Introduction to Machine Learning: Ex4

Or Tal 305793168

1. Let A be a learning algorithm, $\mathcal{D}$ be any distribution, and our loss function is in the range $[0,1]$ (e.g., the 0-1 loss). Prove that the following two statements are equivalent:

   (a) For every $\epsilon, \delta > 0$, there exists $m(\epsilon, \delta)$ such that $\forall m \geqslant m(\epsilon, \delta)$:

   $$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \leqslant \epsilon] \geqslant 1-\delta$$

   (b)

   $$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$$

   Hint: Use Markov's inequality.

⇐:

$\lim_{m\to\infty} \mathbb{E}_{S\sim D^m}\left[L_D\left(A(S)\right)\right] = 0 \Rightarrow from\ markov's\ inequality$: $\forall \epsilon > 0$: $\lim_{m\to\infty} \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right] = 0$

And as $\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right]$ converge to zero, we may conclude:

$$\forall \epsilon, \delta > 0 : \lim_{m\to\infty} \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right] < \delta$$

Meaning that,

$$\forall \epsilon, \delta > 0\ \ \exists m_0 = m(\epsilon,\delta)\ s.t\ \forall m \geq m_0 :\ \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right] < \delta \Rightarrow \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) \leq \epsilon\right] \geq 1-\delta$$

⇒:

$$\forall \epsilon, \delta > 0\ \ \exists m_0 = m(\epsilon,\delta)\ s.t\ \forall m \geq m_0 :\ \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) \leq \epsilon\right] \geq 1-\delta \Rightarrow \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right] < \delta$$

Hence,

$\forall \epsilon, \delta > 0$: $\lim_{m\to\infty} \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right] < \delta \Rightarrow$ choosing a small δ will then result:

$$\lim_{m\to\infty} \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right] \to 0$$

By using Markov's inequality, we will then see that:

$$\frac{\mathbb{E}_{S\sim D^m}\left[L_D\left(A(S)\right)\right]}{\epsilon} \geq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(A(S)\right) > \epsilon\right] \Rightarrow \lim_{m\to\infty} \mathbb{E}_{S\sim D^m}\left[L_D\left(A(S)\right)\right] = 0$$

■

2. **Sample Complexity of Concentric Circles in the Plane** Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$ and let $\mathcal{H}$ be the class of concentric circles in the plane, i.e., $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$, where $h_r(x) = \mathbf{1}[\|x\|_2 \leqslant r]$. Prove that $\mathcal{H}$ is PAC learnable and its sample complexity is bounded by

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant \frac{\log(1/\delta)}{\epsilon} .$$

Note: Please do not use VC dimension arguments but instead prove the claim directly by showing a specific algorithm and analyzing its sample complexity.

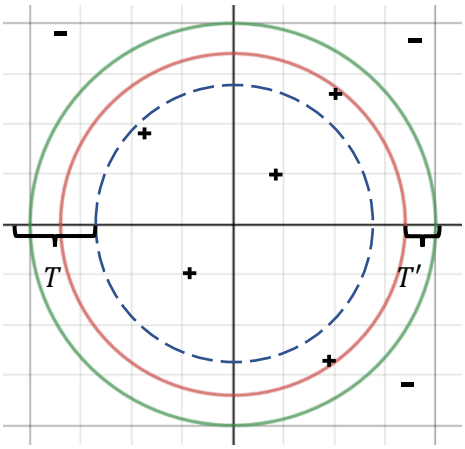
Hint: Remember that for every ϵ,

$$1 - \epsilon \leqslant e^{\epsilon}.$$

Given sample set $S \subset \mathcal{X}, |S| = m$, drawn by some distribution D. We may define $r_0 = \max\{||x||_2 : x \in S\}$ so we could define a classifier: $h_{r_0}(x) = \mathbb{I}\left[||x||_2 \leq r_0\right]$.

Assuming realizability of the hypothesis class (meaning that the ground truth classification is drawn from $\mathcal{H}$ class) means that a ground truth classifier exists = $h_R \in \mathcal{H}$.

Visualization:



Let C = circle with radius R (ground truth, green), C' = circle with radius r_0 (red)

We may conclude that the only region contributing to the loss would then be the region defined by C/C', we would like to show that $L_{D,C}(C') \leq \epsilon$

Let T' be the annulus C/C', and let T be the annulus which encloses the weight ϵ under D.

Clearly, T' has weight exceeding $\epsilon \Leftrightarrow T \subset T'$, and this only happens when no point in S is contained in T.

By the definition of T, the probability that a single draw ($\sim D$) misses the region T is exactly $1 - \epsilon$, therefor the probability of drawing m samples independently($\sim D, iid$), and having all of them all miss the region T is then $(1 - \epsilon)^m$.

given $\delta > 0$, we may choose m that satisfy $(1 - \epsilon)^m \leq \delta$, we saw that $\forall \epsilon > 0$: $(1 - \epsilon) \leq e^{\epsilon}$

$$\delta \geq e^{\epsilon m} \Rightarrow \frac{1}{\delta} \leq e^{-\epsilon m} \Rightarrow \ln\left(\frac{1}{\delta}\right) \leq -\epsilon m \Rightarrow m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon}$$

The above δ, describes the probability where all m drawn samples missed the region T, hence in this case T' has weight (loss) exceeding ϵ. We may then conclude that:

$$\mathbb{P}[T' \subseteq T] \geq 1 - \delta \Rightarrow \forall \epsilon, \delta > 0, m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon} : \mathbb{P}_{S \sim D^m}\left[L_{D,C}(C') \leq \epsilon\right] \geq 1 - \delta.$$

