Assignment 3

Silvan Robert Adrian - zlp432

December 11, 2018

Contents

1 The Role of Independence

So for this question we need to design an example of distributed but *dependent* Bernoulli variables, for which have to hold following condition:

$$P\left(|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i| \geq \frac{1}{2}\right) = 1 \tag{1}$$

By dependent meaning the next random variable in line is dependent on the one before:

$$X_i, X_{i+1}... \tag{2}$$

A good example for that would be markov chains. So we say that conditionally:

$$P(X_{i+1}|X_i) \tag{3}$$

otherwise I might misunderstood something totally.

2 To Split or Not To Split? (And How to Split.)

2.1 Question 1

The hypothesis space $\mathcal{H} = \{h_1, ..., h_M\}$ is finite in this question, which means $|\mathcal{H}| = M$ for which we can use **Theorem 3.2** from which we conclude with probability $1 - \delta$ for all $h \in \mathcal{H}$

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}} \tag{4}$$

where n is equal to $|S_{val}|$, when we insert that we end up in the following:

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{M}{\delta}}{2|s_{val}|}} \tag{5}$$

2.2 Question 2

So first let S^*_{val} be the validation set with which we are testing the hypothesis $\hat{h}^*$ on which we pick. According to our fellow student we test a single hypothesis $\hat{h}^*$ on S^*_{val} and the splitting of the validation set described as $|S_{val^*}| = \frac{n}{M}$. So we can use **Theorem 3.1** to conclude with probability $1-\delta$ for all $h \in \mathcal{H}$

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2\frac{n}{M}}} = \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{M \ln \frac{1}{\delta}}{2n}} \tag{6}$$

No it wasn't a good idea, since the bound is growing linear with M instead of logarithmically.

2.3 Question 3

a) 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 $h \in \mathcal{H}$

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2\frac{n}{2}}} = \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{n}} \tag{7}$$

b) Since we have my hypothesis and the hypothesis of my fellow student, let $\hat{h}^*$ be my hypothesis which I choose and $\bar{h}^*$ for the fellow student. I will use the full S_{val} to choose my hypothesis $\hat{h}^*$ while the fellow student uses S^1_{val} to choose $\bar{h}^*$ From the assignment text we then assume: $\hat{L}(\hat{h}^*, S_{val}) = \hat{L}(\bar{h}^*, S^2_{val})$, from that we know that my bound will be tighter, if:

$$\sqrt{\frac{\ln \frac{M}{\delta}}{2n}} < \sqrt{\frac{\ln \frac{1}{\delta}}{n}} \tag{8}$$

$$= \ln \frac{M}{\delta} < 2 \ln \frac{1}{\delta} \tag{9}$$

$$= \frac{M}{\delta} < \left(\frac{1}{\delta}\right)^2 \tag{10}$$

$$= M\delta < 1 \tag{11}$$

No we can say as an example that we want to have a certainty of $1-\delta = 0.90$, then my bound would only be tighter for $M < 10$.

c) Since I use a a bigger validation set then my fellow student I would have the higher probability of choosing h_i in $\mathcal{H}$ with lowest expected loss $L(h_i)$, even if M is large. Which I would see as a drawback of the chosen method.

2.4 Question 4

a) By choosing a large α we would have the advantage of also having a large validation set S^1_{val}, which means a better chance of choosing the hypothesis in $\mathcal{H}$ with the lowest expected loss. This also means we should expect a lower empirical loss on the test set S^2_{val}. The downside is that with a large α also the

test set gets smaller, so we got more uncertain how well the empirical loss reflect the true expected loss. By choosing a smaller α we end up in similar issues as described for a bigger α but the other way around, lower chance of choosing the hypothesis in $\mathcal{H}$ with the lowest expected loss.
b) I would make this selection according to the size of M, so when M becomes larger I would also choose α larger. Since the larger the hypothesis space gets, this also means a bigger probability of choosing a bad hypothesis $\hat{h}^*$.

