IML Ex4

Eliyahu Strugo [redacted]

June 5, 2020

PAC Learnability

1.Solution:

- $(a) \Rightarrow (b)$:

  Assume that for any $\epsilon, \delta > 0$, there exists $m(\epsilon, \delta)$ s.t $\forall m \geq m(\epsilon, \delta)$ it holds that:

  $$\mathbb{P}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \leq \epsilon \right] \geq 1 - \delta$$

  Let $\epsilon > 0$ and $\delta = \frac{\epsilon}{2}$.

  $\alpha \coloneqq \mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \,|L_{\mathcal{D}}(h_S) > \frac{\epsilon}{2} \right] \mathbb{P}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) > \frac{\epsilon}{2} \right]$

  $\beta \coloneqq \mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \,|L_{\mathcal{D}}(h_S) \leq \frac{\epsilon}{2} \right] \mathbb{P}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \leq \frac{\epsilon}{2} \right]$

  By the Law of Total Expectation:

  $$\mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \right] = \alpha + \beta$$

  - α : Since $L_{\mathcal{D}}(h_S) \leq 1$ then $\mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \,|L_{\mathcal{D}}(h_S) > \frac{\epsilon}{2} \right] \leq 1$
    and there exists $m\left(\frac{\epsilon}{2}, \frac{\epsilon}{2}\right)$ s.t $\forall m \geq m\left(\frac{\epsilon}{2}, \frac{\epsilon}{2}\right)$ $\mathbb{P}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) > \frac{\epsilon}{2} \right] \leq \frac{\epsilon}{2} \Rightarrow \alpha \leq \frac{\epsilon}{2}$
  - β : $\mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \,|L_{\mathcal{D}}(h_S) \leq \frac{\epsilon}{2} \right] \leq \frac{\epsilon}{2}$ and $\mathbb{P}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \leq \frac{\epsilon}{2} \right] \leq 1 \Rightarrow \beta \leq \frac{\epsilon}{2}$

  Hence, $\forall m \geq m\left(\frac{\epsilon}{2}, \frac{\epsilon}{2}\right)$ it holds that $\mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \right] = \alpha + \beta \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$ and therefore:

  $$\mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_S) \right] \underset{m \to \infty}{\longrightarrow} 0$$

- $(b) \Rightarrow (a)$:

  Assume that $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right)\right] \underset{m\to\infty}{\longrightarrow} 0 \Rightarrow \forall\epsilon > 0\, \exists m\left(\epsilon\right) \in \mathbb{N} \text{ s.t } \forall m \geq m\left(\epsilon\right)$:

  $$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right)\right] \overset{*}{=} \left|\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right)\right]\right| < \epsilon$$

  * Since $0 \leq L_{\mathcal{D}}\left(h_S\right)$ then $0 \leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right)\right]$.

  Let $\epsilon > 0$. By the Markov's ineuqality it holds that:

  $$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right) > \epsilon\right] \leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right)\right]}{\epsilon}$$

  Now ,Let $\delta > 0$ then $\epsilon\delta > 0$ and $\exists m\left(\epsilon, \delta\right) \in \mathbb{N}$ s.t $\forall m \geq m\left(\epsilon, \delta\right)$:

  $$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right) > \epsilon\right] \leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right)\right]}{\epsilon} < \frac{\epsilon\delta}{\epsilon} = \delta$$

  Hence:

  $$1 - \delta \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right) \leq \epsilon\right]$$

2.Solution:

Given training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$, define $A = \{r \in \mathbb{R}_+ | \forall (x, 1) \in S$ it holds that $\|x\|_2 \leq r\}$.

Our algorithem is ERM which finds a radius r_{alg} s.t:

$$r_{alg} = \begin{cases} min\{A\} & A \neq \emptyset \\ 0 & else \end{cases}$$

and returns a hypothsis $h_S = A(S)$ .Fix a distribution $\mathcal{D}$ over $\mathcal{X}$. Since we assuming realizablity then $\mathcal{H}$ contains the true hypothsis. Fix ture hypothsis $h_{r*} \in \mathcal{H} \Rightarrow L_{\mathcal{D}}(h_{r*}) = 0$. By the definition of r_{alg} it holds that:

$$r_{alg} < r^*$$

Let $x \in \mathcal{X}$. if $\|x\| \leq r_{alg}$ or $r^* < \|x\|$ then the algorithem is correct and if $r_{alg} < \|x\| \leq r^*$ the algorithem will be incorrect.Therefore, we want to examine the case where the algorithem is incorrect.

Let $\epsilon > 0$.

If $\mathcal{D}(\{x : \|x\| \leq r^*\}) < \epsilon \Rightarrow \mathcal{D}(\{x : r_{alg} < \|x\| \leq r^*\}) < \epsilon \Rightarrow \mathbb{P}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S) \leq \epsilon] = 1$ and we are done, otherwise $\mathcal{D}(\{x : \|x\| \leq r^*\}) \geq \epsilon$.

