A few useful things to know about machine learning

November 9, 2023

1 Summary

0. This article summarizes 12 key lessons that machine learning researchers and practitioners have learned. These include pitfalls to avoid, important issues to focus on, and answers to common questions.

1. Learning = Representation + Evaluation + Optimization.
Representation is formal representation of the model that specifies a functional form of what is to be learnt. This is the *hypothesis space*.
Evaluation is the scoring metric that helps distinguish bad learners from good ones.
Optimization is the method to search for specific function that optimizes the evaluation metric.
The article lists options for each of these components to build a classification learner.

2. It is generalization that counts. Generalization of results beyong training set is a fundamental goal in machine learning. Since we do not know what function to optimize (as we don't have all the data - input space), we use training set errors as a proxy for test set errors.

3. Data alone is not enough. Since real world data is not drawn uniformly from the input data we can make some assumptions about the model (Representation above). The assumptions are based on knowledge beyond training data. Machine learning differs from traditional programming in that programs are not built from scratch but knowledge and data are combined to form learners.

4. Overfitting has many faces. Encoding random quirks in the data is overfitting. This can be decomposed into two parts - high bias is when learner learns wrong thing consistently and high variance is when learner learns random things despite the real signal. Achieving one of these is easy but not both. Cross-validation, significance tests, and regularization can help.

5. Intuition fails in higher dimensions. Input space increases exponentially with new features. Also algorithms which work in lower dimensions may not work in higher

dimensions. However, the examples are usually concentrated in a lower dimensional manifold,that is the positive side of it.

6. Theoretical guarantees are not what they seem. One type of theoretical guarantee is that probability of a classifier or algorithm being 'bad' is less than δ if number of examples is greater than some number. It is important to carefuly interpret what a theoretical guarantee says exactly. The main role of theoretical guarantees in machine learning is not as a criterion for practical decisions, but as a source of understanding and driving force for learner design.

7. Feature engineering is the key. Feature engineering is the most important thing that differentiates succesful machine learning projects from that are not. Machine learning is an iterative process, and feature engineering is the most difficult process in this as sometimes combination of features can be important.

8. More data beats a cleverer algorithm. Fixed size learners cannot take advantage of more data where as variable size learners like Decision Tree can do so. More data has other issues like computation and memory usage.

9. Learn many models, not just one. Practically, it seems learning with many models works better than learning with just the best model. Bagging, boosting and stacking are some of the techniques used to combine methods.

10. Simplicity does not imply accuracy. Fewer parameters or small hypothesis space does not necessarily imply more accurate learner. It can be a goal in itself though.

11. Representable does not imply learnable. A decision tree cannot have more leaves than training examples. Similar examples exist in other learners which show that a representable function is not necessarily learnable.

12. Correlation does not imply causation. Correlation is only a sign of potential causal link.

2 Criticism

None, as this is only a review article.