So $\mathcal{H}$ is PAC learnable, having the requested sample complexity. ■

3. **Boolean Conjunctions** Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = \{0,1\}$, and assume $d \geqslant 2$. Each sample $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ consists of an assignment to d boolean variables $(\mathbf{x})$ and a label (y). For each boolean variable x_k, $k \in [d]$, there are two literals: x_k and $\overline{x}_k = 1 - x_k$. The class $\mathcal{H}_{\text{con}}$ is defined by boolean conjunctions over any subset of these $2d$ literals. For example: let $d = 5$ and consider the hypothesis that labels $\mathbf{x}$ according to the following conjunction

$$x_1 \wedge x_2 \wedge \overline{x}_3$$

For $\mathbf{x} = (0,1,1,1,1)$ the label would be 0, and for $\mathbf{x} = (1,1,0,0,0)$ the label would be 1. Compute the VC dimension of $\mathcal{H}_{\text{con}}$ and prove your answer.

By the above definition $\forall h \in \mathcal{H}_{con}, h: \{x, \bar{x} | x \in C', C' \subseteq \mathcal{X}\} \to \mathcal{Y}$

So, we may conclude that $|\mathcal{H}_{con}| = 2^{2 \cdot 2^d}$ =all possible functions from $\{\mathcal{X}, \overline{\mathcal{X}}\}$ to $\mathcal{Y}$.

For $C = \{\mathcal{X}, \overline{\mathcal{X}}\}$, $|\mathcal{H}_{con}| = 2^{|C|} \Rightarrow$ $\mathcal{H}_{con}$ shatters C

And as C is of maximal size ($\mathcal{X}$ is a finite set), we may than conclude that:

$\sup\{m \in \mathbb{N} : \exists C' with\ |C'| = m, and\ \mathcal{H}_{con}\ shatters\ C'\} = |C| = VCdim(\mathcal{H}_{con})$ ■

4. Prove that if $\mathcal{H}$ has the uniform convergence property with function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ then $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leqslant m^{UC}(\epsilon/2, \delta)$.

Lemma 1: given training set S is $\frac{\epsilon}{2}$ – representative, with respect to a domain Z, hypothesis class $\mathcal{H}$, loss function l and a distribution $\mathcal{D}$, any output of $ERM_{\mathcal{H}}(S)$ will satisfy:

$$for\ h_S \in argmin_{h \in \mathcal{H}} L_s(h): \quad L_D(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon = L_{\mathcal{D}}(h') + \epsilon$$

Proof:

$$L_{\mathcal{D}}(h_S) \leq L_{\mathcal{D}}(h_S) + \frac{\epsilon}{2} \leq^{(*)} L_{\mathcal{D}}(h') + \frac{\epsilon}{2} \leq^{(**)} L_{\mathcal{D}}(h') + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_{\mathcal{D}}(h') + \epsilon$$

(*) - h_s minimizes the ERM

(**) - $\epsilon/2$ representive

Definition. We say that an hypothesis class $\mathcal{H}$ has the **uniform convergence property** if there exists $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $0 < \epsilon, \delta < 1$ and every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$,

$$\mathcal{D}^m(\{S \in (\mathcal{X} \times \mathcal{Y})^m \mid S \text{ is } \epsilon\text{-representative}\}) \geqslant 1 - \delta\,.$$

From the definition above, given that $\mathcal{H}$ has the uniform convergence property we may conclude that with $1 - \delta$ probability we would have S which is $\hat{\epsilon} = \epsilon/2$ representative

And by choosing the $ERM_{\mathcal{H}}(S)$ to be our learning algorithm, from Lemma 1 – we would get that

$$\mathcal{D}^m(\{S \in (\mathcal{X} \times \mathcal{Y})^m : |L_{\mathcal{D}}(h) - L_S(h)| < \epsilon/2\}) \geq 1 - \delta$$

Therefore we may conclude that there exists $m_{\mathcal{H}}^{UC}(\epsilon/2\,, \delta)$ suck that $\forall m \geq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$, using $ERM_{\mathcal{H}}(S)$ we would see that $\mathcal{D}^m(\{S \in (\mathcal{X} \times \mathcal{Y})^m : |L_S(h) - L_{\mathcal{D}}(h)| < \epsilon/2\}) \geq 1 - \delta \Rightarrow$ Agnostic PAC learnable.

6. **Of Sample Complexity** Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot,\cdot)$. Show that $m_{\mathcal{H}}$ is monotonically non-increasing in each of its parameters. That is, show that given $\delta \in (0,1)$, and given $0 < \epsilon_1 \leqslant \epsilon_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon_1,\delta) \geqslant m_{\mathcal{H}}(\epsilon_2,\delta)$. Similarly, show that given $\epsilon \in (0,1)$, and given $0 < \delta_1 \leqslant \delta_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon,\delta_1) \geqslant m_{\mathcal{H}}(\epsilon,\delta_2)$.

Assuming distribution D over a sample set $\mathcal{X}$, and as $\mathcal{H}$ is PAC learnable we may assume realizability such that there exist $\hat{h} \in \mathcal{H}$ which is the optimal hypothesis.