Let us define r' s.t $\mathcal{D}(\{x : r' \leq \|x\| \leq r^*\}) \leq \epsilon$ and define $B = \{x : r' \leq \|x\| \leq r^*\}$. Since $A(S)$ picks the minimal radius that contains all positive instances then:

$$S \cap B \neq \emptyset \Rightarrow r' \leq r_{alg}$$

Which means:

$$\{x : r_{alg} < \|x\| \leq r^*\} \subseteq B$$

$$\Rightarrow \mathcal{D}(\{x : r_{alg} < \|x\| \leq r^*\}) \leq \mathcal{D}(B) \leq \epsilon$$

And we get $S \cap B \neq \emptyset \Rightarrow L_D(h_S) \leq \epsilon$, which holds iff $L_D(h_S) > \epsilon \Rightarrow S \cap B = \emptyset$, and then:

$$\mathbb{P}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S \sim \mathcal{D}^m}[S \cap B = \emptyset] = \mathbb{P}_{S \sim \mathcal{D}^m}[\forall (x, y) \in S, (x, y) \notin B] \underset{i.i.d}{=} \prod_{i=1}^{m} \mathbb{P}_{x_i, y_i \sim \mathcal{D}^m}[(x_i, y_i) \notin B]$$

Moreover, $\mathcal{D}(B) \leq \epsilon \Rightarrow \mathcal{D}(B \cap S) \leq \epsilon \Rightarrow \forall i \in [m], \mathbb{P}_{x_i,y_i \sim \mathcal{D}^m}[\{x_i, y_i\} \in B] \leq \epsilon$ then:

$$\mathbb{P}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \prod_{i=1}^{m} \mathbb{P}_{x_i,y_i \sim \mathcal{D}^m}[(x_i, y_i) \notin B] \leq (1-\epsilon)^m \leq e^{-m\epsilon}$$

And $e^{-m\epsilon} \leq \delta \iff \frac{1}{\epsilon} ln\left(\frac{1}{\delta}\right) \leq m$.

VC dimension

3.Solution:

There is shattered sample of size d:

Let $C = \{v_1, \ldots, v_d\}$ s.t $v_1, \ldots, v_d \in \mathbb{R}^d$ and $\forall i, j \in [d], v_i(j) = 1 - \delta_{i,j}$, where $v_i^{(j)}$ is the j-th coordinate of v_i and $\delta_{i,j}$ is Kronecker delta.

Let $y = (y_1, \ldots, y_d) \in \{0,1\}^d$ and define $h_y = \bigwedge_{i:y_i=0} x_i$ if $y \neq 1_{\mathbb{R}^d}$ else $h_y = 1$. For any $i \in [d]$ it holds that $h_y(v_i) = y_i$. Thus, C is shattered $\Rightarrow VCdim(\mathcal{H}) \geq d$.

There is no shattered sample of size $d+1$:

Assume towards contradiction that there exists $C = \{v_1, \ldots, v_{d+1}\}$ that is shattered by $\mathcal{H}$, which means that points in C can have all possible labeling $\Rightarrow \forall i, j \in [d+1]$ there exists $h_i \in \mathcal{H}$ s.t $h_i(v_j) = 1 - \delta_{i,j}$.

If $h_i(v_j) = 1 - \delta_{i,j}$ then the conjuction h_i must contain some literal $l_i = x_k$ or $l_i = \bar{x}_k$ s.t l_i is false on $v_i(k)$. Since $|C| = d+1$ and $|\{l_1, \ldots, l_d\}| = d$ then by the Pigeonhole principle

there is at least one variable that occurs twise. w.l.o.g assume that l_1 and l_2 contain the same variable .

- If $l_1 = l_2$: then l_1 is false on $v_1(k)$ and $l_2$is false on $v_2(k)$and then $h_1(v_1) = h_2(v_2) = 0$ contadiction.
- If $l_1 = \neg l_2$: since $d \geq 2 \Rightarrow d+1 \geq 3$ then v_3 exists and either l_1 or l_2 is false on $v_3 \Rightarrow$ either $h_1(v_3) = 0$ or $h_2(v_3) = 0$ contradiction.

Hence, $d \leq VCdim(\mathcal{H}) \leq d \Rightarrow VCdim(\mathcal{H}) = d$.

4.Solution:

Let $h_S = \underset{h\in\mathcal{H}}{argmin}\left\{L_S\left(h\right)\right\}$and $h^* = \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}\left(h\right)$.

Claim : For every distribution $\mathcal{D}$ and sample $S \sim D^m$ it holds that:

$$L_{\mathcal{D}}\left(h_S\right) - h^* \leq 2\underset{h\in\mathcal{H}}{sup}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right|$$

Proof:

$$L_{\mathcal{D}}\left(h_S\right) - h^* = L_{\mathcal{D}}\left(h_S\right) - L_S\left(h_S\right) + L_S\left(h_S\right) - h^*$$

$$\leq L_{\mathcal{D}}\left(h_S\right) - L_S\left(h_S\right) + L_S\left(h^*\right) - h^* \leq 2\underset{h\in\mathcal{H}}{sup}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right|$$

Back to the question : Let $\epsilon, \delta > 0$. By the claim we get:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\underset{h\in\mathcal{H}}{sup}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right| \leq \frac{\epsilon}{2}\right] = \mathbb{P}_{S\sim\mathcal{D}^m}\left[2\underset{h\in\mathcal{H}}{sup}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right| \leq \epsilon\right] \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right) - \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}\left(h\right) \leq \epsilon\right]$$

And if $\forall h \in \mathcal{H}\ \left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right| \leq \frac{\epsilon}{2}$ then $\underset{h\in\mathcal{H}}{sup}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right| \leq \frac{\epsilon}{2}$, so we get:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\forall h \in \mathcal{H}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right| \leq \frac{\epsilon}{2}\right] \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[\underset{h\in\mathcal{H}}{sup}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right| \leq \frac{\epsilon}{2}\right] \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right) - \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}\left(h\right) \leq \epsilon\right]$$

Then by the uniform convergence property $\forall m \geq m\left(\frac{\epsilon}{2}, \delta\right)$ it holds that:

$$1 - \delta \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[\forall h \in \mathcal{H}\left|L_{\mathcal{D}}\left(h\right) - L_S\left(h\right)\right| \leq \frac{\epsilon}{2}\right] \leq \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right) - \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}\left(h\right) \leq \epsilon\right]$$

7.Solution:

Let $\mathcal{H}^1 \subseteq \mathcal{H}^2$ be two classes for binary classification and let $C \subset \mathcal{X}$ s.t $C = \{c_1, \ldots, c_m\}$.

