67577 | Introduction to Machine Learning | Exercise 4

Guy Lutsker 207029448

Question 1

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.

Solution :

We need to prove that:

$$\forall \varepsilon, \delta > 0\ \exists m(\varepsilon,\delta)\ \forall m \geq m(\varepsilon,\delta) : \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{\mathrm{m}}}[L_{\mathcal{D}}(A(S) \leq \varepsilon] \geq 1-\delta \Leftrightarrow \lim_{m\to\infty} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^{\mathrm{m}}}[L_{\mathcal{D}}(A(S))] = 0.$$

First we shall prove:

$$b \Rightarrow a :$$

We know that $\lim_{m\to\infty} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^{\mathrm{m}}}[L_{\mathcal{D}}(A(S))] = 0$ that means that $\forall \varepsilon^{'} > 0, \exists N, s.t\ \forall n > N\ \mathop{\mathbb{E}}_{S\sim\mathcal{D}^{\mathrm{n}}}[L_{\mathcal{D}}(A(S))] < \varepsilon^{'}$.
Now, we can choose $\varepsilon^{'} = \varepsilon \cdot \delta$,and derive that
there must exists n s.t $n > m(\varepsilon,\delta)$ s.t $\mathop{\mathbb{E}}_{S\sim\mathcal{D}^{\mathrm{n}}}[L_{\mathcal{D}}(A(S))] < \varepsilon^{'} = \varepsilon \cdot \delta$.
Now, $\forall \varepsilon, \delta > 0\ \exists m(\varepsilon,\delta)\ \forall m \geq m(\varepsilon,\delta) : \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{\mathrm{m}}}[L_{\mathcal{D}}(A(S) \geq \varepsilon] \overset{\text{Markov's inequallity}}{\leq} \frac{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^{\mathrm{m}}}[L_{\mathcal{D}}(A(S))]}{\varepsilon} \leq \delta$
And so we can conclude that $\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{\mathrm{m}}}[L_{\mathcal{D}}(A(S) \leq \varepsilon] \geq 1-\delta$.

Now, lets tend to the other side, lets prove that:

$$a \Rightarrow b :$$

We know that $\forall \varepsilon, \delta > 0 \; \exists m(\varepsilon, \delta) \; \forall m \geq m(\varepsilon, \delta) : \underset{S \sim \mathcal{D}^{\mathrm{m}}}{\mathbb{P}}[L_{\mathcal{D}}(A(S) \leq \varepsilon] \geq 1 - \delta$

Which can we written as statement $\kappa : \forall \varepsilon, \delta > 0 \; \exists m(\varepsilon, \delta) \; \forall m \geq m(\varepsilon, \delta) : \underset{S \sim \mathcal{D}^{\mathrm{m}}}{\mathbb{P}}[L_{\mathcal{D}}(A(S) > \varepsilon] < \delta.$

Now, we can see that $\forall m > m(\varepsilon, \delta)$ it hold that,

$$0 \overset{\text{must be}}{<} \underset{S \sim \mathcal{D}^{\mathrm{m}}}{\mathbb{E}}[L_{\mathcal{D}}(A(S))] \overset{\text{from definition}}{=}$$

$$\sum_{S \in \mathcal{D}^m} \mathbb{P}(S) \cdot L_{\mathcal{D}}(A(S))$$

$$\overset{\text{split into 2 sums}}{=} \sum_{\substack{S \in \mathcal{D}^m \\ s.t \; L_{\mathcal{D}}(A(S)) < \varepsilon}} \mathbb{P}(S) \cdot L_{\mathcal{D}}(A(S)) + \sum_{\substack{S \in \mathcal{D}^m \\ s.t \; L_{\mathcal{D}}(A(S)) \geq \varepsilon}} \mathbb{P}(S) \cdot L_{\mathcal{D}}(A(S))$$

$$\overset{\text{make } L_{\mathcal{D}} \text{ as large as possible}}{\leq} \sum_{\substack{S \in \mathcal{D}^m \\ s.t \; L_{\mathcal{D}}(A(S)) < \varepsilon}} \mathbb{P}(S) \cdot \varepsilon + \sum_{\substack{S \in \mathcal{D}^m \\ s.t \; L_{\mathcal{D}}(A(S)) \geq \varepsilon}} \mathbb{P}(S) \cdot 1$$

$$\overset{\text{Sigma addativity}}{=} \mathbb{P}\Big(\dot{\bigcup_{\substack{S \in \mathcal{D}^m \\ s.t \; L_{\mathcal{D}}(A(S)) < \varepsilon}}} S \Big) \cdot \varepsilon + \mathbb{P}\Big(\dot{\bigcup_{\substack{S \in \mathcal{D}^m \\ s.t \; L_{\mathcal{D}}(A(S)) > \varepsilon}}} S \Big) \overset{\text{definition}}{=} \mathbb{P}(L_{\mathcal{D}}(A(S)) < \varepsilon) \cdot \varepsilon + \mathbb{P}(L_{\mathcal{D}}(A(S)) > \varepsilon)$$

$$\overset{\text{according to statement } \kappa}{<} \mathbb{P}(L_{\mathcal{D}}(A(S)) < \varepsilon) \cdot \varepsilon + \delta \overset{\mathbb{P}(L_{\mathcal{D}}(A(S)) < \varepsilon) < 1}{<} \varepsilon + \delta$$

And so $\forall \varepsilon^{'} > 0$ we can find $m > m(\varepsilon, \delta)$ s.t:

$$\underset{S \sim \mathcal{D}^{\mathrm{m}}}{\mathbb{E}}[L_{\mathcal{D}}(A(S))] < \varepsilon + \delta = \varepsilon^{'}$$

and so by the limit definition we will get that indeed $\underset{m \to \infty}{lim} \underset{S \sim \mathcal{D}^{\mathrm{m}}}{\mathbb{E}}[L_{\mathcal{D}}(A(S))] = 0$

Question 2

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,

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

Solution :

Firstly we have got m samples of $S = (x_i, y_i)_{i=1}^m \in \mathbb{R}^2$.
My algorithm A will look for for the farthermost point in S from the origin.
Such a point will have the largest euclidean norm, and so its a simple maximization problem.
We shall denote our guess as the circle $C_{r_{estimated}}$ centered at the origin, and with radius $r_{estimated} = \max_{\substack{s \in S \\ s \text{ is labled with } y\,=\,1}} \|s\|$.

