Introduction to Machine Learning (67577)

Exercise 4 - PAC & Ensemble Methods

Dor Roter
208772251

June 1, 2021

1 PAC Learnability

1. Let us show the equivalence holds in both directions:
<u>*(a)* ⇒ *(b)*</u>**:** it holds that there exists $m : (0,1)^2 \to \mathbb{N}$ such as:

$$\forall \epsilon, \delta > 0,\ \forall m \geq m(\epsilon, \delta) \qquad 1 - \delta \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) \leq \epsilon\right] = 1 - \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) > \epsilon\right]$$
$$\Leftrightarrow \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) > \epsilon\right] \leq \delta$$

let $\epsilon > 0$.
since $L_{\mathcal{D}}\left(\mathcal{A}(S)\right) \in [0,1]$ thus $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\right] \leq 1$ for any $m \in \mathbb{N}$. As such, assuming $\epsilon \in (0,1)$ it holds that $\forall m \geq m(\epsilon, \epsilon) = N$:

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\right] &\underset{L_{\mathcal{D}}(\cdot)\in[0,1]}{=} \int_0^1 x \cdot \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) = x\right] dx = \\ &= \int_0^\epsilon x \cdot \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) = x\right] dx + \int_\epsilon^1 x \cdot \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) = x\right] dx \leq \\ &\leq \int_0^\epsilon \epsilon \cdot \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) = x\right] dx + \int_\epsilon^1 \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) = x\right] dx\end{aligned}$$

now, since from monotinicity of the probability function,

$$\forall x \in (0, \epsilon]\ \left\{L_{\mathcal{D}}\left(\mathcal{A}(S)\right) = x\right\} \subseteq \left\{L_{\mathcal{D}}\left(\mathcal{A}(S)\right) \leq \epsilon\right\}$$
$$\forall x \in [\epsilon, 1)\ \left\{L_{\mathcal{D}}\left(\mathcal{A}(S)\right) = x\right\} \subseteq \left\{L_{\mathcal{D}}\left(\mathcal{A}(S)\right) > \epsilon\right\}$$

since the integral is not effected by a single point chage we are allowed to exclude $\frac{\epsilon}{2}$, without effecting the value of the integral:

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\right] &\underset{monotonicity}{\leq} \int_0^\epsilon \epsilon \cdot \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) \leq \epsilon\right] dx + \int_\epsilon^1 \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right) > \epsilon\right] dx \\ &\underset{monotonicity}{\leq} \int_0^\epsilon \epsilon \cdot 1 dx + \int_\epsilon^1 \epsilon dx = \epsilon^2 + (1-\epsilon)\cdot\epsilon = \epsilon\end{aligned}$$

and thus directly by limit's defenition there exists $N \in \mathbb{N}$ for any $\epsilon > 0$, such that $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\right] < \epsilon$ and so: $\underset{m\to\infty}{lim}\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\right] = 0$.

$(b) \Rightarrow (a)$**:** let $\epsilon, \delta > 0$.
$\underset{m\to\infty}{lim} \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right)\right] = 0$ and thus for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for any $N < m \in \mathbb{N}$ $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right)\right] < \epsilon$, more specifically, there exists $N \in \mathbb{N}$ such that for each $N < m \in \mathbb{N}$: $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right)\right] < \epsilon \cdot \delta$.
now, as $L_\mathcal{D}$ is bounded between zero and one, it is surely non-negative, and thus using markov's inequality we note that for any such $N < m \in \mathbb{N}$:

$$\underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right) > \epsilon\right] \underset{monotonicity}{\leq} \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right) \geq \epsilon\right] \underset{markov}{\leq} \frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right)\right]}{\epsilon} \leq \frac{\epsilon\cdot\delta}{\epsilon} = \delta$$

and thus denoting $m\left(\epsilon,\delta\right) = N{+}1$ we achieve the required:

$$\forall\epsilon,\delta > 0\ \forall m \geq m\left(\epsilon,\delta\right) \underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right) > \epsilon\right] \leq \delta \Leftrightarrow 1-\delta \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}\left(\mathcal{A}(S)\right) \leq \epsilon\right]$$

2. Firstly, let us describe a proposed ERM algorithm for learning of $\mathcal{H}$.

Algorithm 1 Learner for $\mathcal{H}$-provided a training set $\mathcal{S} = \{(x,y)\}^m$

```
(a) find $\hat{r} = \underset{(x,y)\in\mathcal{S},\, y=1}{max} ||x||_2$
(b) return $h_{\hat{r}} : \mathcal{X} \to \mathcal{Y}$ defined by $h_{\hat{r}}(x) = \mathbb{1}[||x||_2 \leq \hat{r}]$
```