By definition $\mathcal{H}^1_C = \left\{h(c_1), \ldots, h(c_m) : h \in \mathcal{H}^1\right\}$ and since $\mathcal{H}^1 \subseteq \mathcal{H}^2$ it holds :

$$\left\{h(c_1), \ldots, h(c_m) : h \in \mathcal{H}^1\right\} \subseteq \left\{h(c_1), \ldots, h(c_m) : h \in \mathcal{H}^1\right\} \cup \left\{h(c_1), \ldots, h(c_m) : h \in \mathcal{H}^2 \setminus \mathcal{H}^1\right\}$$

$$= \left\{h(c_1), \ldots, h(c_m) : h \in \mathcal{H}^2\right\}$$

Thus, $\mathcal{H}^1_C \subseteq \mathcal{H}^2_C$ .Therefore if $\mathcal{H}^1$shutters C then

$$\left|\mathcal{H}^1_C\right| = 2^{|C|} \leq \left|\mathcal{H}^2_C\right| \leq 2^{|C|} \Rightarrow \left|\mathcal{H}^2_C\right| = 2^{|C|}$$

$\Rightarrow \mathcal{H}^2_C$ shutters C. Now, by the definition of $VCdim$ we get :

$$VCdim\left(\mathcal{H}^1\right) \leq VCdim\left(\mathcal{H}^2\right)$$

8.1.Solution:

$\tau(m)$ is the maximum number of ways that m points can be can be classified by using $\mathcal{H}$.

8.2.Solutions:

If $VCdim(\mathcal{H}) = \infty$ then for all $m \in \mathbb{N}$ there exists C with $|C| = m$ s.t $|\mathcal{H}_C| = 2^m \Rightarrow$for all m it holds that $\tau(m) = 2^m$

8.3.Solution:

If $m \leq d$ then there exists C with $|C| = m$ s.t $|\mathcal{H}_C| = 2^m$ and since $\forall C \subset \mathcal{X}$ it holds that $|\mathcal{H}_C| \leq 2^{|C|}$ then $\tau(m) = 2^m$.

8.4.1.Solution:

Claim : for any finite $C \subset \mathcal{X}$

$$|\mathcal{H}_C| \leq |\{B \subseteq C \,:\, \mathcal{H} \text{ shutters } B\}|$$

Proof : Let $C = \{c_1, \ldots, c_m\}$. We proof the claim by induction on m.

Base : $(m = 1)$ Let $B \subseteq C$ then $B = C = \{c_1\}$ or $B = \emptyset \Rightarrow |\{B \subseteq C \,:\, \mathcal{H} \text{ shutters } B\}| \leq 2$.

- We know that $\mathcal{H}$ shutters $\emptyset \Rightarrow 1 \leq |\{B \subseteq C \,:\, \mathcal{H} \text{ shutters } B\}| \leq 2$
- If $|\mathcal{H}_C| \leq 1$ then $|\mathcal{H}_C| \leq |\{B \subseteq C \,:\, \mathcal{H} \text{ shutters } B\}|$
- Otherwise $|\mathcal{H}_C| = |\{h(c_1) \,:\, h \in \mathcal{H}\}| = 2 = 2^{|C|} \Rightarrow \mathcal{H}$ shutters $\emptyset$ and $\{c_1\} \Rightarrow |\{B \subseteq C \,:\, \mathcal{H} \text{ shutters } B\}| = 2$ and we are done

Step : Assume the statment holds for any $k < m$ and $1 < m$:

Denote $C' = \{c_2, \ldots, c_m\}$ and define $\mathcal{H}^C = \left\{h \in \mathcal{H} \,:\, \exists h' \in \mathcal{H} \text{ s.t } h_{|C'} \equiv h'_{|C'} \text{ and } h(c_1) \neq h'(c_1)\right\}$. We claim that $|\mathcal{H}_C| = \left|\mathcal{H}_{C'}\right| + \left|\mathcal{H}^C_{C'}\right|$:

For any $\bar{y} = (y_2, \ldots, y_m)$ and $i \in \{0, 1\}$ define :

$$I_i(\bar{y}) = \begin{cases} 1 & (i, y_2, \ldots, y_m) \in \mathcal{H}_C \\ 0 & else \end{cases}$$

And, denote $X_k = \left\{\bar{y} \in \mathcal{H}_{C'} \,:\, I_0(\bar{y}) + I_1(\bar{y}) = k\right\}$ for $k \in \{1, 2\}$. Now clearly $|\mathcal{H}_C| = |X_1| + |X_2|$.

Moreover, since $(y_1, y_2, \ldots, y_m) \in \mathcal{H}_C \Leftrightarrow (y_2, \ldots, y_m) \in \mathcal{H}_{C'}$ then $|X_0| = \left|\mathcal{H}_{C'}\right|$,

and $\bar{y} \in X_2 \Leftrightarrow (1, y_2, \ldots, y_m), (0, y_2, \ldots, y_m) \in \mathcal{H}_C \Leftrightarrow \bar{y} \in \mathcal{H}^C_{C'}$. Therefore,$|\mathcal{H}_C| = \left|\mathcal{H}_{C'}\right| + \left|\mathcal{H}^C_{C'}\right|$.

