Introduction to Machine Learning (67577)
Exercise IV – Classification
Hadar Sharvit – 208287599

Theoretical Questions

PAC learnability

1. Prove that $\forall\ \varepsilon, \delta > 0\ \exists\ m(\varepsilon, \delta)$ s.t $\forall\ m \geq m(\varepsilon, \delta)$: $P_{S\sim D^m}(L_D(A(S) \leq \varepsilon) \geq 1 - \delta$ is equivalent to $\lim_{m\to\infty} E_{S\sim D^m}\left(L_D\big(A(S)\big)\right) = 0$:

   Using *Markov's Inequality* $P(X \geq a) \leq E(X)/a$

   $$P_{S\sim D^m}(L_D(A(S) \leq \varepsilon) = 1 - P_{S\sim D^m}(L_D(A(S) \geq \varepsilon) \geq 1 - \frac{E_{S\sim D^m}\left(L_D\big(A(S)\big)\right)}{\varepsilon}$$

   since $\lim_{m\to\infty} \frac{E_{S\sim D^m}\left(L_D(A(S))\right)}{\varepsilon} = 0$ then $\lim_{m\to\infty} P_{S\sim D^m}(L_D(A(S) \leq \varepsilon) \geq 1$, and therefore, there exists some m for which $P_{S\sim D^m}(L_D(A(S) \leq \varepsilon)$ is arbitrarily close to 1, which is equivalent to writing $P_{S\sim D^m}(L_D(A(S) \leq \varepsilon) \geq 1 - \delta$, as needed.

2. Prove that $H = \{h_r(x) = 1[\big|\big|x\big|\big|_2 \leq r\colon r \in \mathbb{R}^+\}]$ is PAC learnable with $m_H(\varepsilon, \delta) \leq \frac{\log(1/\delta)}{\varepsilon}$:

   Similar to what we have seen in class, we will use the following learning algorithm-
   Given points in $X = \mathbb{R}^2$, return the circle with the smallest area that includes all of the $+1$ samples and none of the 0 samples. If there are no such circles, return $\emptyset$.
   For convenience's sake, we will denote a ring with inner radius r_1 and outer radius r_2 as $R(r_1, r_2)$, and the radius of some circle C as r_C.
   Since the tightest fit circle C' is always contained in the true circle C, the error can only come from positive samples $\in R(r_{C'}, r_C)$, and If we are able to guarantee that the weight under D of $R(r_{C'}, r_C)$ is at most ε, then the error of C' is at most ε.
   Suppose the ring weighing exactly ε is $R(r_{C'}, r_{C''})$ for some circle C''. now some points to consider:
   - The ring $R(r_{C'}, r_C)$ has weight exceeding ε iff $C' \subseteq C$.
   - $C \subseteq C'$ iff there are no points of S in $R(r_{C'}, r_C)$.
   - if some point p is in $R(r_{C'}, r_C)$, then the algorithm would have made C' include p.
   - The probability of sampling a point that misses $R(r_{C'}, r_C)$ is $1 - \varepsilon$ (and for m samples if would be $(1 - \varepsilon)^m$)

   From those we understand that if we choose m such that $(1 - \varepsilon)^m \leq \delta$, then with probability $1 - \delta$ over the m random samples the weight of the error is at most ε. Now, since $(1 - \varepsilon)^m \leq e^{-m\varepsilon}$, let us choose $\delta \geq e^{-m\varepsilon}$, thus $m_H(\varepsilon, \delta) \leq \frac{\log(1/\delta)}{\varepsilon}$

VC-dimension

3. For finite class $H = \{h_1, \ldots, h_N\}$: $VCdim(H) \leq \lfloor \log_2(|H|) \rfloor$:

A set $C \subset \{h_1, \dots, h_N\}$ is of maximal size N, therefore $VCdim(H) = \max\{|C| : C \subset \{h_1, \dots, h_N\} \; and \; H_C = 2^{|C|}\} \leq N$.Now, since H is all function $\{h_1, \dots, h_N\} \to \{\pm 1\}$, its size is $|H| = 2^N$, and since $\log_2 |H| = N$ we have $VCdim(H) \leq \lfloor \log_2(|H|) \rfloor$

4. Find $VCdim$ of $H = \{h_I : \{0,1\}^n \to \{0,1\} : \; h_I(x) = (\sum_{i \in I} x_i) mod \; 2, I \subseteq [n]\}$:

   will show that $VCdim(H) = n$
   that is, $\exists$subset $C \subset X$ of size n for which we can generate both 0 and 1. We will choose $C = \{e_1, e_2, \dots e_n\}$ where e_i is a unit vector with 1 in the $i'th$ row and 0 otherwise. Given n labels $y \in \{0,1\}^n$, $h_{[n]}$ can generate them all because when summing over all subsets $[n]$ in $\sum_{i \in [n]} x_i$, we can get any permutation of $\{0,1\}^n$. in other words, by definition of h_I, using the singleton $I = \{i\}$ for some $i \in \{1,2, \dots, n\}$ (see below) we have $h_I = \begin{cases} 1, x_i = 1 \\ 0, \; x_i = 0 \end{cases}$, and from here for any union of singletons $\bigcup_{i=1}^{k} \{i\} = [k]$, the returned label is any permutation of k ones and $n - k$ zeros, such that we will see a 1 in index i if the singleton $\{i\}$ is a part of the union.