Define that $max\, \emptyset = 0$.
Now, lets denote the actual circle with $C_{r_{real}}$ centered at the origin, and with radius r_{real}.
Notice that $r_{estimated} \leq r_{real}$. And so we can only miss-classify points labeled 1 in $MISS = C_{real}/C_{estimated}$.
We need to prove that $\forall \varepsilon, \delta, \mathcal{D} \sim \mathbb{R}^2$, $\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{\mathcal{D}}(A(S) \leq \varepsilon] \geq 1 - \delta$.
To do that we will give the area $MISS$ probability ε over $\mathcal{D} \sim \mathbb{R}^2$.
And so we get that the probability of the m samples to fall in $\mathbb{R}^2/MISS$ is $(1 - \varepsilon)^m$.
And so we would like for $(1 - \varepsilon)^m < \delta$, we know that $(1 - \varepsilon)^m \leq e^{-m\varepsilon} \rightarrow m \geq \frac{log(1/\delta)}{\varepsilon}$.
Which means that $m_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{log(1/\delta)}{\varepsilon}$.

Question 3

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.

Solution :

$\mathcal{H}_{con} = \{h : 2^{2d} \to \{0,1\} | h \text{ is a boolean conjunction}\}$
Let us recall that $VC\mathcal{H}_{con} = max\{|C| \; | C \in \mathcal{X}, |\mathcal{H}_{\mathcal{C}}| = 2^{|C|}\}$
Statement: The VC-dimention of $\mathcal{H}_{con}$ is d!
Proof:
Let us firstly show that $VC\mathcal{H}_{cal} \geq d$:
Let $C = \{x_1 ... x_d | x_i = e_i\}$ then for any $\{y_i \in \{0,1\}^d | i \in [d]\}$
lets define $h(x_j) = \bigwedge_{y_i=0} \overline{x_{ji}}$. when the first coordinate is which vector, and the second is the position within.
If $\exists j$ s.t $h(x_j) = 1$ then we will get a conjugate on `1`'s only, because $x_{ji} = 0$ and so for any j $h(x_j) = y_j = 1$.
If $\exists j$ s.t $h(x_j) = 0$ then we will get a conjugate with at least one `0`, because $x_{ji} = 1$ and so $h(x_j) = y_j = 0$.
Notice that there isn't a j such that $y_j = 0$ we will receive an empty conjugate
which it True by default, and we will get that for any j $h(x_j) = y_j = 1$.
And what we get, is that for any group C of size d we can define a conjugate
which will get to the entire $Image - \{0,1\}$
and so $|\mathcal{H}_{\mathcal{C}}| = 2^{|C|}$ which means C shatters H_C.

Now, let us show that $VC\mathcal{H}_{con} \leq d$:
Let assume towards a contradiction that $\exists C = \{x_i | i \in [d+1]\}$ that shatters $\mathcal{H}_{con}$.
This means that for any $\{y_i \in \{0,1\}^d | i \in [d+1]\}$ s.t $|\{h_i | i \in [d+1], h_i : \{0,1\}^{d+1} \to \{0,1\}| = 2^{d+1}$.
It also means that we can write this as $h_i(x_j) = \delta_{Kronecker}(i,j)$
From the way our Hypotheses class is build we know that there exists a literal that negates x_j and does not negate any other variable in C , and so there exists a coordinate within $x_i \to x_{ij}$ such that is it different then all of the other variables in C .
Now, it follows that each h_i has such a property, and so $\forall x_i \exists j$ s.t x_{ij} is different then all the variables in C
and so we can conclude from the Pigeonhole principle, that we a contradiction on our hands,
because there are $d+1$ in the group, yet there are only d coordinates.
And so we get that a group of size $d+1$ cannot shatter $\mathcal{H}_{con}$, and so we get that $d \leq VC\mathcal{H}_{con} < d+1$
Which means that $VC\mathcal{H}_{con} = d$, as required.

Question 4

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 $m_{\mathcal{H}}(\epsilon, \delta) \leqslant m^{UC}(\epsilon/2, \delta)$.

Solution :

Assume that $\mathcal{H}$ has the uniform convergence property with the function $m_{\mathcal{H}}^{UC}$.
Now according to definition a hypothesis class $\mathcal{H}$ is Agnostic-PAC learnable if:

$$\forall \mathcal{D}, \varepsilon, \delta \; \exists m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}, s.t \; \forall m > m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2}, \delta) \text{ It Holds That:}$$

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon] \geq 1 - \delta$$

Proof:
Let $m > m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2}, \delta)$ Now we know that if S is $\frac{\varepsilon}{2}$ representative we know that:

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \overset{definition}{\leq} L_S(\underset{h' \in \mathcal{H}}{argmin} L_{\mathcal{D}}(h')) + \varepsilon$$

And so because we assumed that $\mathcal{H}$ has the uniform convergence property with the function $m_{\mathcal{H}}^{UC}$
We get that:

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon] \overset{\substack{\text{Because Of} \\ \text{The Definition} \\ \text{Of } \varepsilon \text{ Representative}}}{\geq} \mathbb{P}_{\mathcal{D}}[S \text{ Is } \frac{\varepsilon}{2} \text{ Representative}] \geq 1 - \delta$$

Which means that we get that $\forall m > m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2}, \delta) : \underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon] \geq 1 - \delta$
Which is the definition of $\mathcal{H}$ to be Agnostic-PAC learnable.
as required:)

Question 7

7. **of VC-Dimension** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two classes for binary classification, such that $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Show that $\mathrm{VC}\,\mathcal{H}_1 \leqslant \mathrm{VC}\,\mathcal{H}_2$.