3 Occam's Razor

3.1 Question 1

From the definition of the assignment text we can say that the size of Σ_d is the number of ways to choose d elements from Σ with replacement. Which means that:

$$|\Sigma_d| = |\Sigma|^d = 27^d \tag{12}$$

There is a on to one mapping of the elements of $\mathcal{H}_d$ and the power set $\mathcal{P}(\Sigma_d)$ of Σ_d:

$$|\mathcal{H}_d| = |\mathcal{P}(\Sigma_d)| = 2^{|\Sigma_d|} = 2^{27^d} \tag{13}$$

$\mathcal{H}_d$ in this case is finite, so we can use **Theorem 3.2** for concluding with probability $1 - \delta$ for all $h \in \mathcal{H}_d$:

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{|\mathcal{H}_d|}{\delta}}{2n}} = \hat{L}(h, S) + \sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}} \tag{14}$$

here S is a labeled sample of strings from Σ_d, and n is here $|S|$.

And we see now that the size of $\mathcal{H}_d$ grows exponentially with the size of d:

$$\sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}} \tag{15}$$

3.2 Question 2

Let $\mathcal{H}$ be defined as in the assignment text. For this questions we would use **Theorem 3.3**:

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{p(h)\delta}}{2n}} \tag{16}$$

From here I didn't know how exactly to proceed, any input would be appreciated.

3.3 Question 3

Well one of the trade-offs will be that the longer the strings get the term will grow exponentially, so there has to be found some middle ground to not let it grow too large.

4 Kernels

4.1 Distance in feature space

Take the definitions from the assignment text. From the definition of canonical norm for Hilbert space we get:

$$||\Phi(x) - \Phi(z)||^2 = \langle \Phi(x) - \Phi(z), \Phi(x) - \Phi(z) \rangle \tag{17}$$

But since any inner product of a Hilbert space must be linear for both arguments we get:

$$\langle \Phi(x) - \Phi(z), \Phi(x) - \Phi(z) \rangle = \langle \Phi(x), \Phi(x) \rangle + \langle \Phi(z), \Phi(z) \rangle - 2\langle \Phi(x), \Phi(z) \rangle \tag{18}$$

From the above we get:

$$||\Phi(x) - \Phi(z)|| = \sqrt{\langle \Phi(x), \Phi(x) \rangle + \langle \Phi(z), \Phi(z) \rangle - 2\langle \Phi(x), \Phi(z) \rangle} \tag{19}$$

We know that for all $x_1, x_2 \in \mathcal{X}$

$$k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle \tag{20}$$

From above it follows:

$$||\Phi(x) - \Phi(z)|| = \sqrt{k(x, z) + k(z, z) - 2k(x, z)} \tag{21}$$

4.2 Sum of kernels

Let A and B be the Gram matrix of k_1 and k_2, since they are kernels they are positive definite matrices, from which we can say:

$$\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j}^{m} c_i c_j A_{ij} \geq 0 \tag{22}$$

and the second one:

$$\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j}^{m} c_i c_j B_{ij} \geq 0 \tag{23}$$

Consider now the function $k_3 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$k_3(x, y) = k_1(x, y) + k_2(x, y) \tag{24}$$

As described in the assignment text, we want to prove that: $k_3(x, y) = k_1(x, y) + k_2(x, y)$ is also positive definite. For that we define another Gram matrix C from k_3. So we have to to add A and B together to get C, which looks like that:

$$C_{ij} = k_3(x_i, x_j) = k_1(x_i, x_j) + k_2(x_i, x_j) = A_{ij} + B_{ij} \tag{25}$$

When we move everything together from above we get the following:

$$\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j}^{m} c_i c_j C_{ij} = \sum_{i,j}^{m} c_i c_j (A_{ij} + B_{ij}) \tag{26}$$

$$= \sum_{i,j}^{m} c_i c_j A_{ij} + \sum_{i,j}^{m} c_i c_j B_{ij} \geq 0 \tag{27}$$

Which means that C is also positive-definite.

4.3 Rank of Gram matrix

-