Homework 1

Name: Sennan Liu

NetID: sxl200012

Course #: CS6375.502

Introduction

- A program automatically learn algorithm from data.
- A system that receives a vector of input values and output a class label

Learning

- Represetation, Evaluation and Optimization.
- Representation refers to the fact that machine learning momdels must be represented into formal language such that computer can handle.
- Evaluation refers to the objective function that must be able to distinguish good models from bad ones.
- Optimization refers to the method for searching among all candidate models for the highest score one.
- Information gain calculates the reduction in entropy or surprise from transforming a dataset in some way. let P(i) be the probability of each class i
    - H(S) = -\Sigma_{i \in S}(p(i) * log(P(i))
    - H(S | a) = \Sigma{v \in a}(Sa(v)/S * H(Sa(v)))
    - IG(S, a) = H(S) – H(S | a)

Generalization

- The fundamental goal of machine learning is to generalize beyond the examples in the training set. And at practice, it is nearly 100% for the real examples to be different from data in the training set.
- Randomly deviding the training data into several subsets, holding out each one while training on the rest, testing each learned model on the examples it does not see, and average the testing result as a metric to do model evaluation.
- We have no direct accessibility with the real function that is used in practice

Data alone is not enough

- 1024 possible examples can be there, for 100 examples, the percentage would be roughly 10%
- The "no free lunch" theroem in machine learning refers to the fact that no learner can beat random guessing over all possible functions to be learned.

- Smoothness of samples, similar examples having similar classes, limited dependences, or limited complexity of ground truth. Induction is a knowledge that turns a small amount of input knowledge into a large amount of output knowledge.
- Learning is more like farming in a way that it lets data do most of the work. Farmers combine seeds with nutrients to grow crops. Learners combine knowledge with data to grow programs.

Overfitting

- Hallucinating a model that is not grounded in reality, because it is simply encoding random quirks in the data.
- 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.
- Cross-validation is one way, adding a regularization term is another. Performing a statistical significance test can also be helpful.

Intuition fails in high dimensions

- Because a fixed-size training set covers a dwindling fraction of the input space. Also, in high dimension space the information provided by real signal attributes is easily swamped by noices from other attributes, which spoiled the model induction based on similarity computation.
- In most applications examples are not spread uniformly throughout the instance space, but are concentrated on or near a lowerdimensional manifold.

Theoretical guarantees

- 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.

Feature engineering

- The features used
- Feature engineering is much more time-consuming. ML is an iterative process of running the learner, analyzing the results, modifying the data and/or the learner, and repeating.
- To automate more and more of the feature engineering process

More data beats a cleverer algorithm

- The two options are designing a better learning algorithm and gathering more data. Pragmatically, the quickest way path to success is often to get more data
- The three constraints are time, memory and training data. Today the bottle neck is the time. One way to solve this problem is to come up with fast ways to learn complex classifiers.
- To a first approximation, all learners work in a way to use nearest neighbour of existing data. The difference is simply by how to define nearest. As a result, I would try simplest learners first.

- The two types are those whose representation has a fixed size, like linear classifiers, and those whose representation can grow with the data, like decision trees. The first type can only take advantage of so much data. On the other hand, the second type may take advantege of more features although they often don't pragmatically.

Learn many models, not just one

- It's better to have a combination of different models. The reason was that different learners would cover different aspect of features.

Simplicity does not imply accuracy

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

Correlation does not imply causation

- Data for machine learning is usually observable not experimental, which makes it impossible to do strict learning based on causality. On the other hand, correlation is a good start point to find real causality.