Solution :

Let $\mathcal{H}_1 \subset \mathcal{H}_2$, Now notice that if a sample S of size m shatters $\mathcal{H}_1$ we get that S shatters $\mathcal{H}_2$ as well simply because of the subset relationship between the groups $\to$ because $\mathcal{H}_1 \subset \mathcal{H}_2$.
And so we get for free that $VC\mathcal{H}_1 \leq VC\mathcal{H}_2$.
Notice that this still holds if $VC\mathcal{H}_1 = \infty$.

Question 8

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 $\tau_{\mathcal{H}}$.

Solution :

Given an m, $\tau_{\mathcal{H}}(m)$ measures the maximal cardinal number of a hypotheses class that a samples with size m can "use".

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

Solution :

Suppose $VCdim(\mathcal{H}) = \infty$, then it holds that for any sample C of size m will shatter $\mathcal{H}$ and so we know that $|\mathcal{H}_C| = |2^C| = 2^m$.
and so we get that $\tau_{\mathcal{H}}(m) = 2^m$.

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

Solution :

Let $VCdim(\mathcal{H}) = d$, meaning we know there's a sample C of size d such that $|\mathcal{H}_C| = 2^d$.
And so, we can see that $\forall m \leq d, \forall C^{'}\, s.t\, |C^{'}| = m \to C^{'}$ shatters $\mathcal{H}$.
And so we can conclude that $\tau_{\mathcal{H}}(m) = 2^m$

(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 $C \subset \mathcal{X}$,

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

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

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

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

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

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

$Solution:$

i.
We will prove the claim using induction on the size of C - m:
Base: $m = 0 \rightarrow C = \emptyset$ and so $|\mathcal{H}_C| = 0$.
In addition $\forall B \subset C$ it holds that B must be equal to $\emptyset$.
And so it does indeed hold that:

$$0 = |\mathcal{H}_C| \leq |\{B \subset C : \mathcal{H} \; shutters \; B\}| = 0$$

Step: Suppose the claim is true for any group with size less then m.
Let us prove the claim for m:
Let C be a group of size m, and also let $c \in C$ be an element of C, finally let $B = C/c$
and notice that $|B| = m - 1$.
Now, using the induction hypothesis we know that $|\mathcal{H}_B| \leq |\{A \subset B : \mathcal{H} \; shutters \; A\}|$
In addition we know that c can be tagged as wither ± 1 and so $|\mathcal{H}_C| \leq 2 \cdot |\mathcal{H}_B|$. And finally:

$$\mathcal{H}_C| \leq 2 \cdot |\mathcal{H}_B| \leq 2 \cdot |\{B \subset C : \mathcal{H} \; shutters \; B\}|$$

ii.
The meaning of the inequality is that for any group C the cardinal number of the class $\mathcal{H}_C$ is lesser of equal to the number of all subsets $B \subset C$ that shatter $\mathcal{H}$.

iii.
Let C be of finite size, such that $|C| = m > d$.
Then the number of all subsets of C of size at max d is $\sum_{k=0}^{d} \binom{m}{k}$.
We know that $VCdim(\mathcal{H}) = d$, and so the group $G = \{B \subset C : \mathcal{H} \; shutters \; B\}$
uphold that for any element $E \in G$ we know that $|E| \leq d$ and so :

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

iv.
Let $VCdim(\mathcal{H}) = d < m$, and let B be a group of size m such that $\tau_{\mathcal{H}}(m) = |\mathcal{H}_B|$.
Now, notice that from i,ii we get that:

$$\tau_{\mathcal{H}}(m) = |\mathcal{H}_B| \leq |\{A \subset B : \mathcal{H} \; shutters \; A\}| \leq \sum_{k=0}^{d} \binom{m}{k} \leq (\frac{e \cdot m}{d})^d$$

(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?

$$Solution:$$

Let $d = m$, then the inequality still holds, according the prove above.
Suppose we have a group C of size m that shatters $\mathcal{H}$ then we know that $\tau_{\mathcal{H}}(m) = 2^m$
and so:

$$(\frac{e \cdot m}{d})^d = (\frac{e \cdot d}{d})^d = e^d > 2^d$$

And so, all still holds.

(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 $VCdim(\mathcal{H})$?

