Summary of One Machine learning tutorial - *A few useful things to know about machine learning*

Presented by : Ji Gao

[1]Department of Computer Science, University of Virginia
https://qdata.github.io/deep2Read/

August 26, 2018

Categorical of machine learning data List of relations

features / attribute / variable	
Labels	Supervised/unsupervised/semi-supervised
Sample relationship	i.i.d or spatial dependency, temporal dependency
Approach	numerical function(Linear vs. Non-linear), rule based, ensemble models
Goal	Explanatory vs. Prediction

Table: Categorization of Machine Learning

Outline

1. A few useful things to know about machine learning

A Few Useful Things to Know about Machine Learning

Abstract: Machine learning algorithms can figure out how to perform important tasks by generalizing from examples. This is often feasible and cost-effective where manual programming is not. As more data becomes available, more ambitious problems can be tackled. As a result, machine learning is widely used in computer science and other fields. However, developing successful machine learning applications requires a substantial amount of black art that is hard to find in textbooks. This article summarizes twelve 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.

Machine Learning

Machine Learing: $F : X \to Y$
Component of machine learning:

- Representation: Need to build a hypothesis space
- Evaluation: Score function
- Optimization

Each have lots of different choices. All very important.

Generalization

Generalization is important!
Input space might be super huge. Number of training data provided is small. How to get answers for all other data points?
Generalization is the only way.

Assumptions

Impossible to learn things if no assumptions. No better than random guess. Assumptions often implied:

- Smoothness
- Similar examples lead to similar result (Continuity)
- Limited dependences
- Limited complexities

Overfitting

Overfitting: Rely too much on the training data, often because the training set is too small.
Consequence: Get bad performance on the test set.
Previous way to avoid overfitting: Cross-validation.
Related to our research: Better way?

Dimentionality curse

High dimension is hard to deal with in machine learning: Generalization becomes exponentially harder in higher dimensional space.
Also hard for testing.
However, generally the informational part of the input space is much smaller than the original space.

Theoretical guarantees are hard

As the input space is too large comparing to actual number of training data, it is hard to achieve a theoretical guarantee on the trained result. Direct result will be super loose, that is useless.
PAC-learning: probably approximately correct, with high probability (the "probably" part), the selected function will have low generalization error (the "approximately correct" part).

Feature engineering

Input features decide how well you can learn. The learner only have the information contained in the features.

More data is important

Complex model doesn't necessarily learn better result.
However, more data leads to achieve better generalization.

Multiple models better than one

- Empirically, different models using ensemble technique achieve better result.
- Meaningful considering software testing.
- Test ensemble models?

Simplicity $\neq$ Accuracy

- Less number of parameters doesn't means better generalization: It lead to a smaller hypothesis space, which might be more biased than complex models.
- Also, only a limited number of hypothesis are touched in the training process. The procedure of choosing hypothesis matters!

Representable $\neq$ Learnable

It is often the case that the machine learning models can represent a lot of different functions. However, not all functions are learnable:
Give finite time, data and memory, the machine learning model can only learn a small subset of functions.

Correlation $\neq$ Causality

- Causality is important
- Correlation is only a sign of causality. Unfortunately, machine learning model can only get correlation in many cases.