Learning Note of Generalization Theory

20201102

1 Statement in the chapter 6 of ProbML

Let us initially consider the case where the hypothesis space is finite, which size $dim(\mathcal{H}) = |\mathcal{H}|$. In other words, we are selecting a model/ hypothesis from a finite list, rather than optimizing real-valued parameters. Then we can prove the following.

Theorem For any data distribution p_*, and any dataset $\mathcal{D}$ of size N drawn from p_*, the probability that our estimate of the error rate will be more than ϵ wrong, in the worst case, is upper bounded as follows:

$$P(\max_{h \in \mathcal{H}} |R_{emp}(\mathcal{D}, h) - R(p_*, h) > \epsilon| \leq 2dim(\mathcal{H})e^{2N\epsilon^2}$$

Proof. To prove this, we need two useful results. First, **Hoeffding's inequality**, which states that if $X_1, ..., X_N \in Ber(\theta)$, then, for any $\epsilon > 0$,

$$P(|\overline{x} - \theta| > \epsilon) \leq 2e^{-2N\epsilon^2},$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^{d} P(A_i)$. Second, the **union bound**, which says that if $A_1, ..., A_d$ are a set of events, then $P(\cup_{i=1}^{d} A_i) \leq \sum_{i=1}^{d} P(A_i)$.

Finally, for notational brevity, let $R(h) = R(h, p_*)$ be the true risk, and $\hat{R}_N(h) = R_{emp}(\mathcal{D}, h)$ be the empirical risk.

Using these results we have

$$\begin{aligned} P(\max_{h \in \mathcal{H}} |\hat{R}_N(h) - R(h)| > \epsilon) &= P(\cup_{h \in \mathcal{H}} |\hat{R}_N(h) - R(h)| > \epsilon) \\ &\leq \sum_{h \in \mathcal{H}} P(|\hat{R}_N(h) - R(h)| > \epsilon) \\ &\leq 2e^{2N\epsilon^2} = 2dim(\mathcal{H})e^{-2N\epsilon^2} \end{aligned} \tag{1}$$

If the hypothesis space $\mathcal{H}$ is infinite (e.g. we have real-valued parameters), we cannot use $dim(\mathcal{H}) = |\mathcal{H}|$. Instead, we can use a quantity called the VC dimension of the hypothesis class.

1.1 Problems

- 1. What is the VC dimension? How to compute the VC dimension for common ML models?
- 2. There are more than one version of Hoeffding's inequality. The using of Hoeffding's inequality is not clear here.
- 3. Does the RHS always less than 1?

The answer of the problem1 is as follows. We can understand the VC dimension intuitively with the number of effective parameters of a ML model. For the linear perceptions, the VC dimension is just d+1, which is the number of parameters.

The answer of the problem 2 is as follows. Please refer the definition of Hoeffding's inequality (hoeffding's-inequality.pdf) and this webpage (vc-theory-hoeffding-inequality.pdf) for more details.

The answer of the problem 3 is as follows. If the hypothesis space is finite, then the RHS will less than 1 if N is sufficiently large. If the hypothesis space is infinite, we should seek help from the VC dimension and then further discuss it.

2 Useful references

Blog of VC theory series:

- vc-theory-hoeffding-inequality.pdf
- vc-theory-symmetrization.pdf
- vc-theory-vapnik-chervonenkis-dimension.pdf
- Prof. Yaser's lectures: lecture 5 to lecture 7 link