Let us analize the proposed algorithm sample complexity.
Firstly, denoting by r the true radius of the circle for which h_r has zero generalization error, it is clear any circle portrayed by my proposed algorithm, must be contained within h_r's, as my alogirthm elects the maximal radius circle that contains all $1's$ in the training set, while h_r must contain all $1's$ in $\mathcal{D}$, and so our proposed algorithm may preform only false negative (false 0) erros.
Therefore, let us denote $T = \{x \in \mathcal{X}\,|\hat{r} < ||x||_2 \leq r\}$, and so T includes all $x \in \mathcal{X}$ that our algorithm would tag as 0's which are in-fact 1's.
Let $\epsilon, \delta > 0$. let T' be a ring containing/conatins T such as that $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in T') = \epsilon$. as such, the probability of a given point $x \in \mathcal{X}$ to be misclassifed by $h_{\hat{r}}$ is bounded from above by the probability that $x \in T'$, and therefore by assuming iid, the probability of drawing m independent smaples from $\mathcal{D}$ which all misses T' is at most $(1-\epsilon)^m$.
Finally it holds that:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}\left(h_{\hat{r}}\right) \leq \epsilon\right] \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}\left(h_{\hat{r}}\right) \leq 0\right] \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

and thus by choosing m such that $e^{-\epsilon m} \leq \delta$ will result in $1-\delta$ probability over m random samples that the weight of the error is at most ϵ.

$$e^{-\epsilon m} \leq \delta \Leftrightarrow \epsilon m \leq ln\left(1/\delta\right) \Leftrightarrow m \leq \frac{ln\left(1/\delta\right)}{\epsilon}$$

Therefore $m_\mathcal{H}\left(\epsilon,\delta\right) \leq \frac{ln(1/\delta)}{\epsilon}$. □

2 VC-Dimension

3. let $\mathcal{H} = \{h_1, \dots, h_N\}$,for some $N \in \mathbb{N} \cup \{0\}$.
   denoting $VCdim\left(\mathcal{H}\right) = d$, it holds that there exists a set if size d that is shattered by $\mathcal{H}$. As such, there are at least 2^d different hypothersis in $\mathcal{H}$ and thus

   $$2^d \leq |\mathcal{H}| \Leftrightarrow d \leq log_2\left(|\mathcal{H}|\right) \qquad \square$$

4. Firstly, since $\mathcal{H}_{parity} = \{h_I \,|I \subseteq [n]\}$,$|\mathcal{H}_{parity}| = |\{I\,|I \subseteq [n]\}| = 2^n$, following this, by question 3 it holds that $VCdim\left(\mathcal{H}_{parity}\right) \leq n$
   Now, let us show a group $C \subset \mathcal{X}$ such as that $|C| = n$, and C is shattered by $\mathcal{H}_{parity}$.
   Let $C = \{e_1, \dots, e_n\}$ - (where e_i is the i-th unit vector), therefore we note that for any possible labeling $y = \{y_1, \dots, y_n\}$, defining $I \subset [n]$ as $I = \{i\,|y_i = 1\}$ we get h_I such as that its restriction over the set C results in

   $$\forall c \in C\ \exists i \in [n]\ \ s.t\ h_I\left(c\right) = h_I\left(e_i\right) = \left(\sum_{j\in I} e_i^j\right) mod2 = \mathbb{1}\left[i = j\right] = \mathbb{1}\left[y_i = 1\right] = y_i$$

   and thus for any labeling over C we provided a $h \in \mathcal{H}_{parity}$ resulting in the provided label over C, therefore $|\mathcal{H}_c| = |\text{possible labels}| = 2^{|C|}$i.e C is shattered by $\mathcal{H}_{parity}$.
   Summing it up, n is an upper bound over the shattered groups size which is achieved and thus $VCdim\left(\mathcal{H}_{parity}\right) = n$. $\square$