Let us examine the two sets:

- $\mathcal{H}_{C'}$: By the I.H it holds that $\left|\mathcal{H}_{C'}\right| \leq \left|\left\{B \subseteq C' \,:\, \mathcal{H} \text{ shutters } B\right\}\right|$
- $\mathcal{H}^C_{C'}$: By the I.H it holds that $\left|\mathcal{H}^C_{C'}\right| \leq \left|\left\{B \subseteq C' \,:\, \mathcal{H}^C \text{ shutters } B\right\}\right|$.

  Let $B = \{b_1, \ldots, b_k\} \subseteq C'$.

  If $\mathcal{H}^C$ shutters B then by the definition of $\mathcal{H}^C$ $\forall h \in \mathcal{H}^C$ if $(h(b_1), \ldots, h(b_k)) \in \mathcal{H}^C_B$ then :

$$(1, h(b_1), \ldots, h(b_k)), (0, h(b_1), \ldots, h(b_k)) \in \mathcal{H}^C_{B \cup \{c_1\}}$$

Hence $\left|\mathcal{H}_B^C\right| = 2^k \Rightarrow \left|\mathcal{H}_{B\cup\{c_1\}}^C\right| = 2^{k+1} \Rightarrow \left\{B \subseteq C' : \mathcal{H}^C \text{ shutters } B\right\} \subseteq \left\{B \subseteq C' : \mathcal{H}^C \text{ shutters } B \cup \{c_1\}\right\}$, and thus :

$$\left|\mathcal{H}_{C'}^C\right| \leq \left|\left\{B \subseteq C' : \mathcal{H}^C \text{ shutters } B\right\}\right| \leq \left|\left\{B \subseteq C' : \mathcal{H}^C \text{ shutters } B \cup \{c_1\}\right\}\right|$$

Moreover, since $\mathcal{H}^C \subseteq \mathcal{H}$ then $\left\{B \subseteq C' : \mathcal{H}^C \text{ shutters } B \cup \{c_1\}\right\} \subseteq \left\{B \subseteq C' : \mathcal{H} \text{ shutters } B \cup \{c_1\}\right\}$.

Finally, $\left\{B \subseteq C' : \mathcal{H} \text{ shutters } B\right\} = \{B \subseteq C : \mathcal{H} \text{ shutters } B \text{ and } c_1 \notin B\}$

and $\left\{B \subseteq C' : \mathcal{H}^C \text{ shutters } B \cup \{c_1\}\right\} = \left\{B \subseteq C : \mathcal{H}^C \text{ shutters } B \text{ and } c_1 \in B\right\}$ we get :

$$|\mathcal{H}_C| = \left|\mathcal{H}_{C'}\right| + \left|\mathcal{H}_{C'}^C\right|$$

$$\leq |\{B \subseteq C : \mathcal{H} \text{ shutters } B \text{ and } c_1 \notin B\}| + |\{B \subseteq C : \mathcal{H} \text{ shutters } B \text{ and } c_1 \in B\}| = |\{B \subseteq C : \mathcal{H} \text{ shutters } B\}|$$

8.4.2.Solution:

The number of possible labeling of points in C is not more then the number of subsets of C that there points have all possibe labeling.

8.4.3.Solution:

Assume $d < m$ and $VCdim(\mathcal{H}) = d$.

Let $B \in \{B \subseteq C : \mathcal{H} \text{ shutters } B\}$then $VC(\mathcal{H}) = d \Rightarrow |B| \leq d \Rightarrow |\{B \subseteq C : \mathcal{H} \text{ shutters } B\}|$. Since for fixed $0 \leq k \leq d$ the number subsets of size k of a set of size m is $\binom{m}{k}$ then :

$$|\{B \subseteq C : \mathcal{H} \text{ shutters } B\}| \leq \sum_{k=0}^{d} \binom{m}{k}$$

8.4.4.Solution:

Let $d, m \in \mathbb{N}$ s.t $VCdim(\mathcal{H}) = d$, $d < m$ and let $C^* \subset \mathcal{X}$ s.t $\tau(m) = |\mathcal{H}_{C^*}|$ then according to (8.4.1) it holds that :

$$\tau(m) = |\mathcal{H}_{C^*}| \leq |\{B \subseteq C^* : \mathcal{H} \text{ shutters } B\}| \leq \sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{em}{d}\right)^d$$

8.4.5.Solution:

For non empty $\mathcal{H}$ it holds that $\forall C \subset \mathcal{X}$ it holds that $|\mathcal{H}_C| \leq 2^{|C|}$ then If $m = d$ then there exists $C^* \subset \mathcal{X}$ s.t $\mathcal{H}$ shutters $C^* \Rightarrow \tau(m) = |\mathcal{H}_{C^*}| = 2^m$ .Therefore :

$$\tau(m) = 2^m \leq \sum_{k=0}^{m} \binom{m}{k} = 2^m < e^m$$

Hence, the inequality holds but it is not tight.

