1 The growth function

Question 1: Let $\mathcal{H}$ be any hypothesis set, where $h : \mathcal{X} \to \{-1, 1\}$ for all $h \in \mathcal{H}$. Abu-Mostafa 2015 defines that

$$\mathcal{H}(x_1, ..., x_N) = \{(h(x_1), ..., h(x_N)) \mid h \in \mathcal{H}\} \tag{1}$$

for specific $x_1, ..., x_N \in \mathcal{X}$, and further defines the growth function $m_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ as

$$m_{\mathcal{H}}(N) = \max_{x_1,...,x_N \in \mathcal{X}} |\mathcal{H}(x_1, ..., x_N)| \tag{2}$$

Since $h(x) \in \{-1, 1\}$ for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$, we must have for all $N \in \mathbb{N}$ and all $x_1, ..., x_N \in \mathcal{X}$:

$$\mathcal{H}(x_1, ..., x_N) \subset \{-1, 1\}^N = \{(y_1, ..., y_N) \mid y_1, ..., y_N \in \{-1, 1\}\} \tag{3}$$

which implies that for all $N \in \mathbb{N}$:

$$m_{\mathcal{H}}(N) \leq |\{-1, 1\}^N| = 2^N \tag{4}$$

This is true whether or not $\mathcal{H}$ is finite or infinite.

If we now assume that $\mathcal{H}$ is finite with $\mathcal{H} = \{h_1, ..., h_M\}$, then we get that for all $N \in \mathbb{N}$ and all $x_1, ..., x_N \in \mathcal{X}$:

$$\mathcal{H}(x_1, ..., x_N) = \{(h_i(x_1), ..., h_i(x_N)) \mid i \in \{1, ..., M\}\} \tag{5}$$

which implies that for all $N \in \mathbb{N}$ and all $x_1, ..., x_N \in \mathcal{X}$:

$$|\mathcal{H}(x_1, ..., x_N)| \leq M \tag{6}$$

which implies that for all $N \in \mathbb{N}$:

$$m_{\mathcal{H}}(N) \leq M \tag{7}$$

By line (1) and (7) we now have that for all $N \in \mathbb{N}$:

$$m_{\mathcal{H}}(N) \leq \min(M, 2^N) \tag{8}$$

Question 2: Let $\mathcal{H}$ be any hypothesis set, where $h : \mathcal{X} \to \{-1, 1\}$ for all $h \in \mathcal{H}$. Define the shattered sample sizes $\mathcal{N}_{\mathcal{H}}$ for $\mathcal{H}$ as:

$$\mathcal{N}_{\mathcal{H}} = \{N \in \mathbb{N} \mid m_{\mathcal{H}}(N) = 2^N\} \tag{9}$$

We can now rewrite Abu-Mostafa 2015's definition of the VC-dimension d_{VC} of $\mathcal{H}$ as

$$d_{VC}(\mathcal{H}) = \max \mathcal{N}_{\mathcal{H}} \tag{10}$$

where

$$\max \mathcal{N}_{\mathcal{H}} = \infty \tag{11}$$

if $\mathcal{N}_{\mathcal{H}} = \mathbb{N}$.

Assume that $\mathcal{H}$ is finite with $|\mathcal{H}| = M$. Let me now proof that for all $N \in \mathcal{N}_{\mathcal{H}}$:

$$N \leq \log_2 M \tag{12}$$

Assume that $N \in \mathcal{N}_{\mathcal{H}}$. By definition of $\mathcal{N}_{\mathcal{H}}$ we now have that

$$m_{\mathcal{H}}(N) = 2^N \tag{13}$$

Since $|\mathcal{H}| = M$, we know from question 1 that

$$m_{\mathcal{H}}(N) = \min(M, 2^N) \tag{14}$$

From combining line (13) and (14), we now get that

$$2^N \leq M \tag{15}$$

which implies that

$$N \leq \log_2 M \tag{16}$$

I have now proven that for all $N \in \mathcal{N}_{\mathcal{H}}$:

$$N \leq \log_2 M \tag{17}$$

This clearly means that

$$d_{VC}(\mathcal{H}) = \max \mathcal{N}_{\mathcal{H}} \leq \log_2 M \tag{18}$$

I have now proven that for all hypothesis sets $\mathcal{H}$, where $h : \mathcal{X} \to \{-1, 1\}$ for all $h \in \mathcal{H}$ and $|\mathcal{H}| = M$, we have that

$$d_{VC}(\mathcal{H}) \leq \log_2 M \tag{19}$$

Question 3: Let $\mathcal{H}$ be any hypothesis set, where $h : \mathcal{X} \to \{-1, 1\}$ for all $h \in \mathcal{H}$. As I argued in question 1, we have that for all $N \in \mathbb{N}$:

$$m_{\mathcal{H}}(N) \leq 2^N \tag{20}$$

which implies that for all $N \in \mathbb{N}$:

$$m_{\mathcal{H}}(2N) \leq 2^{2N} = \left(2^N\right)^2 \tag{21}$$

which together with line (20) implies that for all $N \in \mathbb{N}$:

$$m_{\mathcal{H}}(N) \leq \left(m_{\mathcal{H}}(N)\right)^2 \tag{22}$$

Question 4: Unfortunately, I have not had the time to solve this question.

2 VC-dimension

Unfortunately, I have not had the time to solve this part of the assigment.

3 Generalization Bound for Learning Gaussian RBF Kernel Bandwith

Unfortunately, I have not had the time to solve this part of the assigment.

4 Random Forests

Question 1: Is nearest neighbor classification affected by normalizing each feature to having mean equal to 0 and standard deviation equal to 1? Yes. When we use nearest neighbor classification, we need to calculate the distance between each sample point according to some distance metric on the input space $\mathcal{X}$. If each point consists of n real valued features, that is $\mathcal{X} = \mathbb{R}^n$, a very common distance metric is just the normal dot product on $\mathbb{R}^n$. With this metric the order of the distances between a point in $\mathcal{X}$ and the different sample points might be different, if we scale the features differently: The larger the variance and the absolute value of the mean a feature has, the more it will influence the how far two points are from each other.

Question 2: Unfortunately, I have not had the time to solve this part of the assigment.