5. Firstly we note $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, and $\mathcal{H}_{k-interval} = \left\{h : \exists A = \bigcup_{i=1}^{k} [a_i, b_i]\ \ s.t\ h = h_A\right\}$.
   Let $C \subset \mathcal{X}$ such as $C = \{x_1, \dots, x_{2k}\}$, where without limit of generality $x_1 < \dots < x_{2k+1}$.
   Let us show that for any labeling $y = \{y_1, \dots, y_{2k}\} \in \{0, 1\}^k$ over C, we manage to find a $h \in \mathcal{H}_C$ so that $h(x_i) = y_i$.
   We will be constructing $A = \bigcup_{i=1}^{k} [a_i, b_i]$ where the set $(a_i, b_i)_{i=1}^{k}$ is constructed in the following fashion: the i-th interval defined by (a_i, b_i) is bounding the i-th consecutive block of 1's in the labels y, and any other number once there are no more such consecutive blocks of 1's (just 0's for $y_j, ..., y_{2k}$).
   From $\mathbb{R}$ density we have $\aleph$ options for each one of a_i and b_i in between each block and thus the provided construction if indeed plausable.
   Furthermore, the provided constructions provides us with $A = \bigcup_{i=1}^{k} [a_i, b_i]$ where for every $x_i \in C$: $x_i \in A \Leftrightarrow y_i = 1$, and thus $h_A \in \mathcal{H}_{k-intervals}$ enables the required labeling.
   Since there are $2^{|C|}$ different labels, with each label fitting a different function $h_A \in \mathcal{H}_{k-intervals}$ as constructed above, $|\mathcal{H}_C| = 2^{|C|}$ - C is shattered by $\mathcal{H}_{k-intervals} \Rightarrow VCdim\left(\mathcal{H}_{k-intervals}\right) \geq k$.
   On the other hand, for any $C \subset \mathcal{X}$ such as $C = \{x_1, \dots, x_{2k+1}\}$, and the labeling $y = (1, 0, 1, \dots, 0, 1)$, there exist $k + 1$ independandt intervelas with $y_i = 1$, and as such, any function defined by k-intervals, by the pigeonhole principle, would result in at least two different points $y_i, y_j = 1$, such that both need to be placed within the same interval (as any $y_i = 1$ requires x_i to be bounded in one of the k intervals), therefore, by y's defenition, there exist $i < l < j$, such that $y_l = 0$, but as it is placed within the same interval as i, j it must have a labeling of 1 by $\mathcal{H}$defenition, thus the labeling y is unachievable by our hypothesis class, and C is in-fact not shattered by $\mathcal{H}_{k-intervals}$.
   $\Rightarrow 2k \leq VCdim\left(\mathcal{H}_{k-intervals}\right) < 2k + 1 \Leftrightarrow VCdim\left(\mathcal{H}_{k-intervals}\right) = 2k$. $\square$

6. Let $\mathcal{H}_{con} = \left\{h(x) = \bigwedge_{i\in I} \hat{x}^i \left|I \subseteq [d]\right),\ \hat{x}^i \in \left\{x^i, \overline{x}^i\right\}\right\}$.
   Firstly let us show there exists $C \subset \mathcal{X}$ so that $|C| = d$ and is shattered by $\mathcal{H}_{con}$.

Let $C = \{e_1, \dots, e_d\}$ where e_i is the i-th unit vector, and let $y = \{y_1, \dots, y_d\} \in \{0,1\}^d$ be a labeling for each $x \in C$.
Selecting $I = \{i \,|y_i = 0\}$ and for each $i \in I$ $\hat{x}^i = \overline{x}^i$, we are defining an $h \in \mathcal{H}_{con}$ so that:

$$\forall e_i \in C\; h(e_j) = \bigwedge_{i \in I} \hat{e_j}^i = \bigwedge_{i \in I} \overline{e}_j^i = \mathbb{1}\left[\forall i \in I\, e_j^i = 0\right] = \mathbb{1}\left[j \notin I\right] = \mathbb{1}\left[y_i = 1\right]$$

Therefore, we have also found $h \in \mathcal{H}_C$ for any labeling over C and thus C is shattered by $\mathcal{H}_{con}$.
$\Rightarrow VCdim\left(\mathcal{H}_{con}\right) \geq d.$

On the other hand, assuming by way of contradiction that there is a $C = \{x_1, \dots, x_{d+1}\} \subset \mathcal{X}$ so that C is shattered by $\mathcal{H}_{con}$, it holds that there exist for any labeling $y = \{y_1, \dots, y_{d+1}\} \in \{0,1\}^{d+1}$ over C $h \in \mathcal{H}_{con}$ so that for $\forall x_i \in C$ $h(x_i) = y_i$.
Therefore each of the following $d+1$ labelings $y(1), \dots, y(d+1)$ where:

$$\forall j \in [d+1]\;\; y(i)_j = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