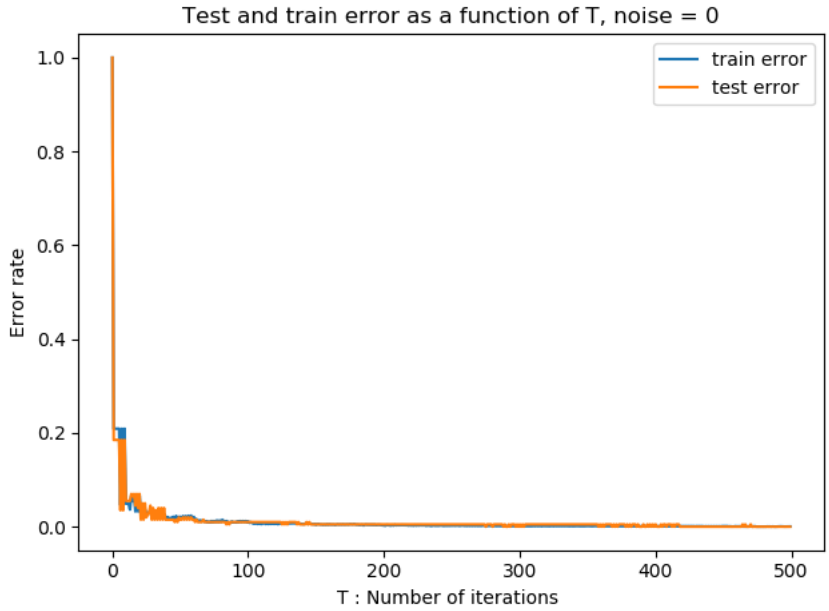
8.4.6.Solution:

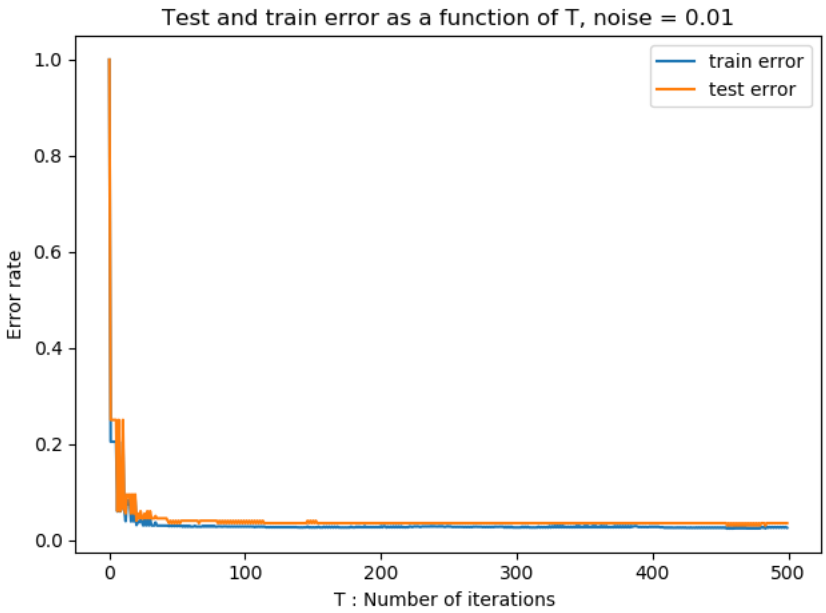
For $VCdim(\mathcal{H}) = d$.

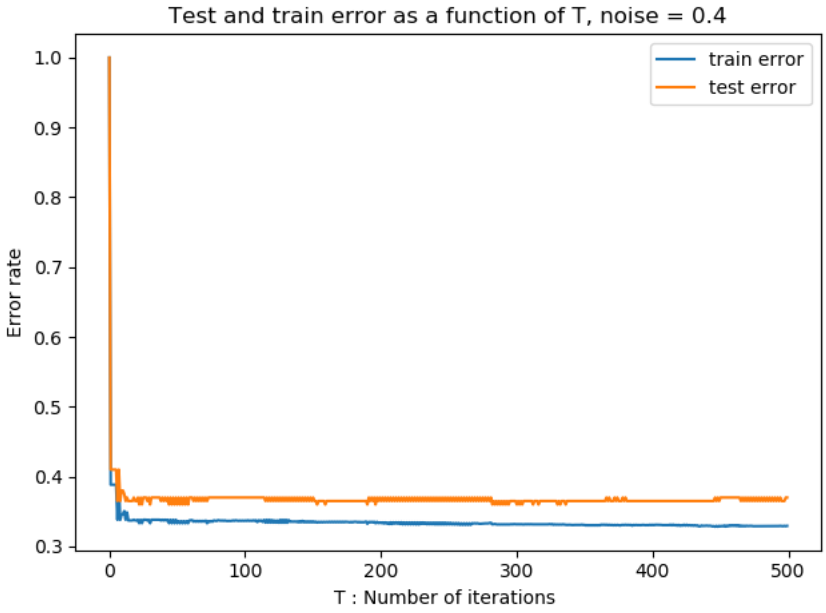
- If $d < m$ then $\tau(m)$ grows polynomially in m : $\tau(m) \leq \left(\frac{em}{d}\right)^d$
- If $m \leq d$ then $\tau(m)$ grows exponentially in m : $\tau(m) = 2^m$

Now we can characterize $VCdim(\mathcal{H}) = d = max\{m : \tau(m) = 2^m\}$ which means there is shattered sample of size d and there is no shattered sample of size greater then d.