Given a learning algorithm A, which learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(.,.)$.

a. Fix some $\delta \in (0,1)$, and choose with no loss of generality ϵ_1, ϵ_2 for which: $0 < \epsilon_1 \leq \epsilon_2 < 1$

   Assuming we draw our samples i.i.d, for S_m= {sized m sample set} , $\forall m \geq m_{\mathcal{H}}(\epsilon_1,\delta)$ and as $\mathcal{H}$ is PAC learnable, A would return an hypothesis $h \in \mathcal{H}$ such that with probability $1-\delta$ we would see that $L_{D,\hat{h}}(h) \leq \epsilon_1 \Rightarrow L_{D,\hat{h}}(h) \leq \epsilon_2$

   By definition $m_{\mathcal{H}}(\epsilon,\delta)$ is the minimal value matching ϵ,δ

   Assume that $m_{\mathcal{H}}(\epsilon_1,\delta) < m_{\mathcal{H}}(\epsilon_2,\delta)$

   We have seen that , $\forall m \geq m_{\mathcal{H}}(\epsilon_1,\delta)$: $D^m\{S_m | L_{D,\hat{h}}(h) \leq \epsilon_2\} \geq 1-\delta$ $\Rightarrow$ contradiction to the minimal property, therefore we may conclude that $m_{\mathcal{H}}(\epsilon_1,\delta) \geq m_{\mathcal{H}}(\epsilon_2,\delta)$.

b. Fix some $\epsilon \in (0,1)$, and choose with no loss of generality δ_1, δ_2 for which: $0 < \delta_1 \leq \delta_2 < 1$

   In a similar manner to (a), A would return a hypothesis $h \in \mathcal{H}$ such that with probability $1-\delta_i$ we would see that $L_{D,\hat{h}}(h) \leq \epsilon$,meaning that $\forall m \geq m_{\mathcal{H}}(\epsilon,\delta_1)$: $D^m\{S_m | L_{D,\hat{h}}(h) \leq \epsilon_2\} \geq 1-\delta_1 \geq 1-\delta_2$

   Assume that $m_{\mathcal{H}}(\epsilon,\delta_1) < m_{\mathcal{H}}(\epsilon,\delta_2) \Rightarrow$ contradiction to the minimal property $\Rightarrow m_{\mathcal{H}}(\epsilon,\delta_1) \geq m_{\mathcal{H}}(\epsilon,\delta_2)$

We may then conclude that $m_{\mathcal{H}}$ is monotonically non-increasing in each of its parameters. ■

8. Let X be a sample space and $\mathcal{Y} = \{\pm 1\}$. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. For $C \subset \mathcal{X}$, recall the notation $\mathcal{H}_C$ for the restriction of $\mathcal{H}$ to the subset C. Define the function $\tau_m(\mathcal{H}) : \mathbb{N} \to \mathbb{N}$ corresponding to $\mathcal{H}$ to be

$$\tau_{\mathcal{H}}(m) := \max\left\{ |\mathcal{H}_C| \,\middle|\, C \subseteq \mathcal{X}\,,\ |C| = m \right\}.$$

(a) Explain, in your own words, the meaning of $\tau_{\mathcal{H}}$.

$\tau_{\mathcal{H}}(m)$ represents the max sized restriction of $\mathcal{H}$ to a subset of size m, intuitively, as we seen in lecture, $\tau_{\mathcal{H}}(m)$ represents the growth rate of $|H_C|$, and if for some m_0: $\forall m > m_0\ \tau_{\mathcal{H}}(m)$ grows polynomially we may then conclude that $VCdim(\mathcal{H}) < \infty \Rightarrow \mathcal{H}$ is Agnostic PAC learnable

(b) Suppose that $VCdim(\mathcal{H}) = \infty$. Find an expression for the value of $\tau_{\mathcal{H}}(m)$ for $m \in \mathbb{N}$.

In this case, as seen in lecture, for every $m \in \mathbb{N}$ we would expect that $\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X}, |C|=m} |\mathcal{H}_C| = 2^m$

(c) Now suppose that $VCdim(\mathcal{H}) = d$. Find an expression for the value of $\tau_{\mathcal{H}}(m)$ for $m \leqslant d$.

In this case, as we know that the $VCdim(\mathcal{H})$ is of finite size d, for every $m \leq d$ we would then expect to have $\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X}, |C|=m} |\mathcal{H}_C| = 2^m$

(d) You will now prove the following important result: suppose that $VCdim(\mathcal{H}) = d$ and let $m > d$. Then

$$\tau_{\mathcal{H}}(m) \leqslant \left(\frac{em}{d}\right)^d,$$

where e is the natural logarithm base. You'll do this in three steps:

i. Using induction, show that for any finite $C \subset \mathcal{X}$,

$$|\mathcal{H}_C| \leqslant \left| \left\{ B \subseteq C \,\middle|\, \mathcal{H} \text{ shutters } B \right\} \right|.$$

Hint: in the induction step divide $\mathcal{H}_C$ to two groups. one of them can be $\mathcal{H}_{C'}$ when $C' = \{c_2, \ldots, c_m\}$.

