1 To Split or Not To Split?

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 $\hat{L}(\hat{h}^*, S_{val}) = \hat{L}(\tilde{h}^*, S^2_{val})$, 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 $\hat{L}(\hat{h}^*, S_{val}) = \hat{L}(\tilde{h}^*, S^2_{val})$, 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^1_{val} - 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 $L(\hat{h}^*, S^2_{val})$ on the test set S^2_{val} 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 $L(\hat{h}^*, S^2_{val})$ 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

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

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

Question 1: Let $d \in \mathbb{N}_0$ and let Σ_d and $\mathcal{H}_d$ be defined as in the assignment text. Σ_d consists of all strings of length d, which can be constructed using letters from the alfabet Σ. Therefore, the size of Σ_d is the number of ways to choose d elements from Σ with replacement. This means that

$$|\Sigma_d| = |\Sigma|^d = 27^d \tag{10}$$

$\mathcal{H}_d$ consists of all functions $f : \Sigma_d \to \{0, 1\}$. There is a one-to-one correspondance between such functions and the subsets of Σ_d. To show this, we can just map any such function f to the subset $A_f = \{s \in \Sigma_d | f(s) = 1\}$, and map any subset A of Σ_d to the function $f_A : \Sigma_d \to \{0, 1\}$, where $f_A(s) = 1$, if and only if $s \in A$. Because there is a one-to-one correspondance between the elements of $\mathcal{H}_d$ and the power set $\mathcal{P}(\Sigma_d)$ of Σ_d, then

$$|\mathcal{H}_d| = |\mathcal{P}(\Sigma_d)| = 2^{|\Sigma_d|} = 2^{27^d} \tag{11}$$

Since $\mathcal{H}_d$ is finite, we can use **Theorem 3.2** to conclude that with probability $1 - \delta$ for all $h \in \mathcal{H}_d$

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{|\mathcal{H}_d|}{\delta}}{2n}} = \hat{L}(h, S) + \sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}} \tag{12}$$

where S is some labeled sample of strings from Σ_d, and $|S| = n$.

Since we have insisted to use a very complex hypothesis space, the size of $\mathcal{H}_d$ grows double exponentially as a function of d. This means that the term

$$\sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}} \tag{13}$$

grows exponentially as a function of d. Therefore, we would have to choose d quite small or have a very large sample size n in order to get a useful bound in practice.

Question 2: Let $\mathcal{H}$ be defined as in the assignment text. Since $\mathcal{H}_d$ is finite for all $d \in \mathbb{N}_0$, then $\mathcal{H}_d$ is countable for all $d \in \mathbb{N}_0$. Therefore, $\mathcal{H}$ is a countable union of countable sets, which means that $\mathcal{H}$ is also countable. Therefore, we can use **Theorem 3.3** to conclude that with probability $1 - \delta$ for all $h \in \mathcal{H}$

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{p(h)\delta}}{2n}} \tag{14}$$

where $p : \mathcal{H} \to (0, 1)$ is some function defined independently of S with $\sum_{h \in \mathcal{H}} p(h) \leq 1$. What set is S in the context of this question? Any specific hypothis $h \in \mathcal{H}$ belongs to $\mathcal{H}_d$ for some specific $d \in \mathbb{N}_0$. In other words, any $h \in \mathcal{H}$ is only defined for strings of a specific length d. We therefore have a many-to-one mapping $d : \mathcal{H} \to \mathbb{N}_0$, where $d(h)$ is the length of strings that h is defined for. With this frasing of the problem we can say that for each $h \in \mathcal{H}$, S must be a labeled sample of strings from $\Sigma_{d(h)}$. We can also use the function $d : \mathcal{H} \to \mathbb{N}_0$ to define a function $p : \mathcal{H} \to (0, 1)$ by

$$p(h) = \frac{1}{2^{d(h)+1}} \frac{1}{2^{27^{d(h)}}} \tag{15}$$

Since $\sum_{d=0}^{\infty} \frac{1}{2^{d+1}} = 1$, then $\sum_{d=0}^{\infty} p(h) \leq 1$. Therefore, we can substitute $p(h)$ in on line (14), by which we get that with probability $1 - \delta$ for all $h \in \mathcal{H}$

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{2^{d(h)+1} 2^{27^{d(h)}}}{\delta}}{2n}} \tag{16}$$

Question 3: The term

$$\sqrt{\frac{\ln \frac{2^{d(h)+1} 2^{27^{d(h)}}}{\delta}}{2n}} \tag{17}$$

grows exponentially as a function of $d(h)$. However, we should also expect the term

$$\hat{L}(h, S) \tag{18}$$

to decrease as a function of $d(h)$, since a h with a higher $d(h)$ uses more information - that is, longer strings - to make its predictions. In terms of picking a h that optimizes the bound, the question is if this decrease outweighs the growth in the other term, also caused by having h defined on longer string lengths.

3 Kernels

3.1 Distance in feature space

Let k be a kernel on input space $\mathcal{X}$ defining the RKHS $\mathcal{H}$, and let $\Phi : \mathcal{X} \to \mathcal{H}$ be the corresponding feature map. Let $\langle \cdot \rangle$ be the inner product on $\mathcal{H}$, which has been defined as part of the construction of $\mathcal{H}$ as the RKHS of k.

Let $x, z \in \mathcal{X}$. By definition of the canonical norm on a Hilbert space, we know that

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

Since any inner product on a Hilbert space must be linear in both arguments, we get that

$$\langle \Phi(x) - \Phi(z), \Phi(x) - \Phi(z) \rangle = \langle \Phi(x), \Phi(x) \rangle + \langle \Phi(z), \Phi(z) \rangle - 2\langle \Phi(x), \Phi(z) \rangle \tag{20}$$

By line (19-20) we know get that

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

By construction of the RKHS of k, we know that for all $x_1, x_2 \in \mathcal{X}$

$$k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle \tag{22}$$

By line (21-22) we know have that

$$||\Phi(x) - \Phi(z)|| = \sqrt{k(x, z) + k(z, z) - 2k(x, z)} \tag{23}$$

which is what I was asked to show.