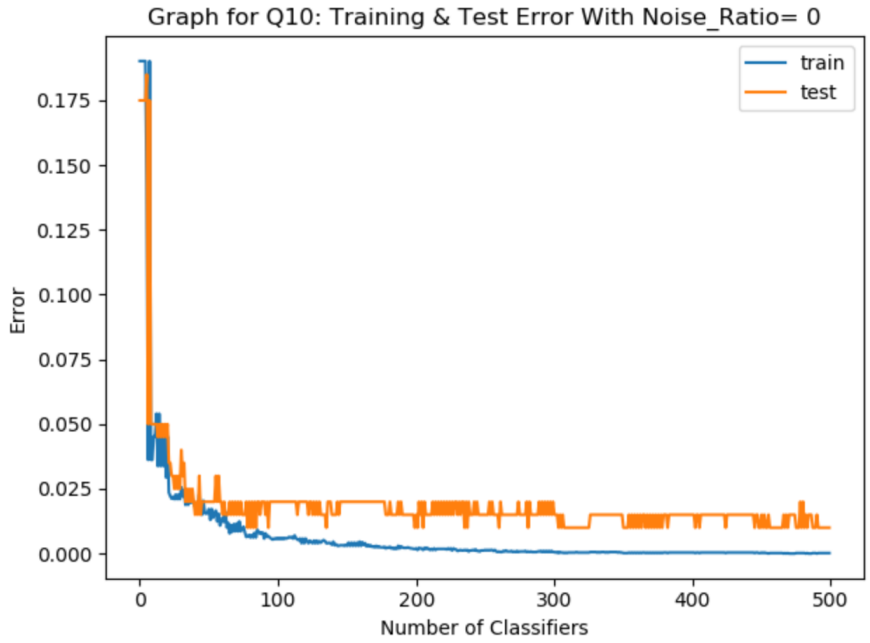
$$Solution:$$

We know that for $m \leq VCdim(\mathcal{H})$ it holds that $\tau_{\mathcal{H}}(m) = 2^m$ as is increases exponentially.
Yet if $m > VCdim(\mathcal{H})$ we saw that $\tau_{\mathcal{H}}(m) \leq (\frac{e \cdot m}{d})^d$ meaning it grows slower- polynomialy.
Meaning we can define $VCdim(\mathcal{H})$ thusly:

$$VCdim(\mathcal{H}) = max\{m \in \mathbb{N} | \tau_{\mathcal{H}}(m) = 2^m\}$$

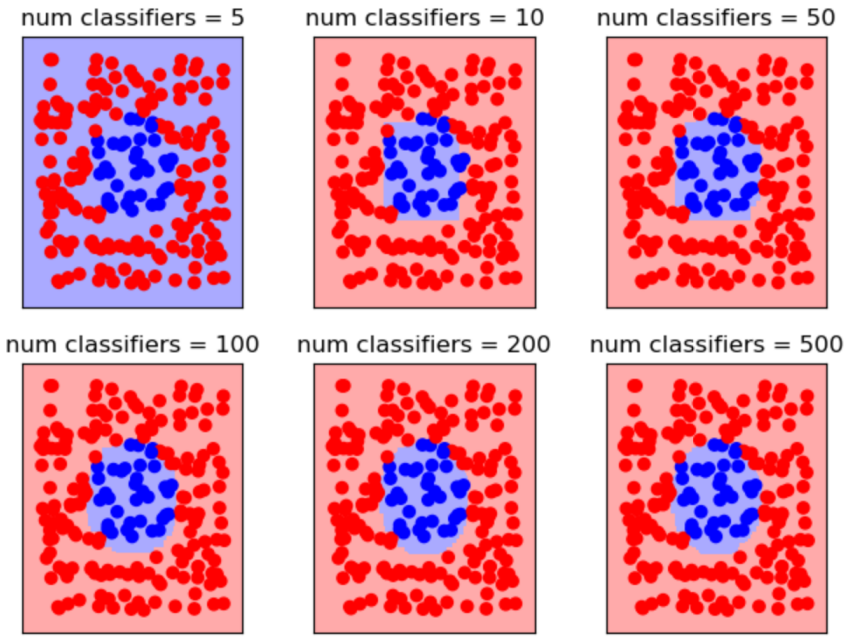
$$Coding \; Part:$$

Question 10



We see here that both the training error, and the test error drops as the number of classifiers rises and they both hit a platou. Also we see that the test error is above the training error.
Here we had 0 noise and so there wasn't noise to overffit to- we will see that in the next questions.

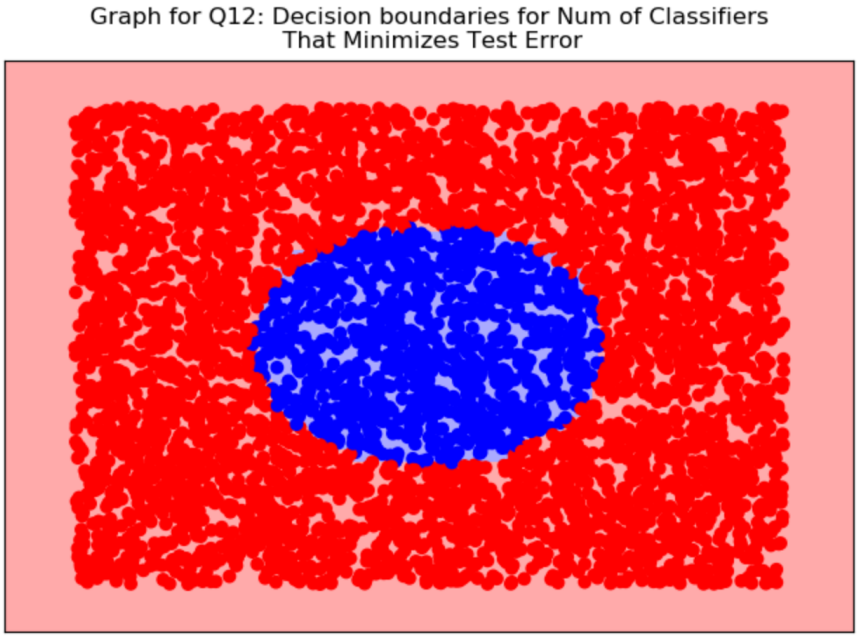
Question 11



We see that using only 5 classifiers yielded a bad trivial classification of the data(all blue).
And as the number if classifiers rises we see a "tighter" classification, which is growing in accuracy.
Until, the accuracy platou's as well.

