A Short Introduction to Statistical Learning

BY JACK YANSONG LI

University of Illinois Chicago

Email: `yli340@uic.edu`

1 Background

To simplify the analysis, we only consider supervised learning in this note. A regression problem is defined as:

- Feature: x.
- Label: y.
- Goal: Given a set of data $\mathcal{D}_N = \{x_i, y_i\}_{i\in[N]}$, find a proper f such that $f(x)$ is close to y.

Assumption 1. There exist an unknown distribution $\mathbb{P}$ such that $(x_i, y_i) \overset{i.i.d}{\sim} \mathbb{P}$.

- Define a loss function $\hat{\mathcal{L}}(f, \mathcal{D}_N) \triangleq \sum_{i\in N} l(f(x_i), y_i)$ such that $\text{minimize}_{f\in\mathcal{F}} \quad \hat{\mathcal{L}}(f, \mathcal{D}_N) \to \hat{f}$.

However, in reality, the real expected loss we faces is defined as

$$\mathcal{L}(\hat{f}) \triangleq \mathbb{E}_{(x,y)\sim\mathbb{P}}[l(\hat{f}(x), y)].$$

The goal of machine learning is to solve the following optimization problem

$$\underset{f}{\text{minimize}} \quad \mathcal{L}(f).$$

Now, let's decomposite the true loss as following:

$$\begin{aligned}
\mathcal{L}(f) &= \mathcal{L}(f) - \mathcal{L}(\hat{f}) + \mathcal{L}(\hat{f}) \\
&= \mathcal{L}(f) - \mathcal{L}(\hat{f}) + \hat{\mathcal{L}}(\hat{f}, \mathcal{D}_N) + \mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f}, \mathcal{D}_N). \\
&= \mathcal{L}(f) - \mathcal{L}(\hat{f}) + \hat{\mathcal{L}}(\hat{f}, \mathcal{D}_N) - \hat{\mathcal{L}}(f, \mathcal{D}_N) + \mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f}, \mathcal{D}_N) + \hat{\mathcal{L}}(f, \mathcal{D}_N).
\end{aligned}$$

- (*Approximation*): minimize the approximation error $\mathcal{L}(f) - \mathcal{L}(\hat{f})$. *Neural net structure, linear or nonlinear?*
- (*Generalization*): minimize the generalization error $\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f}, \mathcal{D}_N)$. *Overfitting*.
- (*Optimization*): minimize the optimization error $\hat{\mathcal{L}}(\hat{f}, \mathcal{D}_N) - \hat{\mathcal{L}}(f, \mathcal{D}_N)$. *Gradient descent*, *Linear programming*.

Before deep learning, researchers in statistical/machine learning theory mainly focusd on generalization part, such as PAC theory. However, in deep learning:

- (*Approximation*): Many options on neural nets. which action function? how many layers? fully-connected or transformer or CNN? (UIUC)
- (*Optimization*): Stochastic gradient descent, nonconvex nonsmooth optimization. (Tengyu Ma)