Important notice: $|\{B \subseteq C | \mathcal{H} \text{ shatters B}\}|$ is the number of subsets of C, where $\mathcal{H}$ shatters the subset

Base case: $|C| = 1$

In this case, $|\mathcal{H}_C| \leq 2$, as for $C = \{c\}$ there are 3 options for functions in $\mathcal{H}_C$:

a. There is no $h \in \mathcal{H}_C$ s.t $h: c \to \{-1,1\}$
b. $\exists h \in \mathcal{H}_C$ s.t $h(c) = 1$
c. $\exists h \in \mathcal{H}_C$ s.t $h(c) = -1$

If b or c occur, then $|\{B \subseteq C | \mathcal{H} \text{ shatters B}\}| = |\{\emptyset, C\}| = 2 \Rightarrow |\mathcal{H}_C| \leq |\{B \subseteq C | \mathcal{H} \text{ shatters B}\}|$

Else, $|\{B \subseteq C | \mathcal{H} \text{ shatters B}\}| = |\{\emptyset\}| = 1 \Rightarrow |\mathcal{H}_C| < |\{B \subseteq C | \mathcal{H} \text{ shatters B}\}|$

We would now assume by induction that the above lemma holds for any $|C| < m$ and will prove for $|C| = m$

Say we choose $C \subseteq \mathcal{X}$, such that $|C| = m$, and we define $C' \subset C \Rightarrow C' = \{c_1, \cdots, c_{m-1}\}, C = \{c_1, \cdots, c_m\}$

We could then split $\mathcal{H}_C$ into two groups:

$\mathcal{H}_{C_1} = \mathcal{H}_{C'}$

$\mathcal{H}_{C_2} = \{h_2 \in \mathcal{H}, h_2 \notin \mathcal{H}_{C'} | \exists\, h_1 \in \mathcal{H}_{C'} s.t\, \forall c_i \in C' : h_1(c_i) = h_2(c_i)\ and\ h_1(c_m) = h_2(c_m)\}$

Intuitively, $\mathcal{H}_{C_2}$ would be the group of functions that would be added to $\mathcal{H}_{C'}$ when adding c_m to C'

It is then simple to see that $\mathcal{H}_C = \mathcal{H}_{C_1} \cup \mathcal{H}_{C_2}$ and that $|\mathcal{H}_C| = |\mathcal{H}_{C_1}| + |\mathcal{H}_{C_2}|$

Define $\mathcal{H}' \subseteq \mathcal{H}$ as $\mathcal{H}' = \{h \in \mathcal{H} | \exists h' \in \mathcal{H}' s.t\, \forall c_i \in C' : h(c_i) = h'(c_i), and\ h(c_m) = -h'(c_m)$

Notice that if $\mathcal{H}'$ shatters $B \subseteq C'$ then it also shatters $B \cup \{c_m\}$, hence we may notice that $\mathcal{H}_{C_2}$ would be the restriction of $\mathcal{H}'$ to $C' \Rightarrow\ \mathcal{H}_{C_2} = \mathcal{H}'_{C'}$

We may than see that from the inductive assumption:

$$|\mathcal{H}_{C_1}| \leq |\{B \subseteq C' | \mathcal{H} \text{ shatters B}\}| = |\{B \subseteq C | c_m \notin B \text{ and } \mathcal{H}\ shatters\ B\}|$$

$$|\mathcal{H}_{C_2}| = |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' | \mathcal{H}' \text{shatters B}\}| = |\{B \subseteq C' | c_m \in B \text{ and } \mathcal{H}' shatters\ B\}|$$

$$\leq |\{B \subseteq C | c_m \in B \text{ and } \mathcal{H}\ shatters\ B\}|$$

Therefore,

$$|\mathcal{H}_C| = |\mathcal{H}_{C_1}| + |\mathcal{H}_{C_2}| \leq |\{B \subseteq C | c_m \notin B \text{ and } \mathcal{H}\ shatters\ B\}| + |\{B \subseteq C | c_m \in B \text{ and } \mathcal{H}\ shatters\ B\}|$$

$\Rightarrow |\mathcal{H}_C| \leq |\{B \subseteq C | \mathcal{H}\ shatters\ B\}|$ as requested.

ii. Explain in your own words the meaning of this inequality.

Proving that the inequality $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ holds, will be stating that the expression $\tau_{\mathcal{H}}(m)$ grows polynomially and not exponentially, hence, as seen in lecture, $|\mathcal{H}_C|$ grows polynomially in $m > d$ which means that it is Agnostic PAC learnable by the ERM learning algorithm.

iii. Show that, for any finite $C \subseteq \mathcal{X}$, we have

$$\left|\{B \subseteq C \mid \mathcal{H} \text{ shutters } B\}\right| \leqslant \sum_{k=0}^{d} \binom{m}{k}$$

The expression $|\{B \subseteq C | \mathcal{H}\ shatters\ B\}|$ would be at most the number of permutations on $C = \sum_{k=0}^{|C|} \binom{|C|}{k}$

As we know that $VCdim(\mathcal{H}) = d$, we may then conclude that for any $C \subseteq \mathcal{X}$, where $|C| > d$, $\mathcal{H}$ does not shatter C (else $VCdim(\mathcal{H}) > d$) and therefore:

$$|\{B \subseteq C | \mathcal{H}\ shatters\ B\}| = |\{B \subseteq C | |C| > d\ and\ \mathcal{H}\ shatters\ B\}| \leq \sum_{k=0}^{d} \binom{|C|}{k} + \sum_{k=d+1}^{|C|} \binom{|C|}{k} + 0$$

Where $\forall k > d, |C| = m > d \Rightarrow \binom{m}{k} = 0 \Rightarrow |\{B \subseteq C | \mathcal{H}\ shatters\ B\}| \leq \sum_{k=0}^{d} \binom{m}{k}$ as requested.

iv. Use the following inequality (which you are not required to prove)

$$\sum_{k=0}^{d} \binom{m}{k} \leqslant \left(\frac{em}{d}\right)^d$$

to finish the proof that $\tau_{\mathcal{H}}(m) \leqslant \left(\frac{em}{d}\right)^d$.

