1: PAC Learning

Rule 1: You are free to combine any of the parts as they are.

Rule 2: You may also cut any of the parts into two distinct pieces before using them.

(1a)

Given N parts, each product that can be made out of these parts is a distinct hypothesis h in the hypothesis space H. From *Rule 1*, a worker can choose to include or not include any of the parts in a product. This can be viewed as a monotone conjunction as a product is defined by choosing to include or not include each of the N parts. There exists 2^N possible products as there are two choices for each of the N parts. We will not consider the product constructed by using none of the parts.

$\boxed{|H| = 2^N - 1}$

(1b)

The experienced worker now creates a product using *Rule 1* and *Rule 2*. There are now four choices that can be made for each of the parts: don't include it, include it, cut the part and use the first half or cut the part and use the second half. A product is now defined as making four choices for each of the N parts. Thus there are 4^N possible products. We will not consider the product constructed by using none of the parts.

$\boxed{|H| = 4^N - 1}$

(1c)

By applying the principles of Occams's Razor we can make a statement about the number of required examples the robot will have to see to have an error of 0.01 with probability 99% on products with 6 available parts.

Given a hypothesis space H, we can say with probability $1 - \delta$, a hypothesis $h \in H$, that is consistent with a training set of size m, will have an error $< \epsilon$ on future examples if

$$m > \frac{1}{\epsilon}(ln(|H|) + ln\frac{1}{\delta})$$

We want an error rate of $\epsilon = 0.01$ with probability $1 - \delta = 0.99$ with a $|H| = 4^6 = 4,096$.

$$m > \frac{1}{0.01}(ln(4,095) + ln\frac{1}{0.01})$$
$$m > 1,292.27$$

The robot will have to see at least $1,293$ examples to guarantee a 0.01 error with probability 99% if there are 6 available parts. We round up as the number of required examples must be an integer value and rounding down would not satisfy the equality.

$\boxed{\text{at least 1,293 examples}}$

(2)

2: VC Dimensions

Shatter: A set of examples S is *shattered* by a set of functions H if for every partition of the examples in S into positive and negative examples there is a function in H that gives exactly these labels to the examples.

VC Dimension: The *VC Dimension* of hypothesis space H over instance space X is the size of the largest finite subset of X that is shattered by H.

(1) We want to prove that a finite hypothesis space $\mathcal{C}$ has a VC dimension at most $log_2|\mathcal{C}|$. That is, $VC(\mathcal{C}) \leq log_2|\mathcal{C}|$.

Proof by contradiction: Assume the opposite that $VC(\mathcal{C}) > log_2|C|$ is true.

Take a finite instance space X of size d. There exists 2^d (the number of possible binary vectors of length d) ways to partition X. Thus $|\mathcal{C}| = 2^d$ as we have 2^d hypothesis functions.

Given that X is of size d, it holds that $0 \leq VC(C) \leq d$ as $\mathcal{C}$ can shatter **at most** d points (the size of the instance space). Visiting the initial assumption:

$$VC(\mathcal{C}) > \log_2 |C|$$

$$VC(\mathcal{C}) > d$$

Arriving at the contradiction $VC(\mathcal{C}) \leq d$ and $VC(\mathcal{C}) > d$. Thus proving $VC(\mathcal{C}) \leq log_2|\mathcal{C}|$.

(2a)

(2b)

(3) The proof for the VC dimension of $\mathcal{H}$ involves two parts. First, we must show that there *exists* any subset of size d that can be shattered (this proves $VC(\mathcal{H}) \geq d$). Second, we must show that *no subset* of size d can be shattered by $\mathcal{H}$ (this proves $VC(\mathcal{H}) < d$). The result of these two bounding inequalities proves $VC(\mathcal{H}) = d$.