Question 12

Graph:

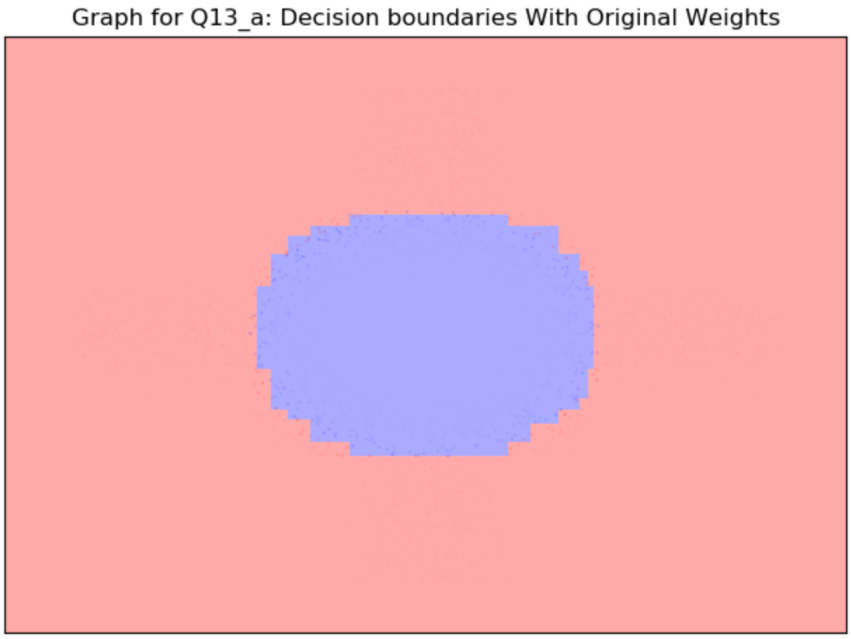


We can see the graph with the minimal test error.
The number of classifiers that yielded this result was 227, and the error was 0

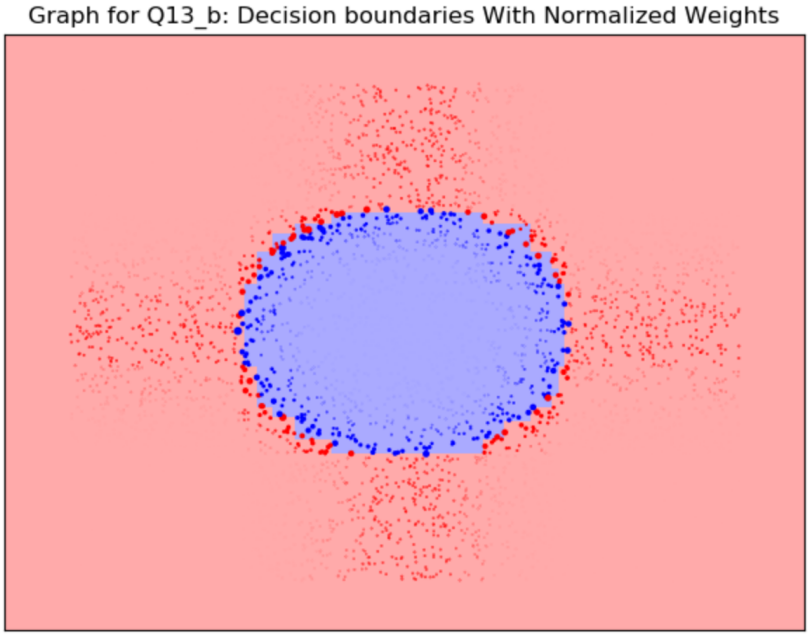
Question 13

Graphs:

Original Weights: The plot of the data with the weight from the algorithm - D^T.



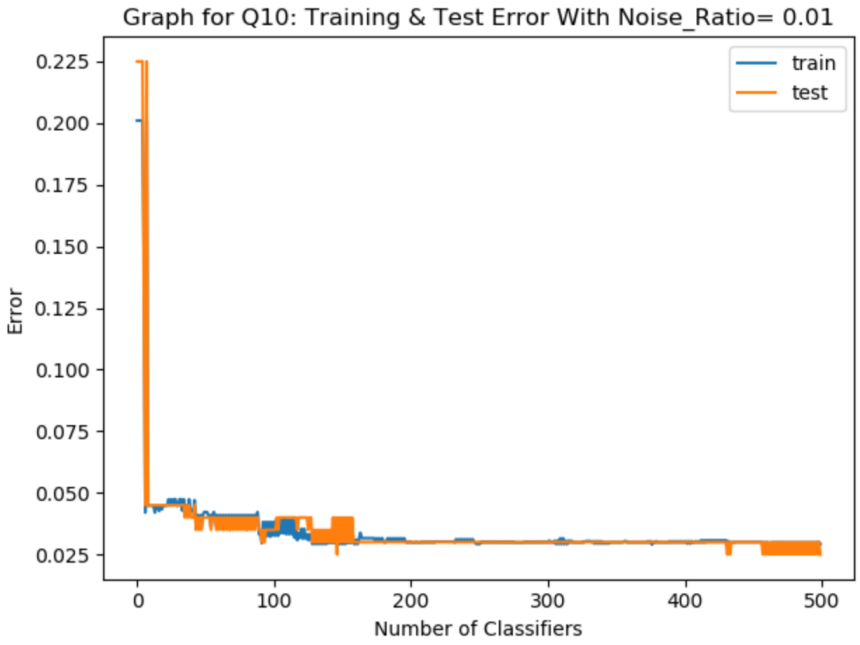
Normalized Weights:



