# Week1 Report

Chengze Cai

caicz18@mails.tsinghua.edu.cn

## 1 Introduction

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

## 2 Basis

### 2.1 Tasks

Given a sample distribution (x,y)$\sim \mathcal{D}$, $x \in X$, $y \in Y$, we want to find a function $f : X \to Y$ in the hypothesis space $\mathcal{H}$, and minimize the expected difference between f(x) and y.

In a classification problem, Y is a finite set, and we need to minimize
$error(f) = E_{(x,y)\sim\mathcal{D}}[\mathbb{I}(f(x) = y)]$,
$\mathbb{I}(cond)$ is 1 if cond is true, and 0 otherwise.

In a regression problem (when using MSE), Y=$\mathbb{R}$, and we need to minimize $error(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$.

### 2.2 Dataset

In practice we only have a finite dataset D sampled from $\mathcal{D}$. We usually divide D into training set and validation set. Use the training set to train the model f, and estimate $error(f)$ by the validation set.
There are several methods to make full use of the dataset:
(a). K-fold cross validation
Divide the dataset into k subsets(mutually exclusive).
Each time, the combination of (k - 1) subsets is used as the training set, the last one as the testing set.

(b). Leave one out
The special case of k-fold cross validation when k is equal to the size of the dataset.
(c) Bootstrapping
Randomly resample D' from D, and let D' has the same size of D.

### 2.3 Inductive bias

Inductive bias is the preference of a learning algorithm on choosing model f from the hypothesis space $\mathcal{H}$. It gives model the ability to generalize on unseen samples.
The "Occam's Razor principle" is a common inductive bias that prefers to choose simpler model. An unbiased learning algorithm knows nothing about the unseen data and is similar to a look-up table. The "No free lunch theorem" says that if a learning algorithm has high performance on some problems, it must pay the price of having low performance on certain tasks.

### 2.4 Bias and variance

On a specific sample distribution $\mathcal{D}$, the performance of a learning algorithm may vary with the change of training set. We use $f_D(x)$ to represent the model generated by training set D.
The bias of a learning algorithm is $\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\mathbb{E}[f_D(x)]-y)^2]}$. It's the inherent defect of the learning algorithm (or its hyperparameters) on this task that cannot be reduced by taking the average of multiple models training on independent training set. Models with a smaller capacity usually have larger biases.
The variance of a learning algorithm is $\mathbb{E}_{(x,y)\sim\mathcal{D}}[Var(f_D(x))]$. Large variance means that the output of the model on the same input varies greatly with the change of training set.
And we have
$\mathbb{E}_{(x,y)\sim\mathcal{D},D\in\mathcal{D}}[(f_D(x)-y)^2] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\mathbb{E}[f_D(x)]-y)^2] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[Var(f_D(x))]$,
it means that the error of a learning algorithm is consist of its bias and variance.

## 3 Methods

### 3.1 Decision Trees

Using divide and conquer to build a tree, each internal node corresponds to a condition, and each leaf node corresponds to a classification result.

### 3.2 Linear regression

When $X = \mathbb{R}^n, Y = \mathbb{R}$, linear regression $f(\vec{x}) = \vec{w}^T\vec{x} + \vec{b}$ and mean square error $\mathbb{E}[(f(\vec{x})-y)^2]$ can be used.

## 3.3 Neural Networks

Since nested linear models are still linear models, we need nonlinear functions to improve the model's capacity. Neural networks usually alternately using linear and non-linear mappings to produce y from x.