3.2 Sum of kernels

Let $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be kernels[1]. Let $x_1, ..., x_m \in \mathcal{X}$, and let A and B be the Gram matrix of k_1 and k_2, respectively, with respect to $x_1, ..., x_m$. Since k_1 and k_2 are kernels, A and B are positive definit matrices, which means

$$\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j}^{m} c_i c_j A_{ij} \geq 0 \tag{24}$$

[1]I omit to say positive definit kernels, since it is a part of the definition of a kernel that it is positive definit.

and

$$\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j}^{m} c_i c_j B_{ij} \geq 0 \tag{25}$$

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

$$k_3(x, y) = k_1(x, y) + k_2(x, y) \tag{26}$$

Let C be the Gram matrix of k_3 with respect to $x_1, ..., x_m$. By definition of C and k_3 we have that

$$C_{ij} = k_3(x_i, x_j) = k_1(x_i, x_j) + k_2(x_i, x_j) = A_{ij} + B_{ij} \tag{27}$$

By line (24-25) and (27) we now get that

$$\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j}^{m} c_i c_j C_{ij} = \sum_{i,j}^{m} c_i c_j (A_{ij} + B_{ij}) \tag{28}$$

$$= \sum_{i,j}^{m} c_i c_j A_{ij} + \sum_{i,j}^{m} c_i c_j B_{ij} \geq 0 \tag{29}$$

This means that C is positive definit.

Since $x_1, ..., x_m$ was arbitrary we now have that for all $m \in \mathbb{N}$ and for all $x_1, ..., x_m$, then the Gram matrix of the function k_3 with respect to $x_1, ..., x_m$ is positive definit. This means that k_3 is a kernel function.

All in all, I have now shown that if k_1 and k_2 are kernels on input space $\mathcal{X}$, then the function $k_3 = k_1 + k_2$ is also a kernel on $\mathcal{X}$.

3.3 Rank of Gram matrix

Let me start by proving a general theorem in linear algegra, namely that for all matrices X with elements in the real numbers, we have that

$$N(X^T X) = N(X) \tag{30}$$

where N is the null space of a matrix.

Let X be a matrix with real elements. Let $x \in N(X)$. By definition of the null space, this means that

$$Xx = \overline{0} \tag{31}$$

By the standard properties of matrices, it hereby follows that

$$(X^T X)x = (X^T)(Xx) = X^T\overline{0} = \overline{0} \tag{32}$$

which means that

$$x \in N(X^T X) \tag{33}$$

We have now shown that

$$N(X^T X) \subset N(X) \tag{34}$$

Now assume that $x \in N(X^T X)$. By definition of the null space, this means that

$$(X^T X)x = \overline{0} \tag{35}$$

By the standard properties of matrices with real elements, it hereby follows that

$$||Xx||^2 = (Xx)^T(Xx) = (x^T X^T)(Xx) = x^T((X^T X)x) = x^T\overline{0} = 0 \tag{36}$$

which implies that

$$||Xx|| = 0 \tag{37}$$

which implies that

$$Xx = \overline{0} \tag{38}$$

which means that

$$x \in N(X) \tag{39}$$

We have now shown that

$$N(X) \subset N(X^T X) \tag{40}$$

Line (34) and (40) together implies that the theorem stated on line (30) is true.

The rank-nullity theorem of linear algebra tells us that if X is some matrix with n columns, then

$$rank(X) + dim(N(X)) = n \tag{41}$$

By the theorem I have just proven, it follows that for all matrices X with real elements

$$dim(N(X)) = dim(N(X^T X)) \tag{42}$$

By this and the rank-nullity theorem we get that for all matrices X with real elements

$$rank(X) = rank(X^T X) \tag{43}$$

Let me now use this generel result to prove a bound on the rank of Gram Matrices arising from a linear kernel, $k(x, z) = x^T z$ for $x, z \in \mathbb{R}^d$, on the input space $\mathbb{R}^d$.

Let $x_1, ..., x_m \in \mathbb{R}^d$. Construct the matrix X by letting the vector x_i by the i^{th} column of X. By the definition of matrix multiplication, this means that for all $i, j \in 1, ..., m$

$$(X^T X)_{ij} = x_i^T x_j = k(x_i, x_j) \tag{44}$$

By the definition of the Gram matrix of k with respect to $x_1, ..., x_m$, this means that for all $i, j \in 1, ..., m$

$$(X^T X)_{ij} = G_{ij} \tag{45}$$

which means that

$$X^T X = G \tag{46}$$

By line (43), this gives us that

$$rank(G) = rank(X^T X) = rank(X) \tag{47}$$

Since X has d rows and m columns, then

$$rank(X) \leq \min(d, m) \tag{48}$$

It hereby follows that

$$rank(G) \leq \min(d, m) \tag{49}$$

We have now proven that, if we define the kernel k as above on the input space $\mathbb{R}^d$, then for all training points $x_1, ..., x_m \in \mathbb{R}^d$, the rank of the Gram matrix G of k with respect to these training points is bounded by

$$rank(G) \leq \min(d, m) \tag{50}$$

In any practical learning problem, we hopefully have that the number m of training points is larger than the number d of features. In that case, we will have that

$$rank(G) \leq d \tag{51}$$

In that case, G will not have full rank, since G is an $m \times m$ matrix.