Lecture 1: Course Overview

Week 1 - Part I

Lecturer: Tianyu Wang

1 Overview

Consider an unknown function $f : \mathcal{X} \to \mathcal{Y}$. The core task of machine learning is to find a good approximation of f...

- When a set of data $\{(x_i, y_i)\}_{i=1}^n$ (often corrupted by noise) generated by the unknown function f is available and $\mathcal{Y}$ is finite (or countable), this is classification. Usually, $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$.
- When a set of data $\{(x_i, y_i)\}_{i=1}^n$ (often corrupted by noise) generated by the unknown function f is available and $\mathcal{Y}$ is a continuum subset of $\mathbb{R}$, this is the task of regression.
- When f is a density function and a set of data $\{x_i\}_{i=1}^n$ generated from the density function is available, this is the task of density estimation, which heavily overlaps with clustering.
- When the function f is "policy" (a mapping from state space to action space), this is a subfield called reinforcement learning.
- $\cdots$

To obtain a good approximation of f, one usually construct a plausible model $\widehat{f}$, which can be...

- linear,
- a neural network,
- non-parametric,
- $\cdots$

After constructing the model $\widehat{f}$, one can solve for $\widehat{f}$ using ...

- Optimization
- Statistical inference

1.1 Example: Linear Regression

Optimization perspective:

Suppose we have data $\{(x_i, y_i)\}_{i=1}^n$ and our model is $\widehat{f}_\theta(x) = \theta^\top x$ for some θ. Fitting the model $\widehat{f}_\theta(x)$ to data $\{(x_i, y_i)\}_{i=1}^n$ can be formulated as an optimization problem:

$$\min_\theta l(\theta),$$

where

$$l(\theta) = \frac{1}{n}\sum_{i=1}^n \left(\widehat{f}_\theta(x_i) - y_i\right)^2.$$

is the mean square error/loss of the model.

Statistical inference perspective:

Suppose we have data $\{(x_i, y_i)\}_{i=1}^n$ ($x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$) and our model assumes that the distribution of x and y is governed by: $y_i \overset{i.i.d.}{\sim} \mathcal{N}\left(\theta^\top x_i, \sigma^2\right)$ for some θ. Fitting the model to data $\{(x_i, y_i)\}_{i=1}^n$ can be formulated as a statistical inference problem. Assuming $\sigma = 1$, the likelihood function is

$$L_\theta(\{(x_i, y_i)\}_{i=1}^n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^n \exp\left(-\frac{1}{2}(y_i - \theta^\top x_i)^2\right).$$

The log-likelihood is

$$\log L_\theta(\{(x_i, y_i)\}_{i=1}^n) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n \left(\theta^\top x_i - y_i\right)^2.$$

The maximum-likelihood estimator (MLE) for θ can be obtained by maximizing the log-likelihood. That is,

$$\theta \in \arg\max_\theta \log L_\theta(\{(x_i, y_i)\}_{i=1}^n).$$

Note that maximizing the likelihood is equivalent to minimizing the aforementioned mean squared error.

Note 1. There is a Bayesian version of the story, which will be covered later.

Note 2. This linear regression example is perhaps the single most important example for this course. We will repeatedly come back to it.