Must have a corresponding $h_i \in \mathcal{H}_{con}$ so that its restriction over C results in $y(i)$ vector.
As such, by defenition each $h_i \in \mathcal{H}_{con}$ is constructed by some literals conjugation and must include some literal $\hat{x}^k \in \left\{x^k, \overline{x}^k\right\}$ $(k \in [d])$ such that for every $x_j \in C$ $(j \in [d+1])$ $\hat{x}_j^k = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$ (or else it would be impossible for h_i to generate the $y(i)$ labeling).
By the pigeonhole principle, as there are $d+1$ different functions but $k \in [d] \to d$ different features, there must be two different functions h_{l_1}, h_{l_2} $(l_1 \neq l_2)$, so that both use x^k (either negated or not) in their conjugation.
Therefore, if x^k or its negation is used for **both** h_{l_1} and h_{l_2},it must hold that $y(l_1)_{l_2} = y(l_2)_{l_2} = 0$ and $y(l_1)_{l_1} = y(l_2)_{l_1} = 0$, which is a contradiction by the labels defenition, as $l_1 \neq l_2$.
If, on the other hand, w.l.o.g h_{l_1} uses x^k and h_{l_2} uses $\overline{x}^k$, it must hold that for all $l_1 \neq j \in [d]$: $y(l_1)_j = 1$ but then also $y(l_2)_j = 0$, once again contradicting the labeling defention.
Thus it is impossible to achieve those specific $d+1$ different labels by d literals conjugation, and $\mathcal{H}_{con}$ cannot shatter any subset of $\mathcal{X}$ of size $d+1$ or greater $\Rightarrow VCdim\left(\mathcal{H}_{con}\right) < d+1$.
$\Rightarrow VCdim\left(\mathcal{H}_{con}\right) = d$ $\square$

3 Agnostic-PAC

7. Let $\mathcal{D}, \epsilon, \delta$.
   Since $\mathcal{H}$ has the uniform convergence property with function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$, then for every $m \geq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$:

$$\mathcal{D}^m\left(\left\{S \in (\mathcal{X} \times \mathcal{Y})^m \,|S \text{ is } \epsilon\text{-representative}\right\}\right) \geq 1 - \delta$$

Assuming S is a $\frac{\epsilon}{2}$-representative of $\mathcal{D}$, and let h_S be any output on $ERM_{\mathcal{H}}(S)$ $\left(h_S \in \underset{h\in\mathcal{H}}{argmin} L_S(h)\right)$, we have shown in recitation that $L_{\mathcal{D}}(h_S) \leq \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}(h) + \epsilon$ and thus

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_S) \leq \underset{h\in\mathcal{H}}{min}L_{\mathcal{D}}(h)+\epsilon\right] \underset{\substack{motonicity \\ \uparrow \\ \frac{\epsilon}{2} - representative \Rightarrow L_{\mathcal{D}}(h_S) \leq \underset{h\in\mathcal{H}}{min}L_{\mathcal{D}}(h)+\epsilon}}{\geq} \mathcal{D}^m\left(\left\{S \in (\mathcal{X}\times\mathcal{Y})^m \middle| S \text{ is } \frac{\epsilon}{2}\text{-representative}\right\}\right) \underset{\substack{uniform \\ convergence}}{\geq} 1-\delta$$

Therefore, directly by defenition it holds that for some $m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2},\delta\right)$ $\mathcal{H}$ is Agnostic-PAC learnable. □

8. Let us refute the claim by show of a counter example.
Let $\mathcal{H} = \{\pm 1\}^{\mathcal{X}}$ for any infinite $\mathcal{X}$. Therefore, clearly $\mathcal{H}$ is not PAC learnable directly by the "no free meals" theorem, yet, for any $\mathcal{D}$ - a given disturbution over Z, for the algorithm $\mathcal{A}_{\mathcal{D}} = \underset{h\in\mathcal{H}}{argmin} L_{\mathcal{D}}(h)$, it holds that for any $m \in \mathbb{N}$, $\epsilon > 0$ and $S \overset{iid}{\sim} \mathcal{D}^m$:

$$L_{\mathcal{D}}(\mathcal{A}_{\mathcal{D}}(S)) = L_{\mathcal{D}}\left(\underset{h\in\mathcal{H}}{argmin} L_{\mathcal{D}}(h)\right) = min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) \leq min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

