1 To Split or Not To Split?

This is some section!

Question 1: In this case, our hypothesis space $\mathcal{H} = \{h_1, ..., h_M\}$ is finite with $|\mathcal{H}| = M$, which means that we can use **Theorem 3.2** to conclude that with probability $1 - \delta$ for all $h \in \mathcal{H}$

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}} \tag{1}$$

where $n = |S_{val}|$.

Question 2: Let S^*_{val} be the validation set om which we are testing the hypothesis $\hat{h}^*$ that we end up choosing. In the setup proposed by our fellow student, we are only testing a single hypothesis, namely $\hat{h}^*$, on S^*_{val}, and $|S_{val^*}| = \frac{n}{M}$. Therefore, we can use **Theorem 3.1** to conclude that with probability $1 - \delta$ for all $h \in \mathcal{H}$

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2\frac{n}{M}}} = \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{M \ln \frac{1}{\delta}}{2n}} \tag{2}$$

Our fellow student has therefore made a bad proposal, since bound is now growing linearly with M instead of logarithmically.

Question 3: Again we only test a single hypothesis, namely $\hat{h}^*$, on S^2_{val}. This time we have that $|S^2_{val}| = \frac{n}{2}$. Therefore, we can use **Theorem 3.1** to conclude that with probability $1 - \delta$ for all $h \in \mathcal{H}$

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2\frac{n}{2}}} = \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{n}} \tag{3}$$

Assume that my fellow student followed this procedure, and I followed the procedure in question 1. Let $\hat{h}^*$ be the hypothesis that I end up choosing, and let $\tilde{h}^*$ be the hypothesis my fellow student chooses. Where I am using the full S_{val} to choose $\hat{h}^*$, my fellow student is only using S^1_{val} to choose $\tilde{h}^*$. Therefore, we cannot not be sure that $\hat{h}^* = \tilde{h}^*$. Apart from not knowing whether we choose the same hypothesis, we also do not test our chosen hypothesis on the same set. Where I am using S_{val}, my fellow student is using S^2_{val}. All in all, it is therefore not very easy to tell know how close my empirical error $\hat{L}(\hat{h}^*, S_{val})$ is to the empirical error $\hat{L}(\tilde{h}^*, S^2_{val})$ of my fellow student. However, we can say that I have a higher probability of choosing the hypothesis h_i in $\mathcal{H}$ with the lowest expected loss $L(h_i)$, since I am using a bigger validation set to inform my decision.

If we assume that $\hat{L}(\hat{h}^*, S_{val}) = \hat{L}(\tilde{h}^*, S_{val}^2)$, then we know that my bound is tighter than my fellow student's, if and only if

$$\sqrt{\frac{\ln \frac{M}{\delta}}{2n}} < \sqrt{\frac{\ln \frac{1}{\delta}}{n}} \tag{4}$$

This is equivalent to

$$\ln \frac{M}{\delta} < 2 \ln \frac{1}{\delta} \tag{5}$$

which is equivalent to

$$\frac{M}{\delta} < \left(\frac{1}{\delta}\right)^2 \tag{6}$$

which is equivalent to

$$M\delta < 1 \tag{7}$$

This means that under the assumption that $\hat{L}(\hat{h}^*, S_{val}) = \hat{L}(\tilde{h}^*, S_{val}^2)$, then if we for instance wanted a certainty $1 - \delta = 0.95$, then my procedure would have a tighter bound, if and only if $M < 20$.

As I had already said, then even if we had a big M, my fellow student would still be less certain than me of picking the best hypothesis in $\mathcal{H}$, which is a drawback of his method.

Question 4: As I have already explained in question 3, then choosing a large α - and thereby a large validation set S_{val}^1 - means having a better chance of choosing the hypothesis in $\mathcal{H}$, which actually has the lowest expected loss, as $\hat{h}^*$. This also means that we should expect a lower empirical loss $L(\hat{h}^*, S_{val}^2)$ on the test set S_{val}^2 than if we had used a smaller validation set to choose $\hat{h}^*$. However, a large α also means a small test set. Therefore, we also get more uncertain how well the empirical loss $L(\hat{h}^*, S_{val}^2)$ on the test set reflects the true expected loss $L(\hat{h}^*)$. This can be seen by the fact that the term

$$\sqrt{\frac{\ln \frac{1}{\delta}}{2(1-\alpha)n}} \tag{8}$$

in our bound

$$L(\hat{h}^*) \leq= \hat{L}(\hat{h}^*, S_{val}^2) + \sqrt{\frac{\ln \frac{1}{\delta}}{2(1-\alpha)n}} \tag{9}$$

grows when α becomes larger. All in all, it therefore not clear whether a larger α will make us choose $\hat{h}^*$, such that the resulting bound on $L(\hat{h}^*)$ becomes larger or smaller. In general, the larger M becomes, the larger I would also choose α, since a large hypothesis space also means a large probability of accidentally choosing a bad hypothesis as $\hat{h}^*$, if the validation set is too small.

2 Occam's Razor

This is some section!

3 Kernels

This is some section!