Optimization Algorithms: Notes

Lo Chun, Chou
R13922136

March 4, 2025

W3

$$\begin{cases} \text{Statistical Learning (Vapnik-Chervonenkis)} \\ \text{PAC Learning (Leslie Valiant)} \\ \text{Online learning} \to \text{Online to batch conversion} \end{cases}$$

Statistical Learning = Stochastic Optimization

- Machine Learning = Decision making under uncertainty

Description in class

If we have data:

- train data: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$
- test data: (x_{n+1}, y_{n+1})

If we calculate the risk function by:

$$R(h) = \lambda(h, (x_{n+1}, y_{n+1}))$$

This is a little bit far from the true risk value $R(h)$, therefore we take the expected value:

$$R(h) = \mathsf{E}[\lambda(h, (x_{n+1}, y_{n+1}))]$$

Description in slides

Since we do not know what the test data is, we cannot know the risk value $R(h)$ for each hypothesis $h \in \mathcal{H}$ directly, therefore we cannot find the exact $h^\star \ni$

$$h^\star \in \arg\min_{h \in \mathcal{H}} R(h)$$

where the risk $R(h)$ is defined as:

$$R(h) = \mathsf{E}_z[\lambda(h, z)] \qquad \text{where } \lambda : \mathcal{H} \times \mathcal{Z} \to \mathbb{R} \text{ is the loss funciton}$$

Which means that the risk value $R(h)$ of a hypothesis h is the expected loss.

Empirical Risk Minimization (ERM)

By law of large numbers, if N is large enough, the empirical risk $\hat{R}_N(h)$ is approximate to the true risk $R(h)$:

$$\hat{R}_N(h) \coloneqq \frac{1}{N} \sum_{n=1}^{N} \lambda(h, z_n) \approx R(h)$$

Therefore, we can approximate $h^\star$ in the original optimization problem by:

$$\hat{h}_N \in \arg\min_{h \in \mathcal{H}} \hat{R}_N(h)$$

Note that here we use $\in$ since there could be multiple hs that achieve the minimum empirical risk.
If

$$\mathbb{P}\{\delta = \delta_i\} \sim \frac{1}{n}$$

then we have:

$$\begin{aligned} \hat{R}_n(h) &= \sum_{i=1}^{n} \mathbb{P}\{\delta = \delta_i\} \cdot \lambda(h, \delta_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \lambda(h, \delta_i) \end{aligned}$$

If we don't know the probability distribution of the random variable, we can use numerical integration to estimate it.

How large is the statistical error $R(\hat{h}_N) - R(h^\star)$?

Check this part:
$\rightarrow$ The difference between the empirical risk (avg. of sum of N sample points' loss), and the expected loss over $\mathcal{Z}$)

$$
\begin{aligned}
R(\hat{h}_N) - R(h^\star) &= R(\hat{h}_N) + \hat{R}_N(\hat{h}_N) - \hat{R}_N(\hat{h}_N) + \hat{R}_N(h^\star) - \hat{R}_N(h^\star) - R(h^\star) \\
&= R(\hat{h}_N) - \hat{R}_N(\hat{h}_N) + \hat{R}_N(\hat{h}_N) - \hat{R}_N(h^\star) + \hat{R}_N(h^\star) - R(h^\star)
\end{aligned}
$$

Both part in seagreen color consists of R and $\hat{R}_N$, with one using $\hat{h}_N$ and the other using $h^\star$.
Check:

- pointwise convergence vs. uniform convergence
- ULLN (Uniform law of large numbers)
- Dadley integral