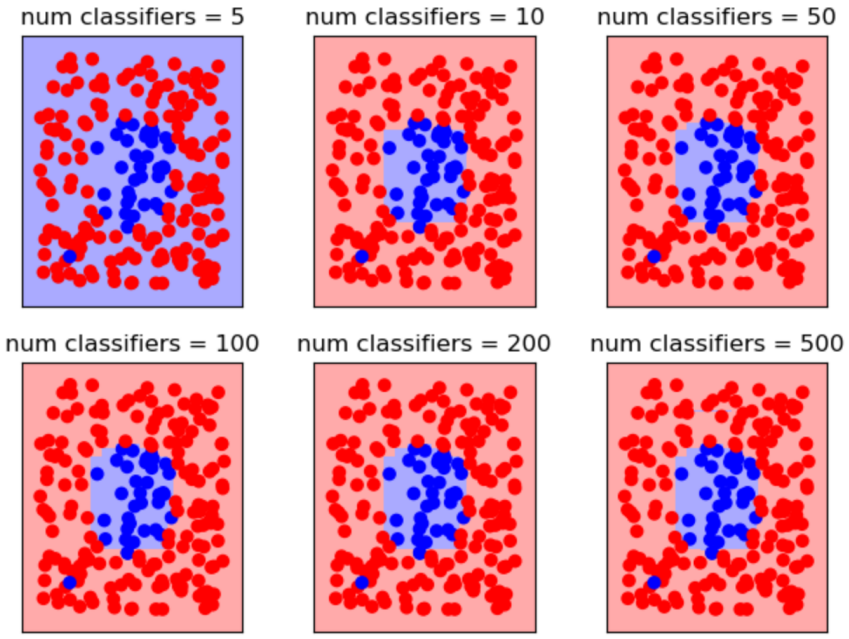
Explanation: The AdaBoost algorithm will re-weight the data point in each iteration, assigning a larger weight to samples we got wrong in the last iteration. In this plot we see the weighted points After the last iteration.
Now, in the first graph we see the raw implementation, and the point are hard to see, only the border line is visible. Yet if we normalize the point, as is done in the second graph we can see the points.
What we can see is that the largest points are the ones closest to the border line, meaning these are the points we got wrong in the latest iterations, and so it makes sense that the border was decided next to them.

Question 14

Graphs for Noise = 0.01:

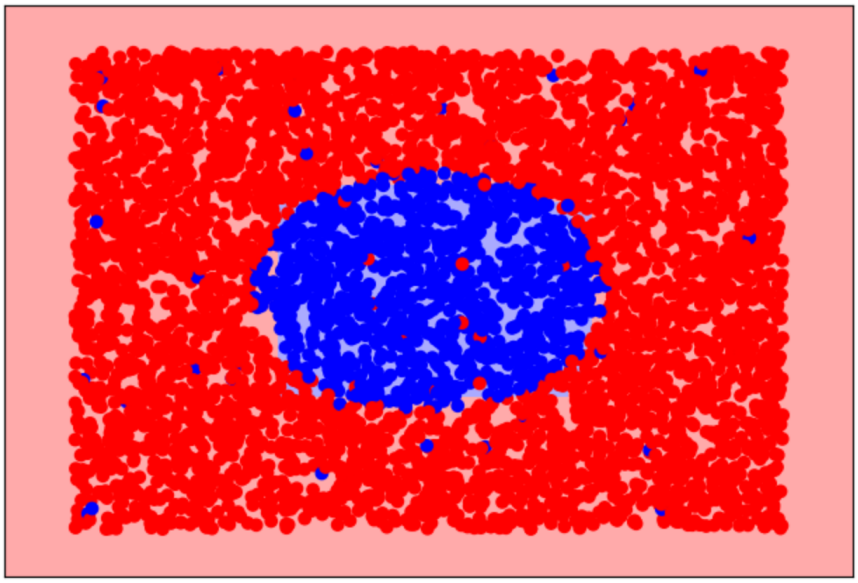


We can see similar results to the graph we saw in the zero noise graph, the main differences are that the platou in the error was in 0.025 with 25 classifiers.
Here we are also yet to see much of a difference in terms of the bias-variance traid off.
Spoilers - in the next noise level we will see the difference.



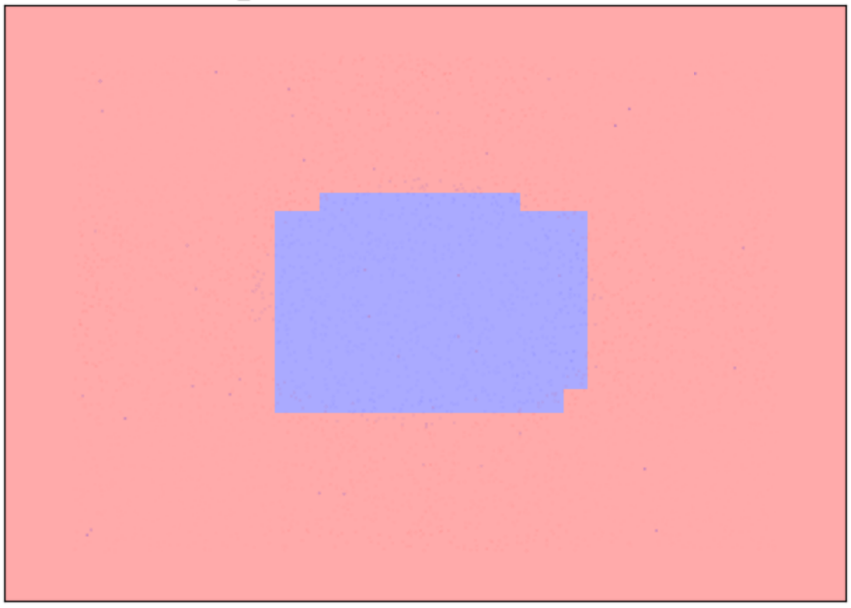
We can start to see the noise "confusing" our model, and making it to start making mistakes.
But still, the model is doing not too badly, and its boundary is reasonable.