$$X = \{0,1\}^n$$
$$C = \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\} \subset X$$
$$I \subseteq \{1,2, \dots, n\} := [n]$$

   $h_I(x \in C) = (\sum_{i \in I} x_i) mod \; 2$ generates every label:

$$I = \emptyset: \; h_\emptyset \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = h_\emptyset \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \cdots h_\emptyset \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = 0 \to label \; 00 \dots 0$$
$$I = \{1\}: \; h_{\{1\}} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 1, h_{\{1\}} \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} = 0 \dots h_\emptyset \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = 0 \to label \; 10 \dots 0$$
$$\dots$$
$$I = \{n\}: \; h_{\{1\}} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0, h_{\{1\}} \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} = 0 \dots h_\emptyset \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = 1 \to label \; 00 \dots 1$$

   And every non singleton is a combination of a set of singleton labels, which generates all the remaining labels.
   This can also be explained as followed: for every $c_i \in C$, to generate $h_I(c_i) = 0$ choose I which does not contain i, and for $h_I(c_i) = 1$ choose I which contains i.
   This in turn shows that $|C| = n$ is shattered by H, so $VCdim(H) \geq n$, but from section (3) we have $VCdim(H) \leq \log_2 |H| = n$, therefore $VCdim(H) = n$.

5. We will show that $VCdim(H_{k \; intervals}) = 2k$, which is ∞ if k is not bounded:

   - $VCdim(H_{k \; intervals}) \geq 2k$: any label of size $2k$ can be generated using k intervals. the worst-case scenario for a label, in terms of the number of intervals we need to use, is

when there are no adjacent 1's. This is because for any repetition of 1's we can simply use one interval to include them all. If so, for k non-adjacent 1s (that are separated with 0's), we must have k intervals (interval for each 1). Furthermore, a label of size $2k$ has at most k non-adjacent 1's, because between two non-adjacent 1's there is at least one 0.

- $VCdim(H_{k\ intervals}) \leq 2k$: assume on a way of contradiction that there exists a group of size $2k+1$ that can be shattered by $H_{k\ intervals}$. If so, the label with $k+1$ non-adjacent 1's and k non-adjacent 0's can be generated. This is a contradiction because $k+1$ non-adjacent 1's can only be generated using $k+1$ intervals (as previously discussed), and we use only k.

6. We will show $VCdim(H_{con}) = d$, where $H_{con} = \left\{ h: \{0,1\}^d \to \{0,1\}: \begin{array}{r} h(\boldsymbol{x}) = \bigwedge_{i=1}^{d} \ell_i \\ \ell_i \in \{x_i, \overline{x_i}\} \\ i \in [d] \end{array} \right\}$.
   - $VCdim(H_{con}) \leq d$: assume by contradiction $VCdim(H_{co}\ ) = d+1$, so a set $C = \{x_1, \dots x_{d+1}\}$ is shattered by H_{con}. Consider the hypothesis
   $$h_i(x_j) = \left\{ \begin{array}{ll} 0, & i = j \\ 1, & i \neq j \end{array} \right\}, \quad i,j \in \{1,2,\dots,d+1\}$$
   Since there are d variables, and C contains $d+1$ of them, there exists two literals, call them ℓ_1 and ℓ_2, that represent the same x_i (either x_i or $\overline{x_i}$).
   If $\ell_1 = \ell_2$: then $h_1(\ell_1) = 1$ but $h_1(\ell_2) = 0$. This is a contradiction because $\ell_1 = \ell_2$
   If $\ell_1 \neq \ell_2$: suppose w.l.o.g that $\ell_1 = x_3\ and\ \ell_2 = \overline{x_3}$. By definition of h_i we have $h_3(\ell_1) = h_3(\ell_2) = 1$, though since $x_j = x_3$ then $h_3(x_3) = 0$. This is a contradiction.
   - $VCdim(H_{con}) \geq d$: consider the group of unit vectors $C = \{e_1, \dots, e_d\}$. For some label vector $(y_1, \dots, y_d) \in \{0,1\}^d$ we can choose the following hypothesis: $h(x) = \bigwedge\limits_{i:y_i=0} \overline{x_i}$
   For such h we have $h(e_i) = y_i$, therefore $VCdim(H) \geq d$