Proof (1) To prove $VC(\mathcal{H}) \geq 4$ we must give one example of points $x_1, x_2, x_3, x_4 \in \mathbb{R}$ that can be shattered by $h \in \mathcal{H}$. There are 16 possible labelings of the four the points and we must show there is a $h \in \mathcal{H}$ that satisfies all of them. The figure below expresses all the possible labelings, excluding labelings that are symmetric to provided labelings to avoid an excessiveness of figures.

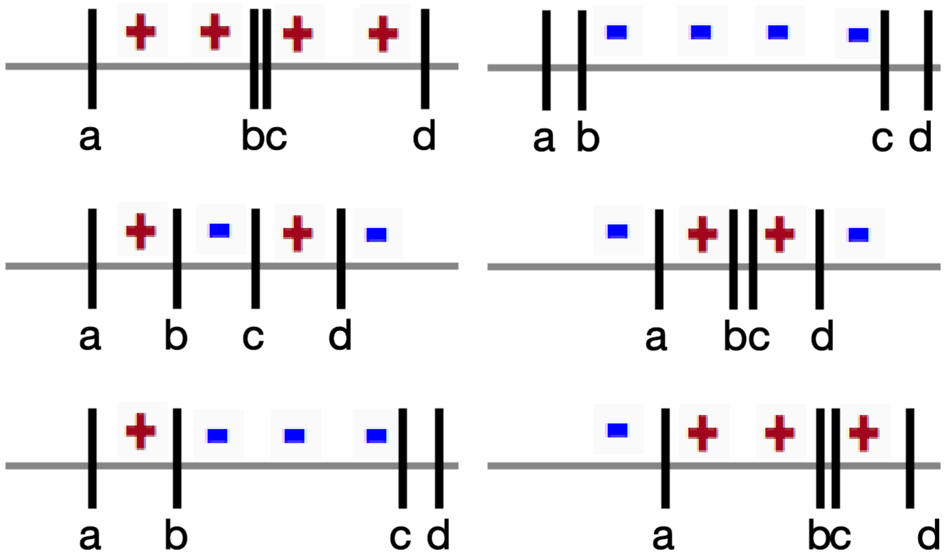

Figure 1: Labeling of four example points and a $h \in \mathcal{H}$ that shatters them.

Proving $VC(\mathcal{H}) \geq 4$. (2) To show $VC(\mathcal{H}) < 5$, we must prove there doesn't exist a set of **any** five points that can be shattered by $\mathcal{H}$. Given any five unique points $x_i \in \mathbb{R}$, there exists a labeling s.t. $\mathcal{H}$ cannot shatter the set of five points.

Figure 2: Labeling of five points that can not be shattered by $\mathcal{H}$

There exists the relationship $x_i < x_j < x_k < x_l < x_m$. That is, choosing any five real numbers, they can be arranged to satisfy the inequality. Starting with x_i, label it positive and alternate labelings moving along the five points in ascending order. There is no way to shatter these labeled five points with $\mathcal{H}$. Thus we have proven the $VC(\mathcal{H}) = 4$.

$\boxed{VC(\mathcal{H}) = 4}$

(4)

(5) Let two hypothesis classes H_1 and H_2 satisfy $H_1 \subseteq H_2$. Prove: $VC(H_1) \leq VC(H_2)$.

Proof by contradiction: Assume the opposite that $VC(H_1) > VC(H_2)$ is true.

Let X be a finite instance space and $VC(H_1) = d$. That is, a set of examples S of size d are shattered by H_1. Meaning for every partition of the examples in S into positive and negative examples there is a function $h \in H_1$ that gives exactly these labels to the examples. More so, S is the largest finite subset of X that is shattered by H_1.

We know that H_1 and H_2 satisfy $H_1 \subseteq H_2$. We know h (the hypothesis that correctly labels all partitions of d points) is also $h \in H_2$. This gives us $VC(H_2)$ is **at least** d (could be greater than d as $h_2 \in H_2$ could exist that shatters a larger subset of points). Thus we have shown $VC(H_1) > VC(H_2)$ is a contradiction, proving $VC(H_1) \leq VC(H_2)$.