Thus the provided claim's conditions hold for $\mathcal{H}$, yet it is **not** PAC learnable.
Assuming by way of contradiction that $\mathcal{H}$ is agnostic-PAC learnable, then there exist a learning algortihm $\mathcal{A}$, and $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and any distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ when running $\mathcal{A}$ on $m \geq m_{\mathcal{H}}(\epsilon,\delta)$: $\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S)) \leq \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}(h) + \epsilon\right] \geq 1-\delta$.
Electing $h^* \in \mathcal{H}$ and $\mathcal{D}$ to be a distribution over $\mathcal{X}$, we denote $\mathcal{D}'$, a distribution over Z which samples $x \sim \mathcal{D}$ and then returns the pair $(x, h^*(x))$ hence - $\mathcal{D}'[\{(x,y)\,|h^*(x) = y\}] = 1$, by defenition, $\underset{h\in\mathcal{H}}{argmin} L_{\mathcal{D}'}(h) = h^* \in \mathcal{H}$ (as $L_{\mathcal{D}'}(h^*) = 0$), and so it holds that for any $m \geq m_{\mathcal{H}}(\epsilon,\delta)$ and $S \overset{iid}{\sim} \mathcal{D}'^m$:

$$\mathbb{P}_{S\sim\mathcal{D}'^m}\left[L_{\mathcal{D}}(\mathcal{A}(S)) \leq \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}'}(h) + \epsilon\right] = \mathbb{P}_{S\sim\mathcal{D}'^m}\left[L_{\mathcal{D}'}(\mathcal{A}(S)) \leq 0 + \epsilon\right] \geq 1-\delta$$

Now, let us note that by defenition, as we defined our loss function to be a 0-1 loss, $L_{\mathcal{D}'} = L_{\mathcal{D},h^*}, S \sim \mathcal{D}'^m \Leftrightarrow S \sim \mathcal{D}^m$ and thus $\mathcal{H}$ is in-fact PAC-learnable:

$$\mathbb{P}_{S\sim\mathcal{D}'^m}\left[L_{\mathcal{D}'}(\mathcal{A}(S)) \leq 0 + \epsilon\right] \geq 1-\delta \Leftrightarrow \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D},h^*}(\mathcal{A}(S)) \leq \epsilon\right] \geq 1-\delta$$

By contradiction to the "no free meals" theorem, therefore $\mathcal{H}$ is not agnostic-PAC-learnable. □

4 Monotonicity

9. Let $\mathcal{H}, \mathcal{D}, m_{\mathcal{H}}$, and $\epsilon, \epsilon_1, \epsilon_2, \delta, \delta_1, \delta_2 \in (0,1)$ such that without limit of generality $\epsilon_1 \leq \epsilon_2$ and $\delta_1 \leq \delta_2$.
Firstly,

$$\epsilon_1 \leq \epsilon_2 \Rightarrow \{L_{\mathcal{D}}(\mathcal{A}(S)) \leq \epsilon_1\} \subseteq \{L_{\mathcal{D}}(\mathcal{A}(S)) \leq \epsilon_2\}$$
$$\delta_1 \leq \delta_2 \Rightarrow \underset{S\sim\mathcal{D}^m}{\mathbb{P}}[L_{\mathcal{D}}(\mathcal{A}(S)) \leq \epsilon] \geq 1-\delta_1 \geq 1-\delta_2$$

Assuming by way of contradiction that $m_{\mathcal{H}}\left(\epsilon_1, \delta\right) < m_{\mathcal{H}}\left(\epsilon_2, \delta\right)$, as $\mathcal{H}$ is PAC learnable it holds that there exists a learning algorithm $\mathcal{A}$ such that for any $m \geq m_{\mathcal{H}}\left(\epsilon_1, \delta\right)$:

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leq \epsilon_2\right] \underset{monotonicity}{\geq} \underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leq \epsilon_1\right] \underset{PAC}{\geq} 1-\delta$$

Thus more specifically for any $m_{\mathcal{H}}\left(\epsilon_1, \delta\right) \leq m < m_{\mathcal{H}}\left(\epsilon_2, \delta\right)$ the above must hold, in contradiction to $m_{\mathcal{H}}\left(\epsilon_2, \delta\right)$ minimality $\Rightarrow m_{\mathcal{H}}\left(\epsilon_1, \delta\right) \geq m_{\mathcal{H}}\left(\epsilon_2, \delta\right)$.