Graph for Q12: Decision boundaries for Num of Classifiers That Minimizes Test Error

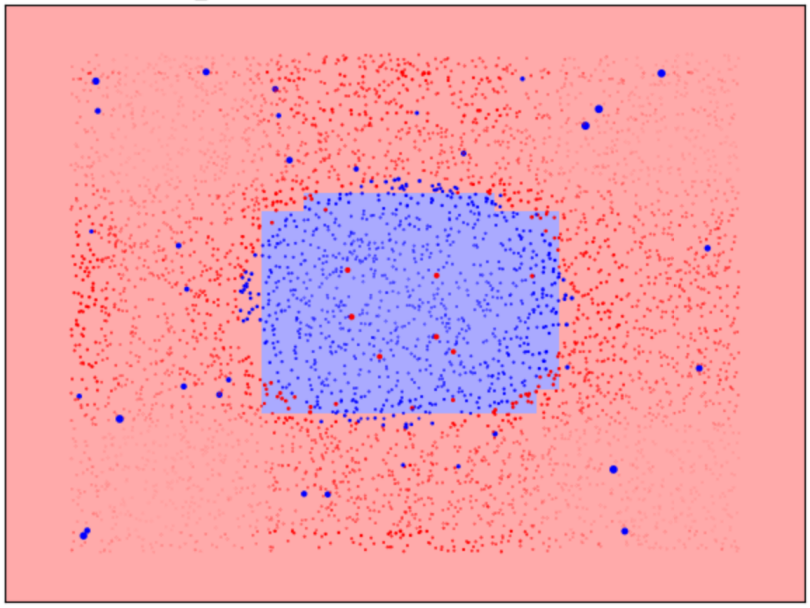


Here we can see that the noise is making it harder for the model to correctly decide the boundary and there are indeed mistakes it makes, but overall it made an accurate boundary.

Graph for Q13_a: Decision boundaries With Original Weights

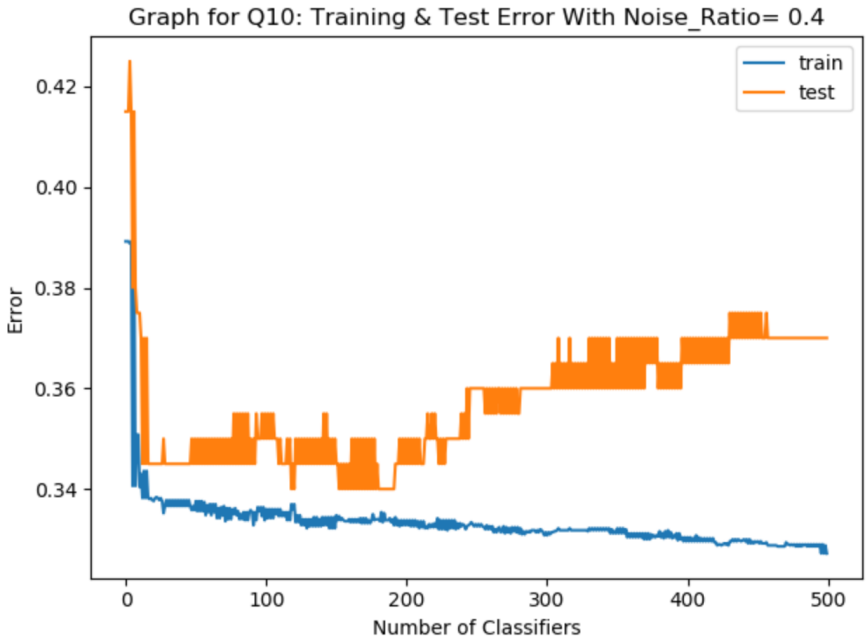


Graph for Q13_b: Decision boundaries With Normalized Weights

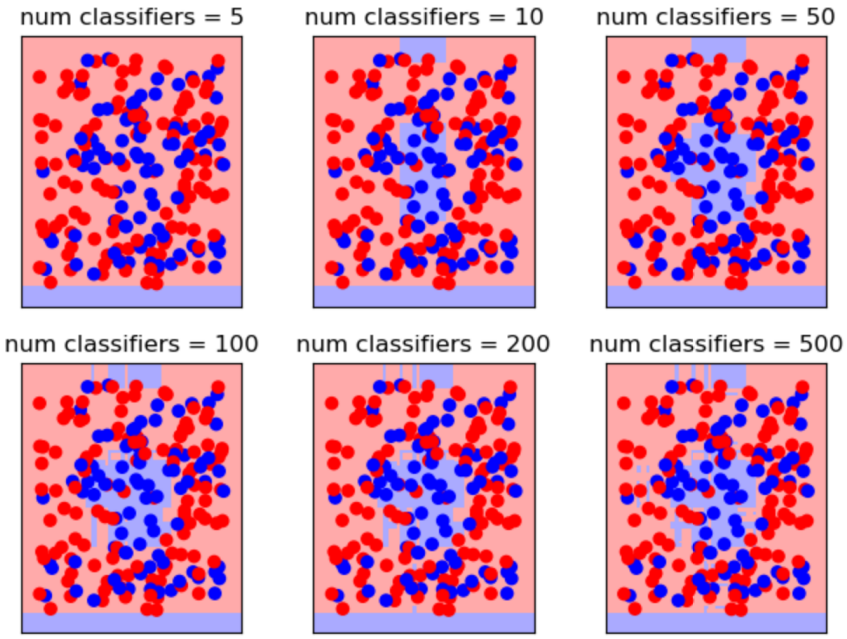


We can see the weighted data points, and now its harder to assign meaning to the size of the point in a meaningful way, those are simple noise points that the model got wrong.
We can also see the decision boundary becoming more “boxy” and less round.