10.Solution:

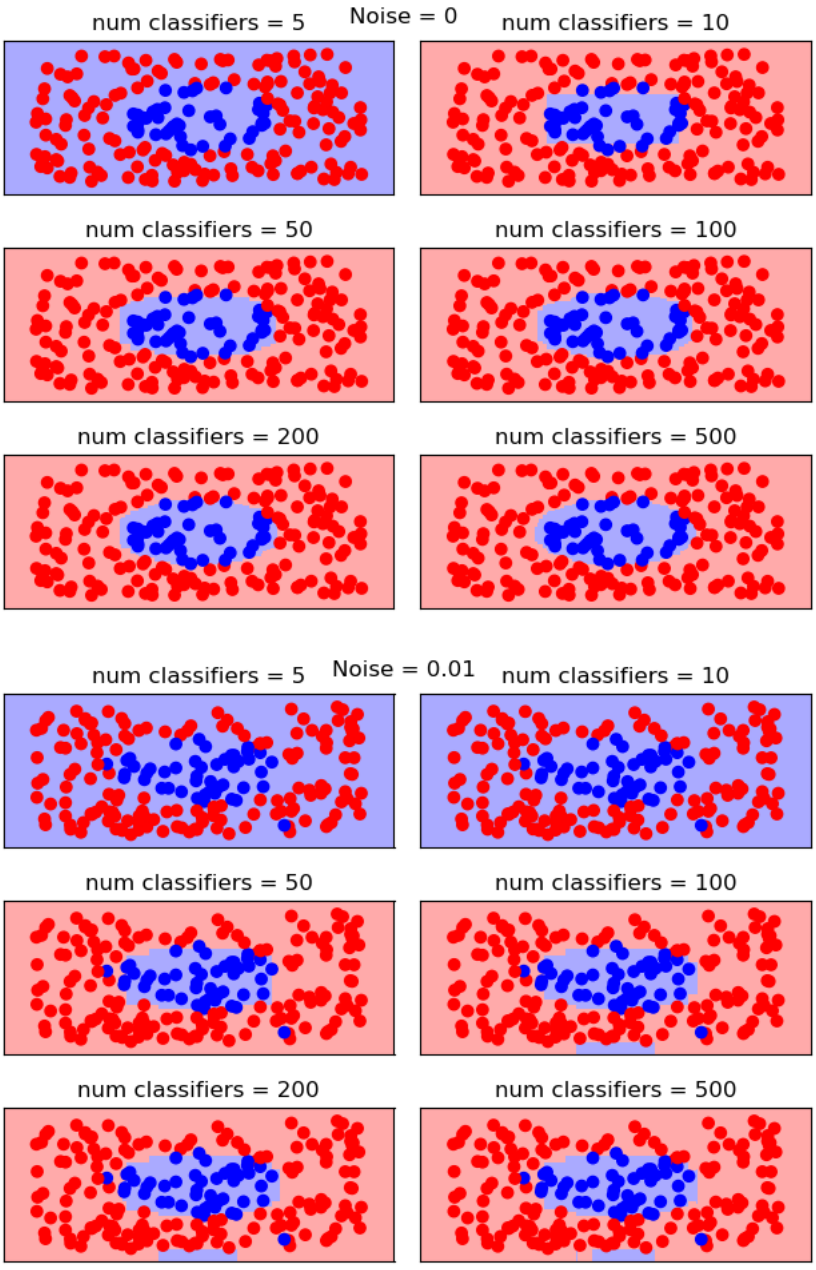


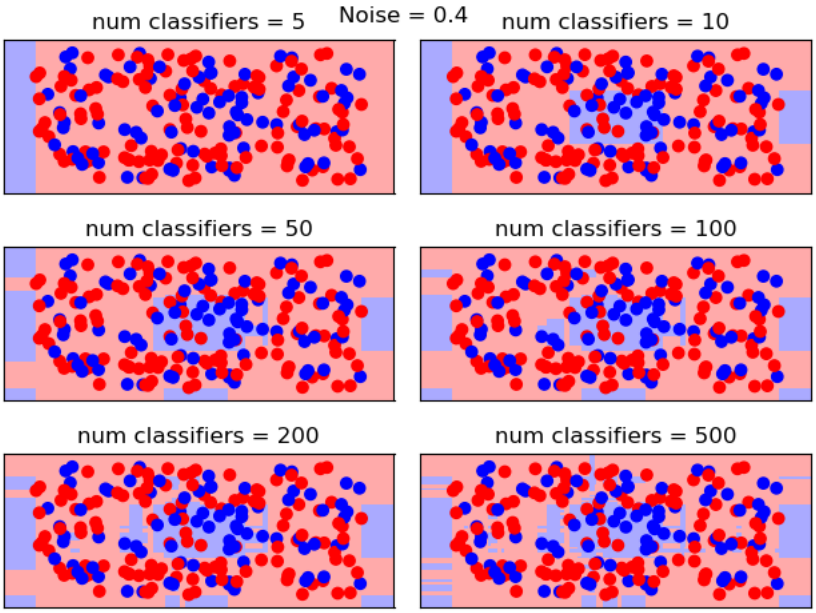




By looking at the graphs we can see that for data without noise both the bias and the variance are small and after a drastic change for small vlaues of T there is only slight change as T grows (namely for T greater or equal to 200). Moreover, we can see that AdaBoost is sensitive to label noise. when the noise grows to 0.01 we can see a slight increase in the variance, but when the noise grows to 0.4 we can see clearly an increase in the bias and the variance grows with T. It seems that AdaBoost is sensitive to label noise. An explanation for this might be that AdaBoost is fitting a classification model to an exponential loss function, so every one mislabeled point could have a very strong influence on the final model learned.