By definition $\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X}, |C|=m} |\mathcal{H}_C|$ and as we have seen in (i), we may then conclude that

$\tau_{\mathcal{H}}(m) \leq |\{B \subseteq C | \mathcal{H}\ shatters\ B\}| \Rightarrow$ from $(iii + iv)$ we may conclude that $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$

(e) If $m = d$, does the inequality $\tau_{\mathcal{H}}(m) \leqslant \left(\frac{em}{d}\right)^d$ hold? If it does hold, is it tight?

In case where $m = d$ we would get: $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{m}\right)^m = e^m > 2^m$

Hence this bound would not be as tight as the number of possible permutations is 2^m therefor we know from (i) that $\tau_{\mathcal{H}}(m) \leq 2^m < e^m$

(f) Characterize in words the behavior of $\tau_{\mathcal{H}}(m)$ for $m \leqslant VCdim(\mathcal{H})$ and for $m > VCdim(\mathcal{H})$. Can you use your characterization to offer an alternative definition of the VC-dimension $VCdim(\mathcal{H})$?

For $m \leq d = VCdim(\mathcal{H})$: we may see that $\tau_{\mathcal{H}}(m) = 2^m$ as $\max_{C, |C|=m} |\mathcal{H}_C| = 2^m$

Define $C \subseteq \mathcal{X}$ s.t $|C| = d$ and $C' = C \cup \{x\}, where\ x \in \mathcal{X}\ and\ x \notin C$

For every choice of x, $|\mathcal{H}_{C'}| \leq 2^{d+1} - 1$ (else $\mathcal{H}$ shatters C')

We may then define $VCdim(\mathcal{H}) = \log_2 \tau_{\mathcal{H}}(m)$, where $m = argmax_{m'}\{\tau_{\mathcal{H}}(m') = 2^{m'}\}$ ■

Programming Questions – Results

AdaBoost class implementation under adaboost.py
Main function to run Q10 to Q14 is under ex4_ans.py
Result figures are in next pages

Bias-complexity tradeoff: (Q10 – different noise_ratios)
In fig 1, we see a plot of the classification error as a function of num of base classifiers (committee size). It is clear the in all cases the error decreases as committee size increases, and as none of these base classifiers had increased its complexity (they are all decision stumps), we may conclude that this is due to decrease in bias.

Differences in minimal test error cases with different noise ratios
As seen in fig 5, as noise ratio increases it becomes harder for the model to predict the label of the samples as for 0.4 case, it appears that the positive and negative samples are intertwined in such way that without having a background color that marks the decision of the model, it's extremely hard to decide where should be classified positive or negative with the naked eye. Such scatter may result an overfitting to the train set as it may seen that the model colors inaccurate sections of the space (see bottom left blue corner for example at noise=0.4 fig)
On the other hand, small noise ratio seemed to accurately color the space, as only a small mass of samples was measured from an area that does not match their label.

Why did we had to scale the weights in order to see them?
The input weights are used as marker size for plotting. These weights are initialized uniformly, hence as there are 5000 training samples, the initial weight for each sample is 0.0002, and even if some weight only grows within the 500 iterations the train phase performs, it's still relatively small compared to the figure dimensions, therefore scaling was necessary for visibility.

Changes in results

1. **As committee size increases** – we can notice an improvement in prediction in every case, as stated above, the bias goes down as committee size goes up, the improvement in precision could also be seen in Fig 2-4, as the "division" of the space grows accurately.
2. **As noise ratio increases** – as stated above, as noise ratio increases, we see a decrease in precision as it becomes harder for the model to divide the space, we see unexpected "divisions" of the space in the 0.4 noise ratio case, which is very dependent on the actual train set, and estimates the distribution inaccurately.