Assuming by way of contradiction that $m_{\mathcal{H}}\left(\epsilon, \delta_1\right) < m_{\mathcal{H}}\left(\epsilon, \delta_2\right)$, as $\mathcal{H}$ is PAC learnable it holds that there exists a learning algorithm $\mathcal{A}$ such that for any $m \geq m_{\mathcal{H}}\left(\epsilon, \delta_1\right)$:

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leq \epsilon\right] \geq 1-\delta_1 \geq 1-\delta_2$$

Thus more specifically for any $m_{\mathcal{H}}\left(\epsilon, \delta_1\right) \leq m < m_{\mathcal{H}}\left(\epsilon, \delta_2\right)$ the above must hold, in contradiction to $m_{\mathcal{H}}\left(\epsilon, \delta_2\right)$ minimality $\Rightarrow m_{\mathcal{H}}\left(\epsilon, \delta_1\right) \geq m_{\mathcal{H}}\left(\epsilon, \delta_2\right)$.

10. Assuming by way of contradiction that $d_1 =: VCdim\left(\mathcal{H}_1\right) > VCdim\left(\mathcal{H}_2\right) := d_2$, directly by the VC-dimension defenition it holds that there are sets $C_1, C_2 \subset dom\left(\mathcal{H}_1\right) = dom\left(\mathcal{H}_2\right) = \mathcal{X}$ such that $|C_1| = d_1, |C_2| = d_2$ and C_1, C_2 are shattered by $\mathcal{H}_1, \mathcal{H}_2$ respectively - thus there exists $\mathcal{H}_1' \subseteq \mathcal{H}_1, \mathcal{H}_2' \subseteq \mathcal{H}_2$ such that their restrictions over C_1, C_2 are of sizes $2^{|C_1|}, 2^{|C_2|}$ respectively.
    Since $\mathcal{H}_1 \subseteq \mathcal{H}_2$, it holds that $\mathcal{H}_1' \subseteq \mathcal{H}_1 \subseteq \mathcal{H}_2$, and so C_1 is also shattered by $\mathcal{H}_2$ (we have found a subset of $\mathcal{H}_2$ functions which their restriction over C_1 covers all possible labeling) $\Rightarrow VCdim\left(\mathcal{H}_2\right) = d_1$, by contradiction to our claim that $d_1 > d_2$.
    And thus, $\mathcal{H}_1 \subseteq \mathcal{H}_2 \Rightarrow VCdim\left(\mathcal{H}_1\right) \leq VCdim\left(\mathcal{H}_2\right)$. $\square$

5 Separate the Inseparable - Adaboost

13.

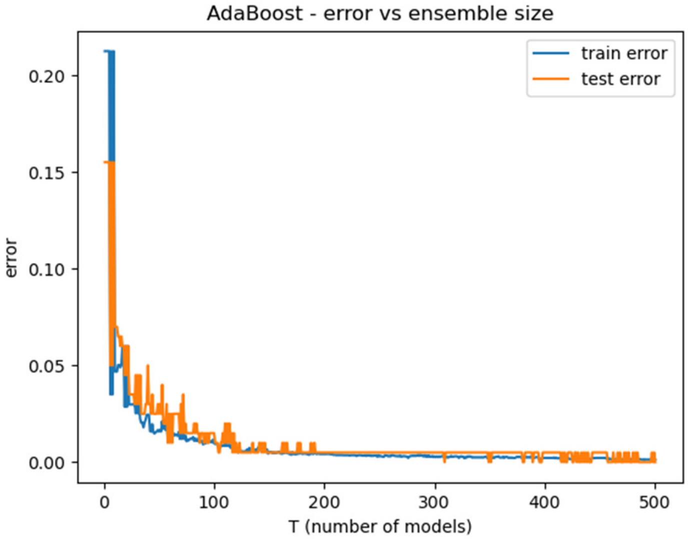


Figure 1: AdaBoost error rate relative to varying ensemble size

14.