3: AdaBoost

We can calculate D_2 given h_a, α_1, and D_1 for each example in the training set.

$$\boxed{D_{t+1}(i) = \frac{D_t(i)}{Z_t} * exp(-\alpha_t * y_i h_i(x_i))} \tag{1}$$

$$D_2(i) = \frac{D_1(i)}{Z_1} * exp(-\alpha_1 * y_i h_a(x_i))$$

$$D_2(1) = \frac{1}{2},\ D_2(2) = \frac{1}{6},\ D_2(3) = \frac{1}{6},\ D_2(4) = \frac{1}{6}$$

$x = [x_1, x_2]$	y_i	$h_a(x)$	D_1	$D_1(i) y_i h_t(x_i)$	D_2
[1,1]	-1	1	1/4	-1/4	1/2
[1,-1]	1	1	1/4	1/4	1/6
[-1,-1]	-1	-1	1/4	1/4	1/6
[-1,1]	-1	-1	1/4	1/4	1/6

Table 1: $h_a(x) = sgn(x_1),\ \epsilon_1 = 1/4,\ \alpha_1 = \frac{\ln 3}{2},\ Z_1 = \frac{\sqrt{3}}{2}$

Choosing $\mathbf{h_d(x) = -sgn(x_2)}$ for iteration 2. We calculate the weighted classification error to determine if it's better than chance.

$$\boxed{\epsilon_t = \frac{1}{2} - \frac{1}{2}(\sum_{i=1}^{m} D_t(i) * y_i h_i(i))} \tag{2}$$

$$\epsilon_2 = \frac{1}{2} - \frac{1}{2}(\sum_{i=1}^{m} D_2(i) * y_i h_d(i)) = \frac{1}{6}$$

Given ϵ_2, we calculate α_2 which is the weight the current hypothesis has on the final hypothesis.

$$\boxed{\alpha_t = \frac{1}{2} \ln(\frac{1 - \epsilon_t}{\epsilon_t})} \tag{3}$$

$$\alpha_2 = \frac{1}{2} \ln(\frac{1 - \epsilon_2}{\epsilon_2}) = \frac{1}{2} \ln(\frac{1 - \frac{1}{6}}{\frac{1}{6}}) = \frac{\ln 5}{2}$$

We calculate Z_2 which is a normalization constant to ensure all of the D_3 weights add up to 1.

$$\boxed{Z_t = \sum_{i=1}^{m} D_t(i) * exp(-\alpha_t * y_i h_i(x_i))} \tag{4}$$

$$Z_2 = \sum_{i=1}^{m} D_2(i) * exp(-\alpha_2 * y_i h_d(x_i)) = \frac{\sqrt{5}}{3}$$

Finally we calculate a new weight D_3 for each example in the training set.

$$D_3(i) = \frac{D_2(i)}{Z_2} * exp(-\alpha_2 * y_i h_d(x_i))$$

$$D_3(1) = \frac{3}{10},\ D_3(2) = \frac{1}{10},\ D_3(3) = \frac{1}{2},\ D_3(4) = \frac{1}{10}$$

The results for iteration 2 using hypothesis $h_d(x)$ are recorded in the table below.

$x = [x_1, x_2]$	y_i	$h_d(x)$	D_2	$D_1(i)y_ih_t(x_i)$	D_3
[1,1]	-1	-1	1/2	1/2	3/10
[1,-1]	1	1	1/6	1/6	1/10
[-1,-1]	-1	1	1/6	-1/6	1/2
[-1,1]	-1	-1	1/6	1/6	1/10

Table 2: $h_d(x) = -sgn(x_2),\ \epsilon_2 = 1/6,\ \alpha_2 = \frac{\ln 5}{2},\ Z_2 = \frac{\sqrt{5}}{3}$

Choosing $\mathbf{h_b(x) = sgn(x_1 - 2)}$ for iteration 3. We calculate the weighted classification error to determine if it's better than chance.