Fig 1：classification error as a function of committee size

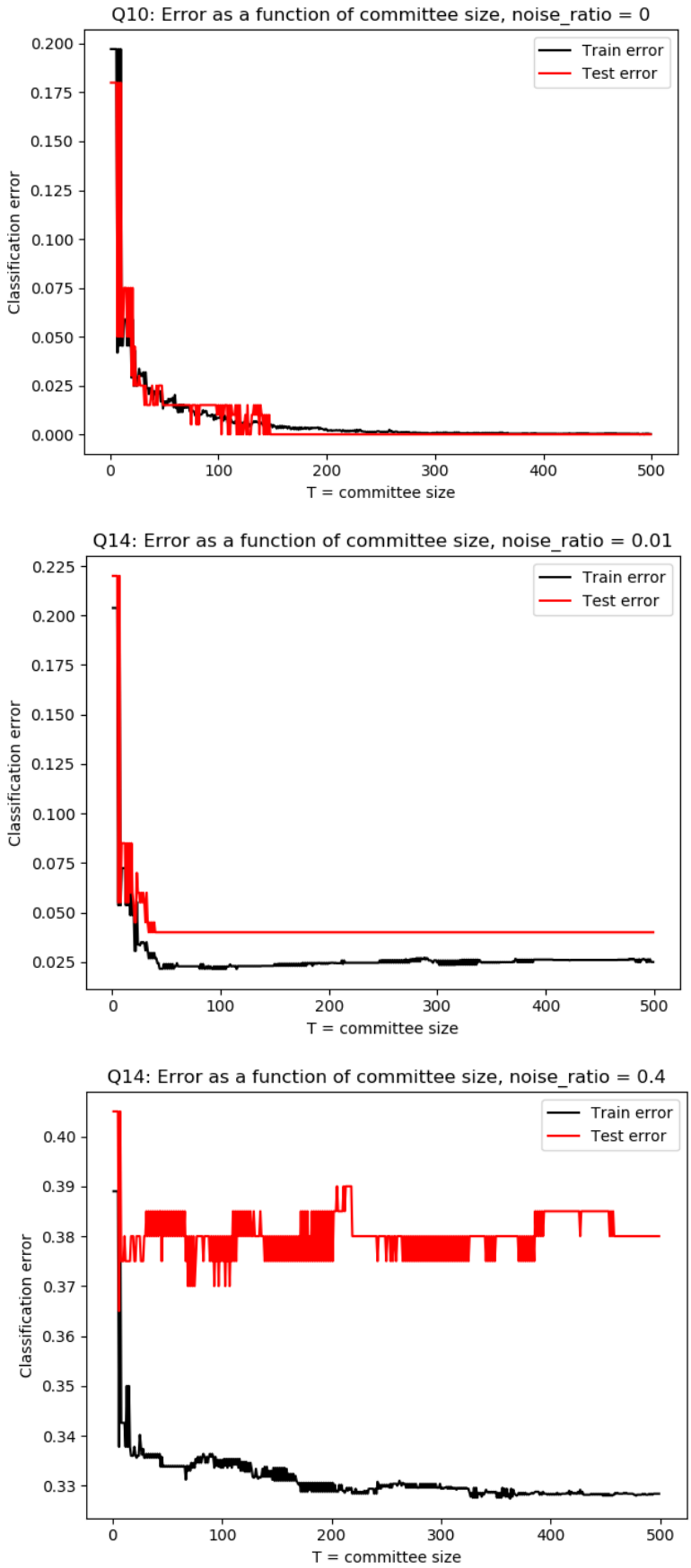


Fig 2-4：decision boundaries over train set for different num of classifiers, and different noise ratios