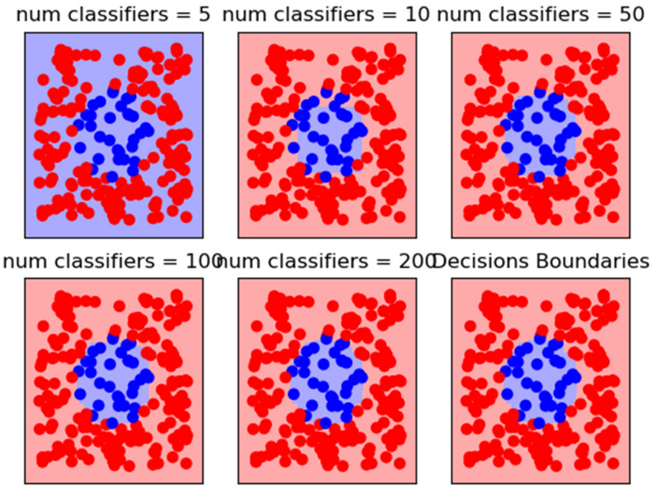


Figure 2: Decision boundary for different ensemble sizes

15.

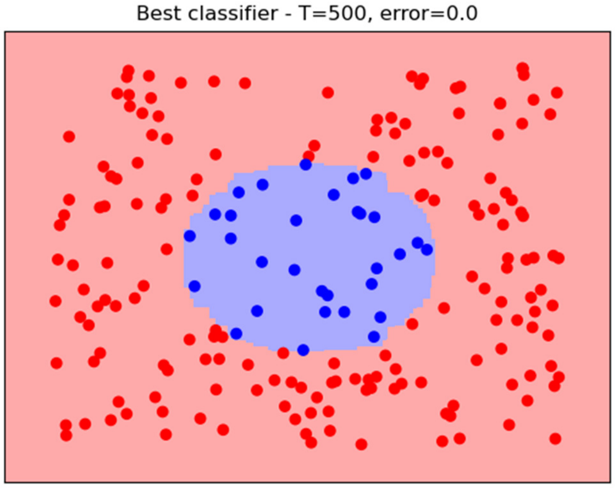


Figure 3: Best preforming ensemble size

16.

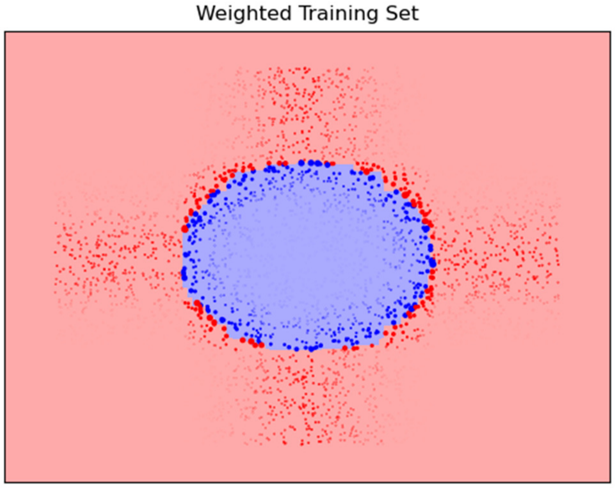
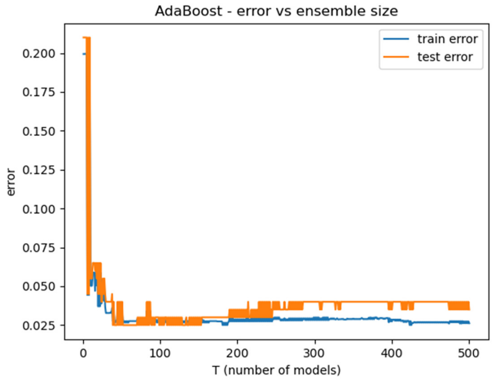
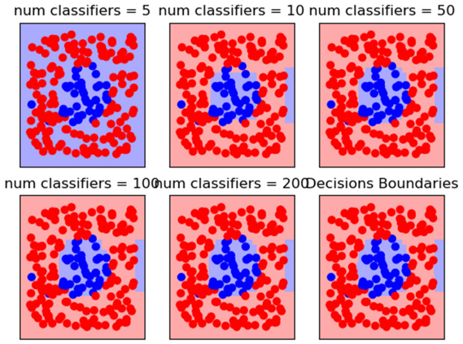


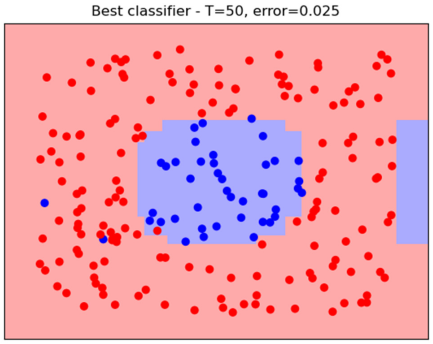
Figure 4: Weighting of the final training set disturbution

