# Assignment 3

Silvan Robert Adrian - zlp432

December 11, 2018

## Contents



# 1 The Role of Independence

# 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 $h \in \mathcal{H}$

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

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

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

## 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 $h \in \mathcal{H}$

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

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 $h \in \mathcal{H}$

$$L(\hat{h}^*) \leq \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2\frac{n}{2}}} = \hat{L}(\hat{h}^*, S_{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{n}} \tag{4}$$

**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:

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

$$= \ln \frac{M}{\delta} < 2 \ln \frac{1}{\delta} \tag{6}$$

$$= \frac{M}{\delta} < \left(\frac{1}{\delta}\right)^2 \tag{7}$$

$$= M\delta < 1 \tag{8}$$

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 $\alpha$ 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 $\alpha$ 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 $\alpha$ we end up in similar issues as described for a bigger $\alpha$ but the other way around, lower change 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 $\alpha$ larger. Since the larger the hypothesis space gets, this also means a bigger probability of choosing a bad hypothesis $\hat{h}^*$, if the validation set is too small.

## 3 Occam's Razor

### 3.1 Question 1

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

$$|\Sigma_d| = |\Sigma|^d = 27^d \tag{9}$$

$\mathcal{H}_d$ consists of all functions $f : \Sigma_d \rightarrow \{0,1\}$. There is a one-to-one correspondance between such functions and the subsets of $\Sigma_d$. 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 $\Sigma_d$ to the function $f_A : \Sigma_d \rightarrow \{0,1\}$, 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 $\Sigma_d$, then

$$|\mathcal{H}_d| = |\mathcal{P}(\Sigma_d)| = 2^{|\Sigma_d|} = 2^{27^d} \tag{10}$$

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

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{|\mathcal{H}_d|}{\delta}}{2n}} = \hat{L}(h,S) + \sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}} \tag{11}$$

where $S$ is some labeled sample of strings from $\Sigma_d$, 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

$$\sqrt{\frac{\ln \frac{2^{27^d}}{\delta}}{2n}} \tag{12}$$

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.

### 3.2 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 $h \in \mathcal{H}$

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{1}{p(h)\delta}}{2n}} \tag{13}$$

where $p : \mathcal{H} \rightarrow (0,1)$ 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 $d : \mathcal{H} \rightarrow \mathbb{N}_0$, 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

$$p(h) = \frac{1}{2^{d(h)+1}} \frac{1}{2^{27^{d(h)}}} \tag{14}$$

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 $h \in \mathcal{H}$

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{2^{d(h)+1} 2^{27^{d(h)}}}{\delta}}{2n}} \tag{15}$$

### 3.3 Question 3

The term

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

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

$$\hat{L}(h, S) \tag{17}$$

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.

## 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:

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

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

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

From the above we get:

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

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

$$k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle \tag{21}$$

From above it follows:

$$||\Phi(x) - \Phi(z)|| = \sqrt{k(x, z) + k(z, z) - 2k(x, z)} \tag{22}$$

## 4.2 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

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

and

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

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

$$k_3(x, y) = k_1(x, y) + k_2(x, y) \tag{25}$$

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

$$C_{ij} = k_3(x_i, x_j) = k_1(x_i, x_j) + k_2(x_i, x_j) = A_{ij} + B_{ij} \tag{26}$$

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

$$\forall c_1, ..., c_m \in \mathbb{R} : \sum_{i,j}^{m} c_i c_j C_{ij} = \sum_{i,j}^{m} c_i c_j (A_{ij} + B_{ij}) \tag{27}$$

$$= \sum_{i,j}^{m} c_i c_j A_{ij} + \sum_{i,j}^{m} c_i c_j B_{ij} \geq 0 \tag{28}$$

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 $\mathcal{X}$.

## 4.3 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

$$N(X^T X) = N(X) \tag{29}$$

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

$$Xx = \overline{0} \tag{30}$$

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

By the standard properties of matrices, it hereby follows that

$$(X^T X)x = (X^T)(Xx) = X^T \overline{0} = \overline{0} \tag{31}$$

which means that

$$x \in N(X^T X) \tag{32}$$

We have now shown that

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

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

$$(X^T X)x = \overline{0} \tag{34}$$

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

$$||Xx||^2 = (Xx)^T(Xx) = (x^T X^T)(Xx) = x^T((X^T X)x) = x^T \overline{0} = 0 \tag{35}$$

which implies that

$$||Xx|| = 0 \tag{36}$$

which implies that

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

which means that

$$x \in N(X) \tag{38}$$

We have now shown that

$$N(X) \subset N(X^T X) \tag{39}$$

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

$$rank(X) + dim(N(X)) = n \tag{40}$$

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

$$dim(N(X)) = dim(N(X^T X)) \tag{41}$$

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

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

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 $i, j \in 1, ..., m$

$$(X^T X)_{ij} = x_i^T x_j = k(x_i, x_j) \tag{43}$$

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

$$(X^T X)_{ij} = G_{ij} \tag{44}$$

which means that

$$X^T X = G \tag{45}$$

By line (43), this gives us that

$$rank(G) = rank(X^T X) = rank(X) \tag{46}$$

Since $X$ has $d$ rows and $m$ columns, then

$$rank(X) \leq \min(d, m) \tag{47}$$

It hereby follows that

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

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

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

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

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

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