Fig 2

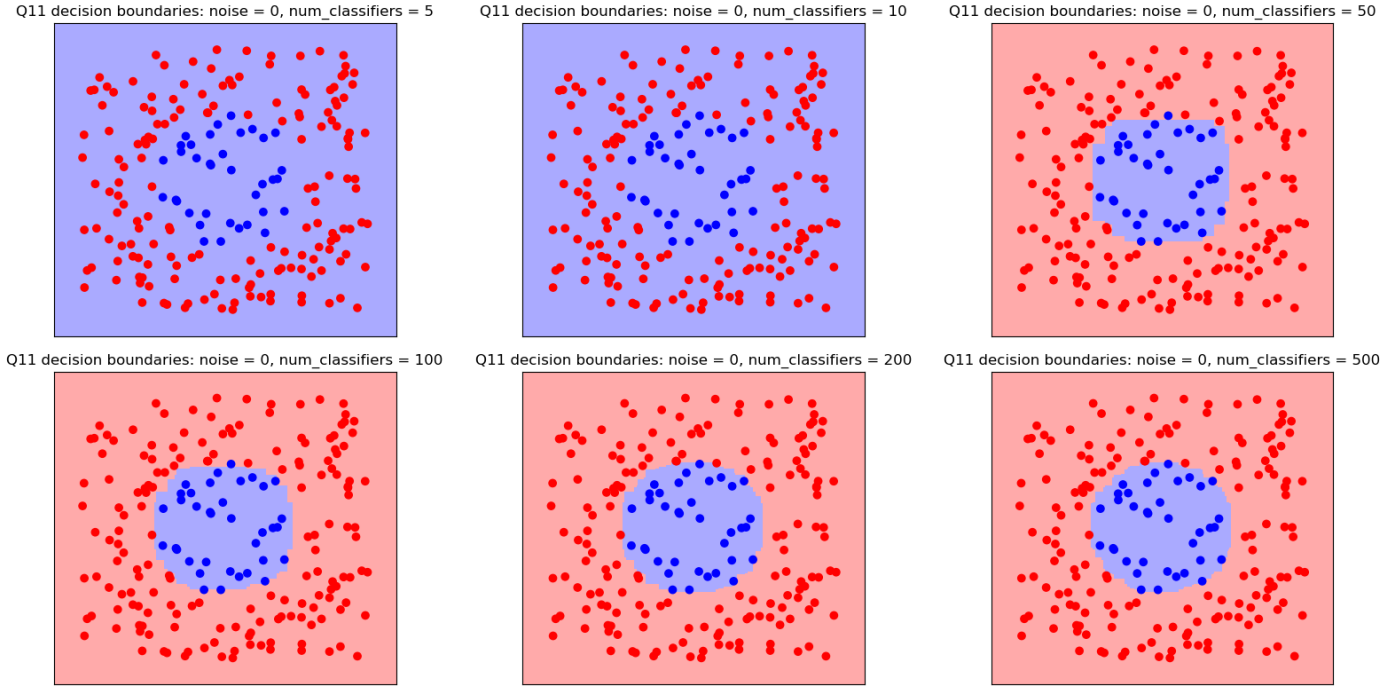


Fig 3

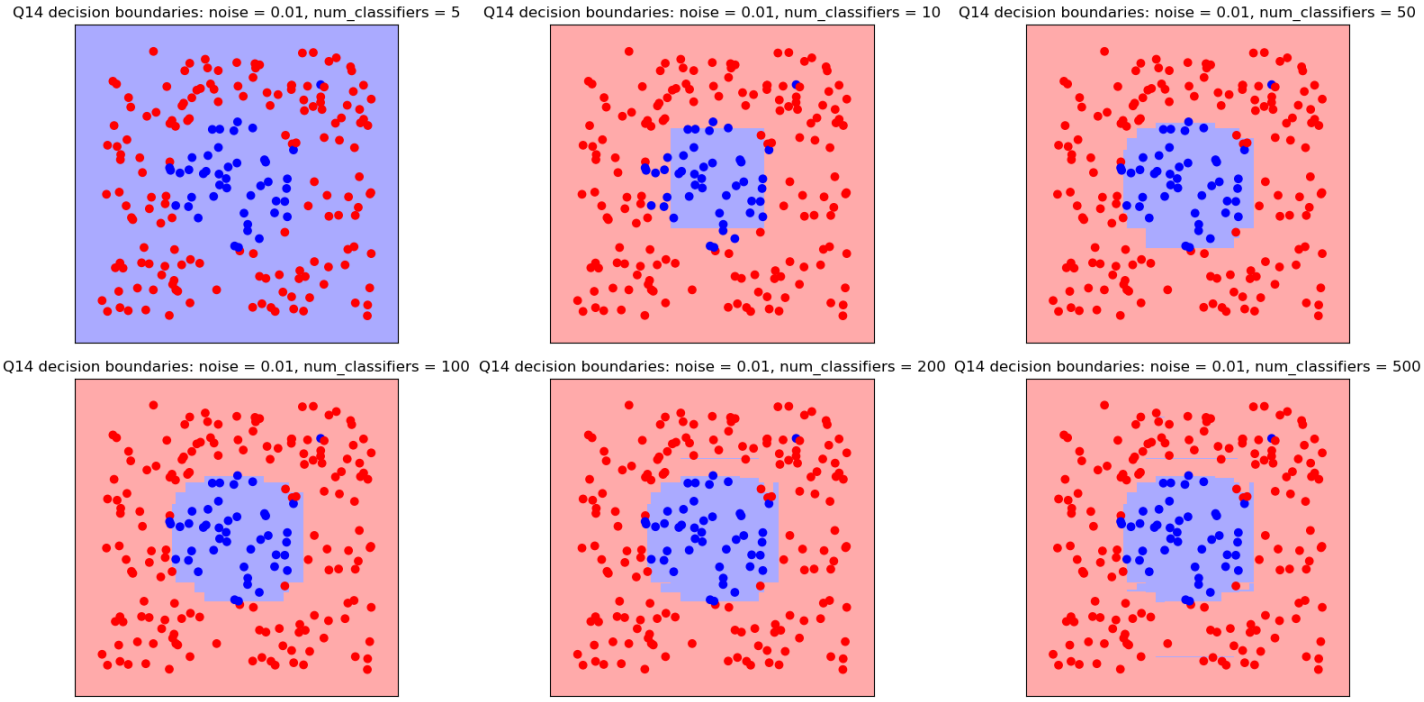


Fig 4

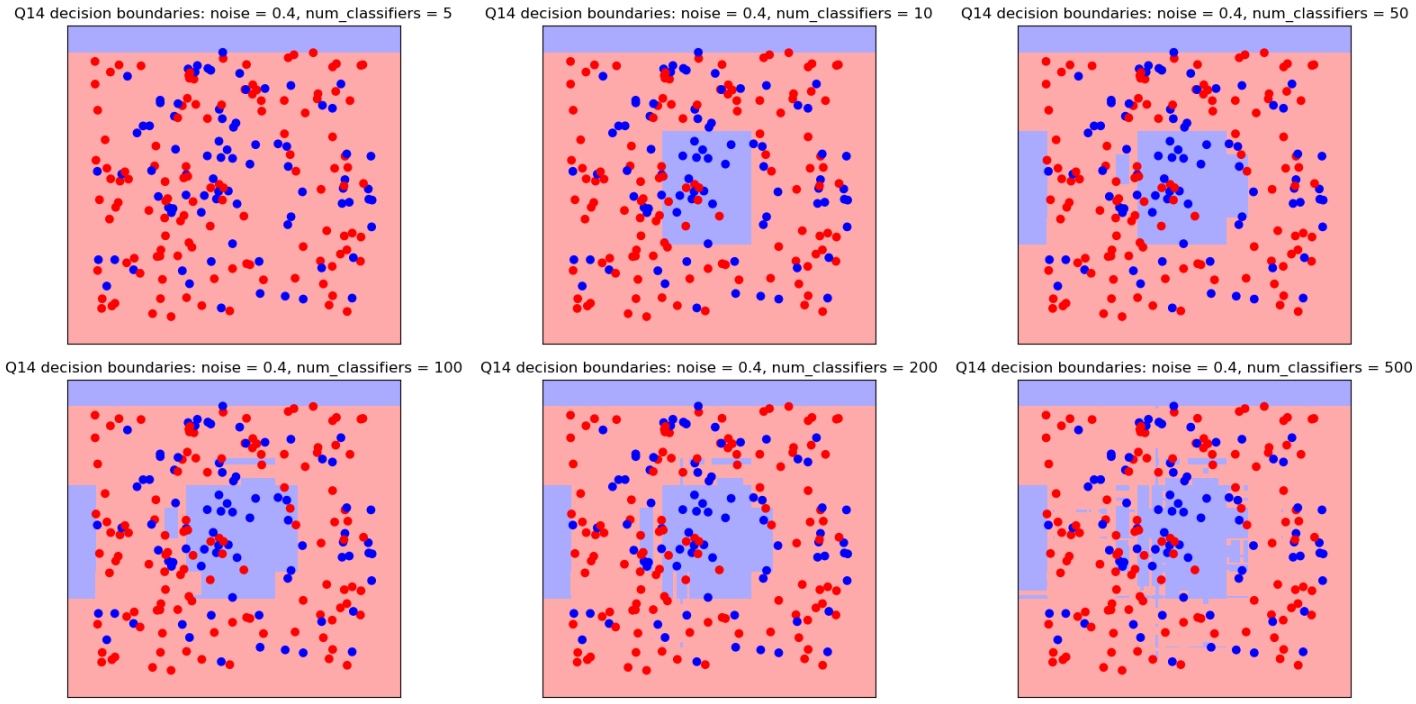

Q14 decision boundaries: noise = 0.4, num_classifiers = 5
Q14 decision boundaries: noise = 0.4, num_classifiers = 10
Q14 decision boundaries: noise = 0.4, num_classifiers = 50
Q14 decision boundaries: noise = 0.4, num_classifiers = 100