11.Solution:

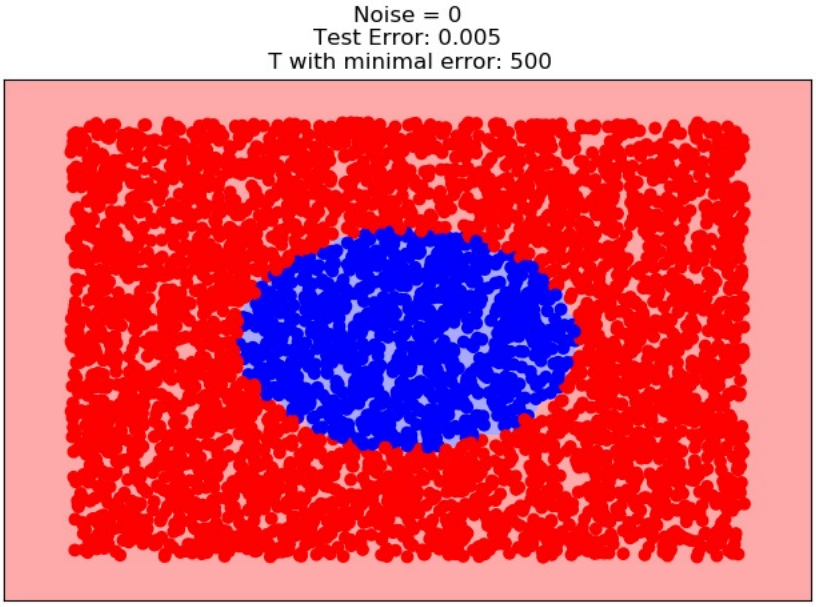

Noise = 0
num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500
Noise = 0.01
num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500

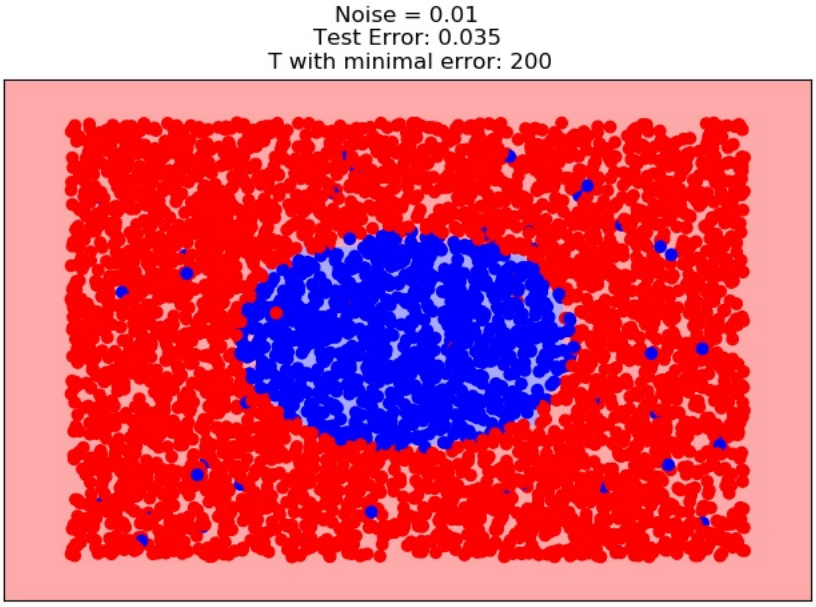


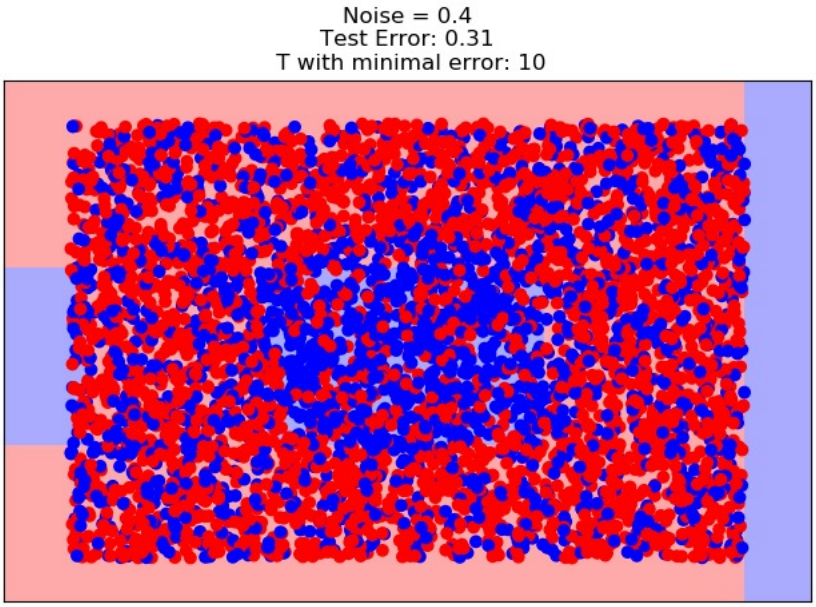
Here we can see a demostration to the previous explanation. For not noisy data we can see that there are great results for only 100 classifiers. For noise of 0.01 we can see an outlier and it's effect on the final model learned.

there is great improvment when the number of classifiers is 50 and then it's getting worse and seems to overfit and for noise of 0.4 we see very poor performance.

