**Questions from paper "A Few Useful Things**

**to Know about Machine Learning"**

**1. Introduction**

**1. What is the definition of ML?**

Machine Learning is automatically learning programs from data.

**2. What is a classifier?**

A classifier is a system that takes a vector of continuous or discrete features as an input, and outputs a single discrete value, which is the class

**2. Learning**

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

**briefly.**

1. Representation
   1. Classifier must be represented in a computer-understandable language
2. Evaluation
   1. Function that distinguishes good classifiers from bad classifiers
3. Optimization
   1. A method to search for the highest-scoring classifier among multiple

**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.**

Information gain is the measure of how much information a feature gives us about the target variable (how much entropy about the target class is reduced).

Its equation is: A black text on a white background

Description automatically generated

**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 training data likely won’t be seen again.

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

Randomly dividing training data into subsets. Its advantages are that there will not be overfitting, and the learner will have new data each time.

**3. How is generalization different from other optimization problems?**

Generalization is different from other optimization problems because we do NOT have access to the function that 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 2100**

**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 can be 210 = 1,024 possible examples. If you se 100, you have seen 9.77%.

**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 in ML is that there is no one-size-fits-all algorithm that is universally best for all possible problems

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

**meaning of induction?**

Assumptions such as smoothness, similar examples having similar classes, limited dependences, and limited complexity.

Induction turns a small amount of input knowledge into a large amount of output knowledge

**4. How is learning like farming? J**

In learning, we combine knowledge with data to grow programs, similar to how farmers combine seeds with nutrients to grow crops

**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 when a classifier is determined based on random quirks of data and is not grounded in reality. It can have 100% accuracy on a training dataset, and show only 50% on data that has not been seen, you think that your model is 100% accurate on all data.

**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 learning the same wrong thing over and over again (same wrong answer each time). Variance is learning random things regardless of the correct signal (different wrong answers each time)

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

Cross validation, adding a regularization term to the evaluation function, and statistical significance tests

**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**

Because the more features there are, the more data there is, and the training data’s fractional representation of the overall data becomes smaller and smaller.

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

Most examples are concentrated around a lower-dimensional manifold, so the lower dimension can be used to the advantage of the learner

**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?**

That we can have guarantees on the results of induction 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?**

The features used

**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

ML is an iterative process because you run the learner, analyze the results, and then modify the learner/data and repeat until you get the desired outcome

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

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?**

1. Design a better algorithm

2. Gather more data – this is the better option

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

The 3 limited resources are time, memory, and training data.

Today the bottleneck is data

One of the solutions is to come up with fast ways to learn complex classifiers

**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?**

All learners essentially work by grouping nearby examples into the same class

You should try simple learners first

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

**summary.**

Fixed representation size: can only take advantage of so much data

Representation size grows with the data: can learn any function given sufficient enough data, but no amount of data may be enough due to limitations of the algorithm

**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 a combination of different models because it greatly reduces variance while slightly increasing bias.

**11. Simplicity does not imply accuracy**

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

**hypotheses.**

Because simpler algorithms and hypotheses follow the virtue of simplicity.

**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.**

This is because the correlation between independent and dependent variables can indicate a POTENTIAL causal connection, and researchers use this as a guide for further investigation.