Q14 decision boundaries: noise = 0.4, num_classifiers = 200
Q14 decision boundaries: noise = 0.4, num_classifiers = 500

Fig 5: decision boundaries over train set matching committee size ($\hat{T}$) that minimizes the set error

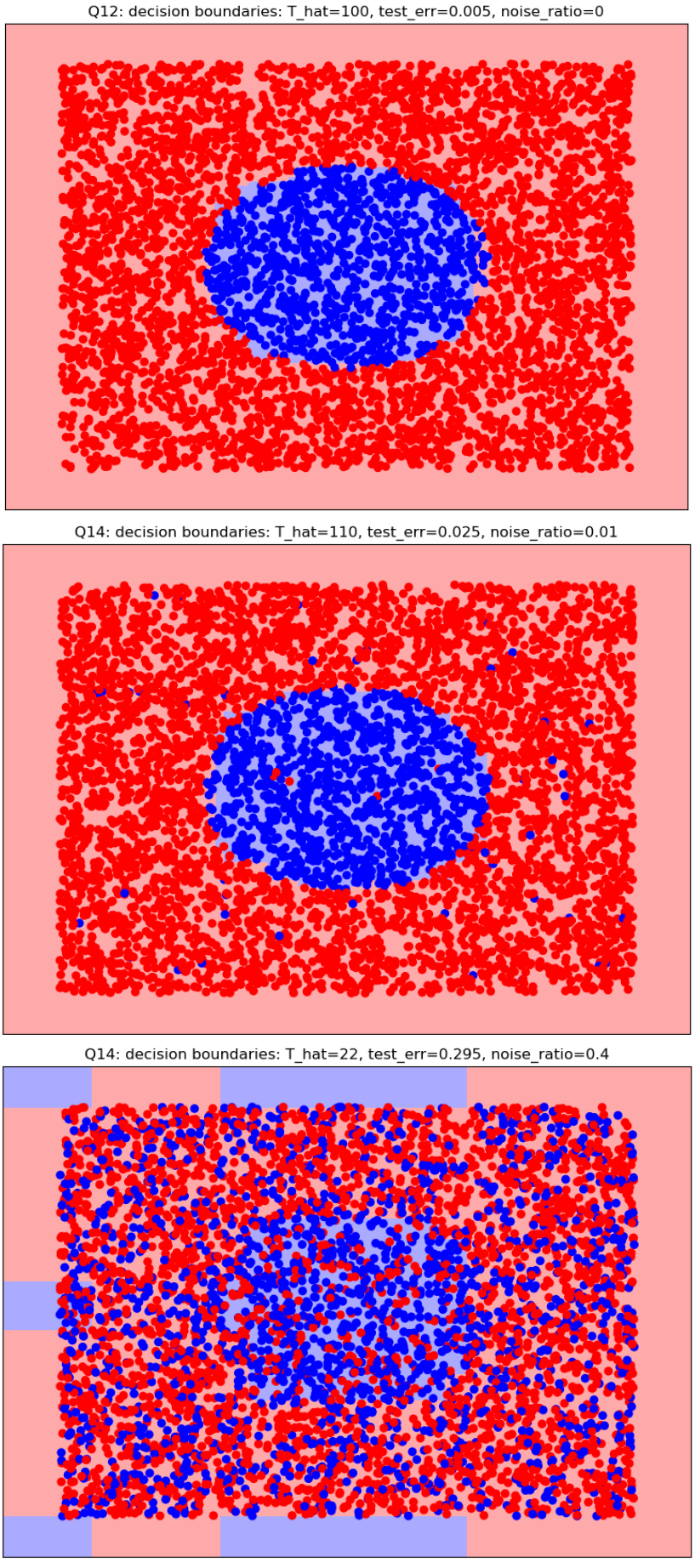


Fig 6: decision boundaries over train set matching committee size = 500, samples scaled by weight