Exercise 12

To show:

Consider the class $\mathcal{A}$ of all sets of the form

$$A_\alpha = \{x \in \mathbb{R} : sin(\alpha x) > 0\}$$

where $\alpha > 0$. What is VC dimension of $\mathcal{A}$? (Note that $\mathcal{A}$ has one free parameter.)

Solution:

The VC-Dimension of the class is **infinity**. In order of showing that, start by defining n different points (x_i) with corresponding labels y_i. In paritcular let x_i be $x_i = 2^{-i}$ and choose y to be $y_i \in \{0, 1\}$. In order to show the proposed statement one needs to show that it is possible to shatter any 2^n combinations of points for a specific level of α. Hence, start by choosing also alpha in the following way:

$$\alpha = \pi\left(1 + \sum_{i=1}^{n} 2^i\big(1 - y_i\big)\right)$$

Use both α and x and proceed as follows:

$$\begin{aligned}\alpha x_j &= \left(\pi\big(1 + \sum_{i=1}^{n} 2^i(1 - y_i)\big)\right)2^{-j} \\ &= \pi\left(2^{-j} + \sum_{i=1}^{n} 2^{i-j}(1 - y_i)\right)\end{aligned}$$

In the next step split the sum. Furthermore, notice that the last term can be dropped since it only adds some multiple of π, which doesn't affect the value of the sin function we are interested in:

$$\begin{aligned}&= \pi\left(2^{-j} + \sum_{i=1}^{j-1} 2^{i-j}\big(1 - y_i\big) + 2^0(1 - y_j) + \sum_{i=}^{n-j} 2^i\big(1 - y_i\big)\right) \\ &\implies \pi\left(2^{-j} + \sum_{i=1}^{j-1} 2^{i-j}(1 - y_i) + (1 - y_j)\right) \\ &= \pi\left(2^{-j} + \sum_{i=1}^{j-1} 2^{-i}(1 - y_i) + (1 - y_j)\right) > \pi(1 - y_j) \qquad \textbf{(1)}\end{aligned}$$

Continou by upper bounding the left hand side of the last inequality as follows:

$$\begin{aligned}
&\pi\Big(2^{-j} + \sum_{i=1}^{j-1} 2^{-i}(1-y_i) + (1-y_j)\Big) \\
&\leq \pi\Big(2^{-j} + \sum_{i=1}^{j-1} 2^{-i} + (1-y_j)\Big) \\
&= \pi\Big(\sum_{i=1}^{j} 2^{-i} + (1-y_j)\Big)
\end{aligned}$$

Finally, using the last expression to obtain a second bound:

$$\textbf{(2)} \quad \pi\Big(\underbrace{\sum_{i=1}^{j} 2^{-i}}_{<1} + (1-y_j)\Big) < \pi\big(2-y_i\big)$$

Recall the previous result from **(1)**, which is:

$$\textbf{(1)} \quad \pi\Big(2^{-j} + \sum_{i=1}^{j-1} 2^{-i}(1-y_i) + (1-y_j)\Big) > \pi(1-y_j)$$

Use those two bounds to conclude that is possible to any subsets of points using a specific α:

$$\begin{aligned}
&\text{For: } y_j = 0 \quad \overset{\text{By (1)}}{\Longrightarrow} \quad \alpha x_j > \pi\big(1-0\big) = \pi \quad \Longrightarrow \quad \sin(\alpha x_j) > 0 \\
&\text{For: } y_j = 1 \quad \overset{\text{By (2)}}{\Longrightarrow} \quad \alpha x_j < \pi\big(2-1\big) = \pi \quad \Longrightarrow \quad \sin(\alpha x_j) < 0
\end{aligned}$$

Exercise 13

To show:

Let $\mathcal{A}_1, \ldots, \mathcal{A}_k$ be classes of sets, all of the with VC dimension at most V. Show that the VC-dimension of $\cup_{i=1}^{k} \mathcal{A}_i$ is at most $4V \log_2(2V) + 4k$. You may use the fact that for $\alpha \geq 1$ and $b > 0$,if $x \geq 4a \log(2a) + 2b$ then $x \geq a \log x + b$.
Can you bound the VC-dimension of the class of all sets of the form

$$A_1 \cup \cdots \cup A_k \quad \text{with} \quad A_1 \in \mathcal{A}_1, \ldots, A_k \in \mathcal{A}_k?$$

Solution:

As a start, let $\mathcal{A}$ be the union of classes $\mathcal{A}_i$ ($\mathcal{A} = \cup_{i=1}^{k} \mathcal{A}_i$). The Shatter Coeffiecient for the union can be bounded as follows:

$$S_{\mathcal{A}}(n) \leq S_{\mathcal{A}_1}(n) + \cdots + S_{\mathcal{A}_k}(n) \leq k(n+1)^V$$

Consider now the VC-Dimension of the union class to be sucht that $S_{\mathcal{A}}(u) = 2^u$. Combine results and rearrange the terms:

$$\begin{aligned} 2^u &= k(u+1)^V \\ u &= \log_2(k(u+1)^V) \\ &= \log_2 k + V log_2(u+1) \end{aligned}$$

Notice that $\log k \leq 2k$ and use that fact for the next step:

$$\leq 2k + V \log_2(u+1)$$

Furthermore, notice that $\log_2(u) \approx \log_2(u+1)$ for sufficiently large u. Sustitute that in the former equation to obtain:

$$\approx 2k + V \log_2(u)$$

Notice that this expression matches in the structure the fact provided in the exercise text. Consider in this particular case: $a = V$, $x = u$ and $b = 2k$, to obtain the final result:

$$2k + V \log_2(u) \quad \rightarrow \quad 4V \log_2(2V) + 4k$$

And hence it follows:

$$u \leq 4V \log_2(2V) + 4k$$

In the second part there shall be found a bound for the union of sets A_i, where each set belongs to a class $\mathcal{A}_i$. Hence, start as follows:

$$S_{\mathcal{A}}(n) \leq S_{\mathcal{A}_1}(n) \cdot, \ldots, \cdot S_{\mathcal{A}_k}(n) \leq (u+1)^{Vk}$$

The modus operandi follows the same structural procedure :

$$\begin{aligned} 2^u &= (u+1)^{Vk} \\ u &= Vk \log_2(u+1) \\ &\approx Vk \log_2(u) \end{aligned}$$

Notice, some slight modification. Since b has to be greater than zero, we add a very small (positive) number (say δ) to the right hand side. Naturally this is an upper bound to the last expression.

$$\leq Vk \log_2(u) + \delta$$

Now use the same property as before and let in this case the variables be $a = Vk$,$b = \delta$ and $x = u$ to obtain the final result:

$$u \leq 4Vk \log_2(2Vk) + 2\delta$$

Exercise 14

To show:

(PERCEPTRON CONVERGENCE.) Consider the "normalized" version of the perceptron algorithm in which one starts with a nonzero vector $w_0 \in \mathbb{R}^d$ and, cycling through the data, one sets, for $t = 1, 2, \ldots$

$$w_t = \begin{cases} w_{t-1} & w_{t-1}^T X_t Y_t \geq 0 \\ w_{t-1} + Y_t X_t / ||X_t|| & otherwise \end{cases}$$

Suppose that the data are linearly separable. This means that there exists $w_* \in \mathbb{R}^d$ such that $w_*^T X_t Y_t \geq 1$ for all $i = 1, ;n$. Prove that the algorithm finds a classifier that separates the data in at most $||w_* - w_0||^2$ updates.
Hint: Prove that whenever w t is updated (i.e $w_{t-1}^T X_t Y_t \geq 0$) one has $||w_t - w_*||^2 \leq ||w_{t-1} - w_*||^2 - 1$

Solution:

The first part of the exercise aims to show the following:

$$||w_t - w_*||^2 \leq ||w_{t-1} - w_*|| - 1$$

In doing so assume that the algorithm updates, such that w_t becomes:

$$w_t = w_{t-1} + \frac{Y_t X_t}{||X_t||}$$

Substitute w_t in the left hand side of the first equation to obtain the following:

$$\left|\left| w_{t-1} - w_* + \frac{Y_t X_t}{||X_t||} \right|\right|^2$$

Expanding the last expression leads to:

$$\begin{aligned} &= ||w_{t-1} - w_*||^2 + \underbrace{\left|\left| \frac{Y_t X_t}{||X_t||} \right|\right|^2}_{=1} + 2(w_{t-1} - w_*)^T \left(\frac{Y_t X_t}{||X_t||} \right) \\ &= ||w_{t-1} - w_*||^2 + 1 + 2(w_{t-1} - w_*)^T \left(\frac{Y_t X_t}{||X_t||} \right) \\ &= ||w_{t-1} - w_*||^2 + 1 + 2 \underbrace{w_{t-1} \frac{Y_t X_t}{||X_t||}}_{<0} - 2 \underbrace{w_* \frac{Y_t X_t}{||X_t||}}_{\geq 1} \\ &\leq ||w_{t-1} - w_*||^2 - 1 \end{aligned}$$

The second part of the exercise aims to show the number of steps till termination. Hence, start with the previous result:

$$||w_t - w_*||^2 \leq ||w_{t-1} - w_*||^2 - 1$$

Consider in the following the iterations:

$$\begin{aligned} ||w_t - w_*||^2 &\leq ||w_{t-2} - w_*||^2 - 2 \\ &\vdots \\ ||w_t - w_*||^2 &\leq ||w_0 - w_*||^2 - t \end{aligned}$$

Denote $w_t = w_*$, since the algorithm terminates to obtain the final step and the solution:

$$\begin{aligned} 0 &\leq ||w_0 - w_*||^2 - t \\ t &\leq ||w_0 - w_*||^2 \end{aligned}$$

Exercise 15

To show:

Let g_n be an abitrary (data-dependent) classifier. The leave-one-out error estimate is defined as

$$R_n^{(D)}(g_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{g_{n-1}(X_i, Dn, i) \neq Y_i\}}$$

where $D_{n,i} = ((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n))$. Show that the estimate is nearly unbiased in the sense that

$$\mathbb{E}R_n^{(D)}(g_n) = \mathbb{E}R^{(D)}(g_{n-1})$$

Use this to derive a bound for the expected risk of a perceptron classifier of the previous exercise when the data are linearly separable (i.e., $L^* = 0$ and the Bayes classifier is liear).

Solution:

Start with the definition of the risk given in the exercise:

$$\begin{aligned}
\mathbb{E}\big[R_n^{(D)}(g_n)\big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{g_{n-1}(X_i, Dn, i) \neq Y_i\}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}_{\{g_{n-1}(X_i, Dn, i) \neq Y_i\}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{g_{n-1}(X_i \neq Y_i\}}\middle| D_n\right]\right] \\
&= \frac{1}{n} n \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{g_{n-1}(X) \neq Y\}}\middle| D_n\right]\right] \\
&= \mathbb{E}\Big[\mathbb{P}\big(g_{n-1}(X) \neq Y) | D_n\big)\Big] \\
&= \mathbb{E}\big[R(g_{n-1})\big]
\end{aligned}$$

Assuming for the second part of the exercise :

$$\mathbb{E}\left[R_n^{(D)}(g_n)\right] = \mathbb{E}\left[R^{(D)}(g_{n-1})\right] \approx \mathbb{E}\left[R_n^{(D)}(g_{n-1})\right]$$