Graphs for Noise = 0.4:

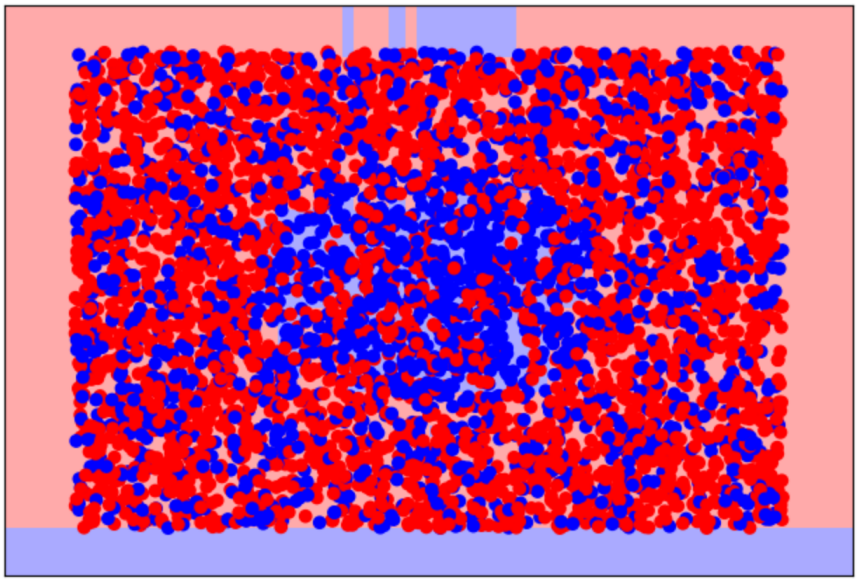


Here we can finally see the variance-bias trade off in action.
The data set was filled with noise, and so as we added more and more classifiers we see the training error falling as it has in previous attempts.
Yet the test error is more interesting, firstly we see a dip in the error, but quickly we can start to see a rise in error, what could that be? Well the variance-bias trade-off might have something to do with it ;)
As we trained our model more and more, it started to try and fit to every singly noise data -
and that a recipe for disaster, we can start to see our model overffiting the data, and the test data error rising and rising.
The min error that was achieved was in 0.325 with 12 classifiers.



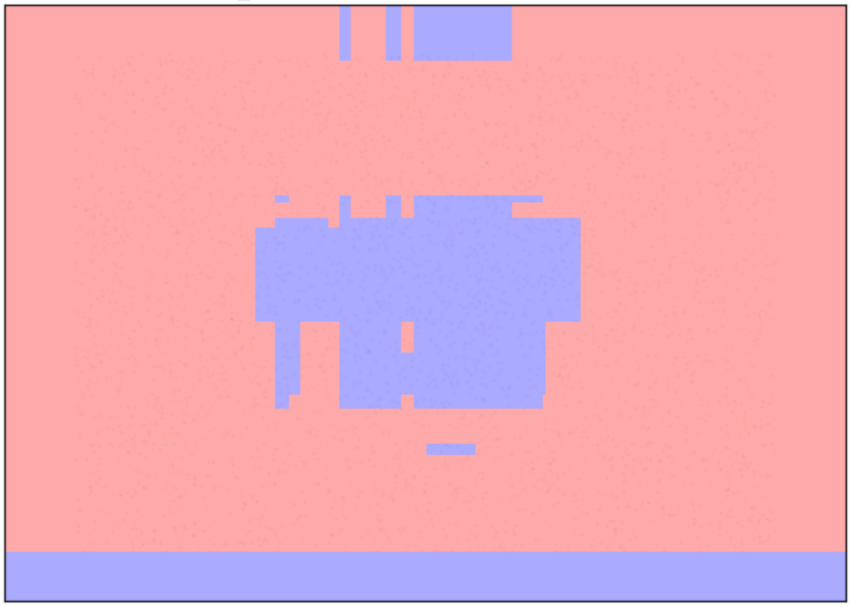
As the noise level was so high we cant really see a trend in the number of classifiers.
All of them seem to do poorly.

Graph for Q12: Decision boundaries for Num of Classifiers That Minimizes Test Error

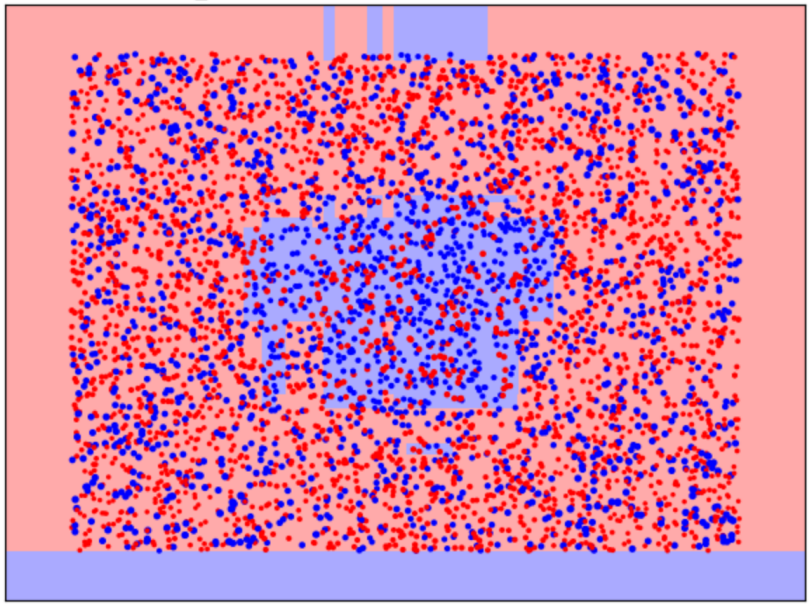


Here we can also see that the noise level was way too high and the model couldn't really decide a logical decision boundary. It tried its best, and it does resemble the previous graphs, but its not accurate.

Graph for Q13_a: Decision boundaries With Original Weights



Graph for Q13_b: Decision boundaries With Normalized Weights



The noise level was way too high and the model couldn't really decide
a logical decision boundary.
Yet one interesting thing to note is that the weights of the points are kind of uniform, which might suggest that there were so many mis-classified points that they all got a similar weight.