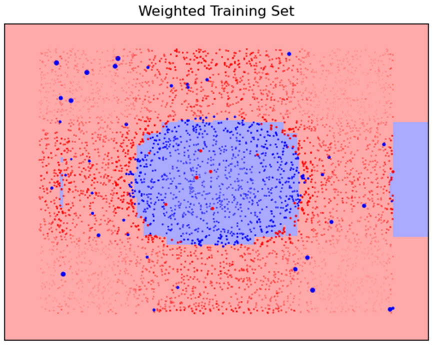
In the provided plot we see the most "problematic" samples. As expected, the boundary between the two classes includes the most difficult samples to classify, and as such, they get misclassified often and thus achieve a higher weight. Furthermore, as our model is a decision stump, it can split the data along the x or y axis, therefore we can notice that weights of corner samples are small, as they can be distiguished well by both types of splits, as oppose to the plus like shape of samples which are more effected by the x,y split limitations, and thus get a higher weight.

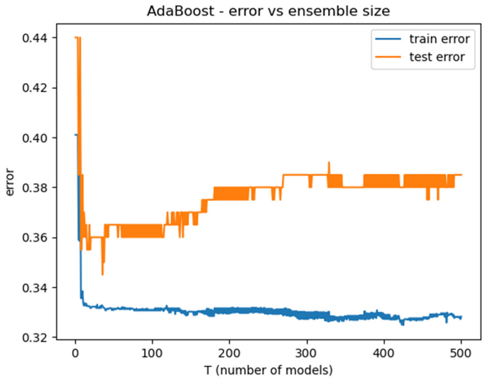
17. (a) First for noise value of 0.01: For noise value of 0.4:

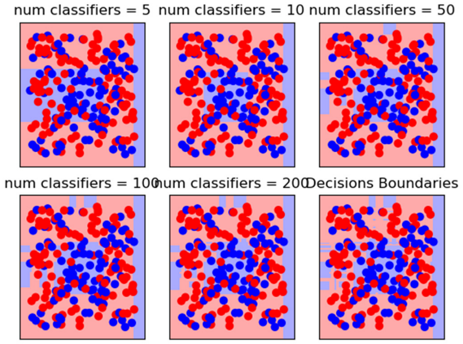

AdaBoost - error vs ensemble size
train error
test error
error
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0
100
200
300
400
500
T (number of models)

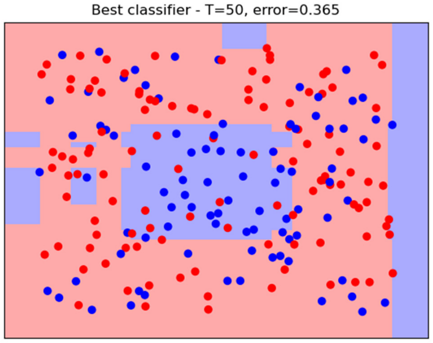

num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
Decisions Boundaries

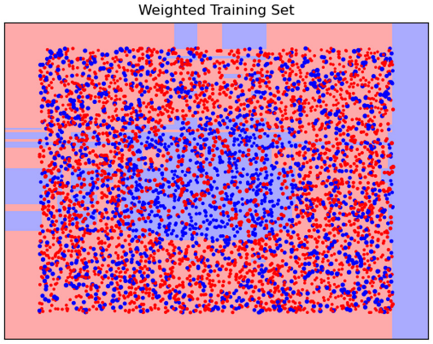

Best classifier - T=50, error=0.025


Weighted Training Set


AdaBoost - error vs ensemble size
train error
test error
error
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0
100
200
300
400
500
T (number of models)


num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
Decisions Boundaries


Best classifier - T=50, error=0.365


Weighted Training Set

(b) As more and more noise is introduced, the two classes are getting mixed, resulting in it becoming harder for the model to generalize well, and to differentiate between them.

(c) Since we introduced noise, the data now is not a perfect representation of the dataset.
As so, the more complex the model gets (by the use of more stumps), the more it is able to overfit and achieve better training error - lower bias, yet as it fits itself to the noised data, it gets further away from a more general solution, failing to generalize and achieving a more substantial estimation error.

(d) As oppose to the noiseless case, where the more complex the model was, the better it was able to learn the data, in the noised case larger ensembles have worse generalization error due to overfitting and as such the best model is one with a rather small ensemble.