small fraction of this suffices for accurate learning

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

➔ "no free lunch" theorems are presented which establish that for any algorithm, any elevated performance over one class of problems is offset by performance over another class. These theorems result in a geometric interpretation of what it means for an algorithm to be well suited to an optimization problem.

➔ Occam's razor famously states that entities should not be multiplied beyond necessity. In machine learning, this is often taken to mean that, given two classifiers with the same training error, the simpler of the two will likely have the lowest test error. Purported proofs of this claim appear regularly in the literature, but in fact, there are many counterexamples to it, and the "no free lunch" theorems imply it cannot be true

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

- ➔ Machine learning is not a one-shot process of building a dataset and running a learner, but rather an iterative process of running the learner, analyzing the results, modifying the data and/or the learner, and repeating.
- ➔ Induction (what learners do) is a knowledge lever: it turns a small amount of input knowledge into a large amount of output knowledge.

4. How is learning like farming? ☺

- ➔ Learning is more like farming, which lets nature do most of the work. Farmers combine seeds with nutrients to grow crops. Learners combine knowledge with data to grow 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?

- ➔ What if the knowledge and data we have are not sufficient to completely determine the correct classifier? Then we run the risk of just hallucinating a classifier (or parts of it) that is not grounded in reality and is simply encoding random quirks in the data. This problem is called overfitting,
- ➔ When your learner outputs a classifier that is 100% accurate on the training data but only 50% accurate on test data, when in fact it could have output one that is 75% accurate on both, it has overfit.

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

- ➔ Bias is a learner's tendency to consistently learn the same wrong thing.
- ➔ Variance is the tendency to learn random things irrespective of the real signal.

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

- ➔ Cross-validation, adding a regularization term

6. Intuition fails in high dimensions

Nam Nguyen – npn190000

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

➔ In most applications, examples are not spread uniformly throughout the instance space, but are concentrated on or near a lower dimensional manifold. For example, k-nearest neighbor works quite well for handwritten digit recognition even though images of digits have one dimension per pixel, because the space of digit images is much smaller than the space of all possible images. Learners can implicitly take advantage of this lower effective dimension, or algorithms for explicitly reducing the dimensionality can be used.

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

➔ Fortunately, there is an effect that partly counteracts the curse, which might be called the "blessing of nonuniformity."

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?

➔ One of the major developments of recent decades has been the realization that in fact, we can have guarantees on the results of induction, particularly if we are willing to settle for probabilistic guarantees.

8 Feature engineering

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

➔ Like any discipline, machine learning has a lot of "folk wisdom" that can be difficult to come by, but is crucial for success
➔ Machine learning researchers are mainly concerned with the former, but pragmatically the quickest path to success is often to just get more data

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

- ➔ Running a learner with a very large number of features to find out which ones are useful in combination may be too time-consuming, or cause overfitting. So there is ultimately no replacement for the smarts you put into feature engineering
- ➔ Machine learning is not a one-shot process of building a dataset and running a learner, but rather an iterative process of running the learning analyzing the results, modifying the data and/or the learner, and repeating

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

- ➔ One of the holy grails of machine learning is to automate more and more of the feature engineering process.

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?

- ➔ Design a better learning algorithm, or gather more data (more examples, and possibly more raw features, subject to the curse of dimensionality).
- ➔ Machine learning researchers are mainly concerned with the former, but pragmatic the quickest path to success is often to just get more data

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

- ➔ The two main limited resources are time and memory. In machine learning, there is a third one: training data.
- ➔ Today it is often time.

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?

- ➔ Part of the reason using cleverer algorithms has a smaller payoff than you might expect is that, to a first approximation, they all do the same.
- ➔ This is surprising when you consider representations as different as, say, sets of rules and neural networks. But in fact, propositional rules are readily encoded as neural networks, and similar relationships hold between other representations.

➔ All learners essentially work by grouping nearby examples into the same class; the key difference is in the meaning of "nearby." With nonuniformly distributed data, learners can produce widely different frontiers while still making the same predictions in the regions that matter (those with a substantial number of training examples, and therefore also where most test examples are likely to appear)

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

➔ Learners can be divided into two major types: those whose representation has a fixed size, like linear classifiers, and those whose representation can grow with the data, like decision trees.

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.

➔ As the competition progressed, teams found they obtained the best results by combining their learners with other teams' and merged into larger and larger teams. The winner and runner up were both stacked ensembles of over 100 learners and combining the two ensembles further improved the results.

11. Simplicity does not imply accuracy

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

➔ Because simplicity is a virtue in its own right, not because of a hypothetical connection with accuracy.

12. Representable does not imply learnable

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

Nam Nguyen – npn190000

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.

- ➔ the goal of learning predictive models is to use them as guides to action. If we find that beer and diapers are often bought together at the supermarket, then perhaps putting beer next to the diaper section will increase sales.
- ➔ Machine learning is usually applied to observational data, where the predictive variables are not under the control of the learner, as opposed to experimental data, where they are.
- ➔ Some learning algorithms can potentially extract causal information from observational data, but their applicability is rather restricted.
- ➔ On the other hand, correlation is a sign of a potential causal connection, and we can use it as a guide to further investigation.
- ➔ Many researchers believe that causality is only a convenient fiction. For example, there is no notion of causality in physical laws. Whether or not causality really exists is a deep philosophical question with no definitive answer in sight, but there are two practical points for machine learners. First, whether we call them “causal,” we would like to predict the effects of our actions, not just correlations between observable variables. Second, if you can obtain experimental data then by all means do so