12.Solution:

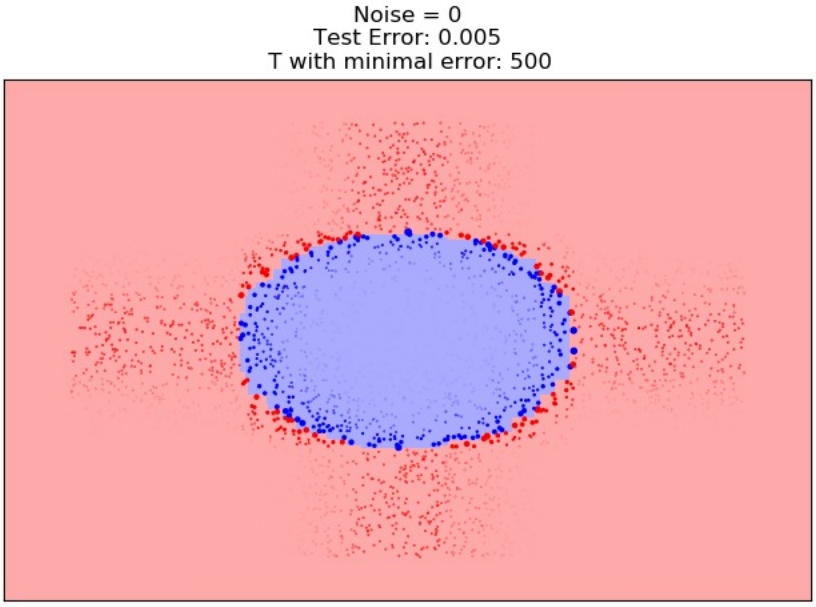

Noise = 0
Test Error: 0.005
T with minimal error: 500

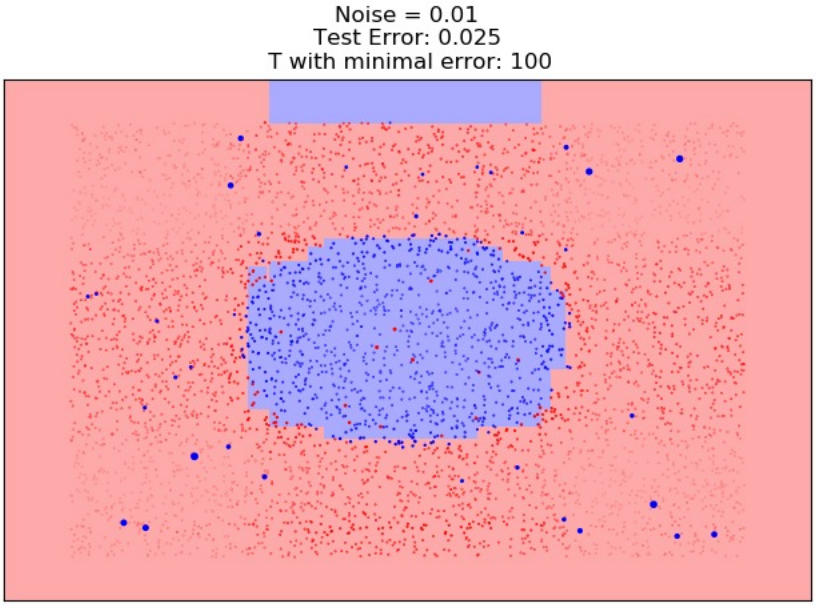

Noise = 0.01
Test Error: 0.035
T with minimal error: 200

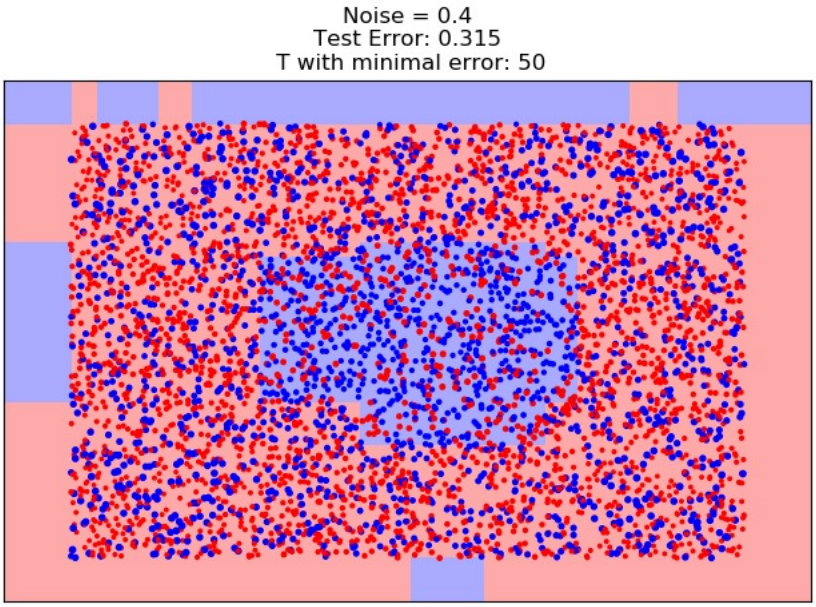


After running the algorithem few times it seems that typically when the noise grows the number of classifiers needed in order to gain minimum error decreasing. Again since every one mislabeled point could have a very strong influence. Since it is an additive model then at some point,for noisy date, it seems better to redunce the number of classifiers in order to reduce influence of the outliers.

13.Solution:


Noise = 0
Test Error: 0.005
T with minimal error: 500


Noise = 0.01
Test Error: 0.025
T with minimal error: 100



We can that for not noisy data to points around the circle get greater values which makes sense ,but since there are no outliers the the increasing of the wieghts around the circle seems to help and to get very well results.On the other hand the interesting part is a slight increase in noise (0.01), where only few outliers with large negative margin a strong effect and it can be seen that there surrounding suffers more from misclassification. Finally for noise of 0.4 the date is too noisy and too many points have large value of weights.