$$\epsilon_3 = \frac{1}{2} - \frac{1}{2}(\sum_{i=1}^{m} D_3(i) * y_i h_b(i)) = \frac{1}{10}$$

Given ϵ_3, we calculate α_3 which is the weight the current hypothesis has on the final hypothesis.

$$\alpha_3 = \frac{1}{2}\ln(\frac{1-\epsilon_3}{\epsilon_3}) = \frac{1}{2}\ln(\frac{1-\frac{1}{10}}{\frac{1}{10}}) = \frac{\ln 9}{2}$$

We calculate Z_3 which is a normalization constant to ensure all of the D_4 weights add up to 1.

$$Z_3 = \sum_{i=1}^{m} D_3(i) * exp(-\alpha_3 * y_i h_b(x_i)) = \frac{3}{5}$$

Finally we calculate a new weight D_4 for each example in the training set.

$$D_4(i) = \frac{D_3(i)}{Z_3} * exp(-\alpha_3 * y_i h_b(x_i))$$

$$D_4(1) = \frac{1}{6},\ D_4(2) = \frac{1}{2},\ D_4(3) = \frac{5}{18},\ D_4(4) = \frac{1}{18}$$

The results for iteration 3 using hypothesis $h_b(x)$ are recorded in the table below.

$x = [x_1, x_2]$	y_i	$h_b(x)$	D_3	$D_1(i)y_ih_t(x_i)$	D_4
[1,1]	-1	-1	3/10	3/10	1/6
[1,-1]	1	-1	1/10	-1/10	1/2
[-1,-1]	-1	-1	5/10	1/2	5/18
[-1,1]	-1	-1	1/10	1/10	1/18

Table 3: $h_b(x) = sgn(x_1 - 2),\ \epsilon_3 = 1/10,\ \alpha_3 = \frac{\ln 9}{2},\ Z_3 = \frac{3}{5}$

Choosing $\mathbf{h_c(x) = -sgn(x_1)}$ for iteration 4. We calculate the weighted classification error to determine if it's better than chance.

The weighted classification error ϵ_4 for $h_c(x)$ is not better than chance.

$$\epsilon_4 = \frac{1}{2} - \frac{1}{2}(\sum_{i=1}^{m} D_4(i) * y_ih_c(i)) = \frac{5}{6}$$

The classification error ϵ_4 is not better than chance. As a result hypothesis $h_c(x)$ is not considered.

Finally, we consider the final hypothesis $H_{final}(x)$ which takes a weighted average of the classification of $h_a(x)$, $h_b(x)$, and $h_d(x)$.

$$\boxed{H_{final}(x) = sgn(\sum_t \alpha_t h_t(x))} \tag{5}$$

Using $H_{final}(x)$ we classify each example from the training set.

$$H_{final}(1) = sgn(\frac{\ln 3}{2}(1) + \frac{\ln 5}{2}(-1) + \frac{\ln 9}{2}(-1)) = -1$$
$$H_{final}(2) = sgn(\frac{\ln 3}{2}(1) + \frac{\ln 5}{2}(1) + \frac{\ln 9}{2}(-1)) = 1$$
$$H_{final}(3) = sgn(\frac{\ln 3}{2}(-1) + \frac{\ln 5}{2}(1) + \frac{\ln 9}{2}(-1)) = -1$$
$$H_{final}(4) = sgn(\frac{\ln 3}{2}(-1) + \frac{\ln 5}{2}(-1) + \frac{\ln 9}{2}(-1)) = -1$$

$x = [x_1, x_2]$	y_i	$H_{final}(x)$
[1,1]	-1	-1
[1,-1]	1	1
[-1,-1]	-1	-1
[-1,1]	-1	-1

Table 4: Classification of the training set by $H_{final}(x)$

Using $H_{final}(x)$ we have properly classified all the examples in the training set.