Using now the the definition for the empirical risk given by the exercise to bound the perceptron classifier. Notice that the summation term can be interpreted as the number of updates the perceptropn classifier is doing. Using Novikoff's theorem outlined in the lectures the perceptron algorithm terminates after at most $\left(\frac{R}{\gamma}\right)^2$ iterations (where R is $\max ||X_i||$, and γ is the margin). Using that one gets the following:

$$\begin{aligned}\mathbb{E}\left[R_n^{(D)}(g_n)\right] &= \mathbb{E}\Bigg[\frac{1}{n}\underbrace{\sum_{i=1}^{n}\mathbf{1}_{\{g_{n-1}(X_i,Dn,i)\neq Y_i\}}}_{\leq \left(\frac{R}{\gamma}\right)^2}\Bigg] \\ &\leq \frac{1}{n}\mathbb{E}\left[\left(\frac{R}{\gamma}\right)^2\right]\end{aligned}$$

Exercise 16

To show:

Consider the majority classifier

$$g_n(x, D_n) = \begin{cases} 1 & \text{if } \sum_{i=1}^n Y_i \geq n/2 \\ 0 & \text{otherwise} \end{cases}$$

(Thus, g_n ignores x and the x_i's) Assume that n is odd. What is the expected risk $\mathbb{E}R(g_n) = \boldsymbol{P}\{g_n(X) \neq y\}$ of this classifier? Study the performance of the leave-one-out error estimate. Show that for some distributions $Var(R_n^{(D)}(g_n)) \geq c/\sqrt{n}$ for some constant c. Hint: Strange things happen when the number of 0's and 1's is about the same in the data.

Solution:

Start with the definition of the risk given in the exercise:

$$\begin{aligned} \mathbb{E}\big[R(g_n)\big] &= \mathbb{P}\big(g_n(X) \neq y\big) \\ &= \mathbb{P}\big(g_n(X) = 1, Y = 0\big) + \mathbb{P}\big(g_n(X) = 0, Y = 1\big) \\ &= \mathbb{P}\big(g_n(X) = 1\big)\mathbb{P}\big(Y = 0\big) + \mathbb{P}\big(g_n(X) = 0\big)\mathbb{P}\big(Y = 1\big) \end{aligned}$$

Let $N = \sum_{i=1}^n Y_i$ and use this for the probability of classifying 1 or 0. Furthermore let, $p = \mathbb{P}(Y = 0)$ and $(1 - p) = \mathbb{P}(Y = 1)$. Applying this the risk can be expressed as:

$$= \mathbb{P}\bigg(N \geq \frac{n}{2}\bigg)p + \mathbb{P}\bigg(N < \frac{n}{2}\bigg)(1 - p)$$

Performance issues:

In case of perfomance consider the following case; Assume that there are 100 zeros and 101 ones in the data set. Moreover, assume, that we rule in favor of the minority class in case of a tie. However, with the leave-one-out error estimator we misclassify all the time and the risk becomes 1.

Finally for the last part of the exercise we are interested in a lower bound for the variance:

$$\begin{aligned} Var\big(R_n^{(D)}(g_n)\big) &= \mathbb{E}\bigg[\bigg(R_n^{(D)}(g_n) - \mathbb{E}\big[R_n^{(D)}(g_n)\big]\bigg)^2 \\ &= \mathbb{E}\bigg[\bigg(R_n^{(D)}(g_n) - \frac{1}{2}\bigg)^2\bigg] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{P}\bigg(R_n^{(D)}(g_n) = i\bigg)\bigg(i - \frac{1}{2}\bigg)^2 \\ &\geq \frac{1}{n^2}\mathbb{P}\bigg(Bin\bigg(n, \frac{1}{2}\bigg) = \frac{n+1}{2}\bigg)\bigg(\frac{n+1}{2} - \frac{1}{2}\bigg) \end{aligned}$$

Now assume that $n \approx n+1$. Furthermore, using Sterlings approximation to obtain the final result:

$$\begin{aligned}
\mathbb{P}\left(Bin\left(n, \frac{1}{2}\right) = \frac{n}{2}\right) &= \binom{n}{\frac{n}{2}} 2^{-n} \\
&\geq \frac{2^{n-1+\frac{1}{2}-n}}{\sqrt{n}} = \frac{1}{\sqrt{2}\sqrt{n}} \\
&= c * \frac{1}{\sqrt{n}}
\end{aligned}$$