Agnostic-PAC

7. H has uniform convergence property with $m_H^{UC}: (0,1)^2 \to \mathbb{N} \Rightarrow H$ is agnostic-PAC learnable with sample complexity $m_H(\varepsilon, \delta) \leq m_H^{UC}(\varepsilon/2, \delta)$:
   H has uniform convergence property with $m_H^{UC}: (0,1)^2 \to \mathbb{N} \Rightarrow \forall(\varepsilon,\delta) \in (0,1)\ \forall D$ on $X \times Y$ $D^m(\{S \in (X \times Y)^m : S \text{ is } \varepsilon \text{ representative}\}) \geq 1-\delta$. Since that claim is true $\forall \varepsilon \in (0,1)$ let us choose $\varepsilon \mapsto \varepsilon/2$, therefore $D^m(\{S \in (X \times Y)^m : S \text{ is } \varepsilon/2 \text{ representative}\}) \geq 1-\delta$. Now, since S is $\varepsilon/2$ representative, we know that $L_D(h_S) \leq \min\limits_{h \in H} L_D(h) + \varepsilon$, where $h_S = \underset{h \in H}{argmin}\ L_S(h)$.
   Thus, we have $D^m\left(\left\{S: L_D(h_S) \leq \min\limits_{h' \in H} L_d(h) + \varepsilon\right\}\right) \geq 1-\delta$, which is the definition of Agnostic-PAC learnability, w.r.t sample complexity $m_H(\varepsilon,\delta) \leq m_H^{UC}(\varepsilon/2,\delta)$

8. H is not agnostic PAC learnable:
   The claim suggests that A can depend on D, even though it should not according to the definition of PAC learnability. This dependance (of A on D) suggests that any hypothesis class satisfies the statement. As a counter example, consider the hypothesis class H of all functions over some non-finite domain X. We have seen that such H is not PAC learnable. Furthermore, we have seen in class that H is PAC learnable iff H is agnostic PAC learnable,

therefore, since the class above is not PAC learnable, it is not agnostic PAC learnable even though it satisfies the question's claim.

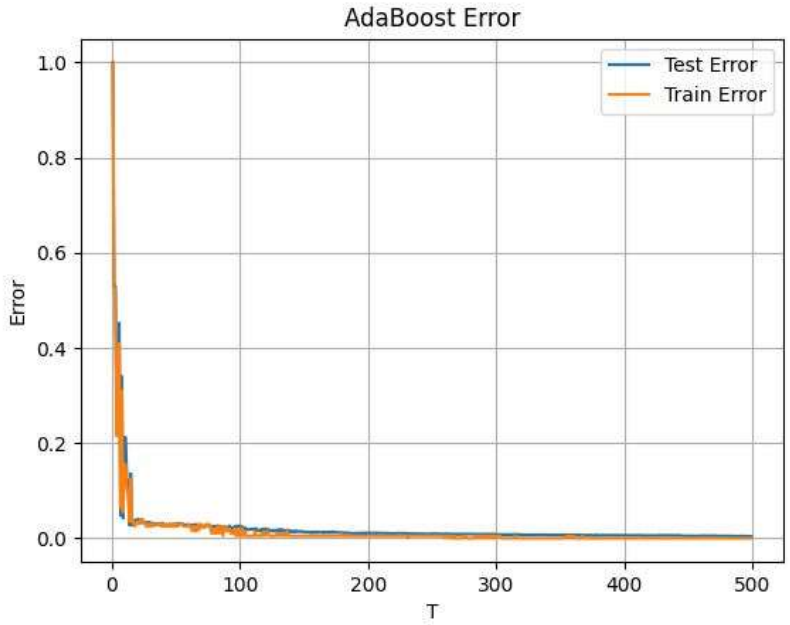
Monotonicity

9. $\varepsilon_1 \leq \varepsilon_2 \rightarrow m_H(\varepsilon_1, \delta) \geq m_H(\varepsilon_2, \delta)$: assume by contradiction that $\varepsilon_1 \leq \varepsilon_2$ but $m_H(\varepsilon_1, \delta) < m_H(\varepsilon_2, \delta)$. By definition, $m_H(\varepsilon_2, \delta)$ is the minimal number of samples needed to achieve $D^m(\{S: L_{D,f}(h_S) \leq \varepsilon_2\}) \geq 1 - \delta$, but from our assumption we have $D^m(\{S: L_{D,f}(h_S) \leq \varepsilon_1 \leq \varepsilon_2\}) \geq 1 - \delta$ using $m_H(\varepsilon_1, \delta) < m_H(\varepsilon_2, \delta)$ samples. In other words, we have used less samples to achieve better resolution, which is a contradiction.
   The claim for δ is very similar:
   $\delta_1 \leq \delta_2 \rightarrow m_H(\varepsilon, \delta_1) \geq m_H(\varepsilon, \delta_2)$: assume by contradiction that $\delta_1 \leq \delta_2$ but $m_H(\varepsilon, \delta_1) < m_H(\varepsilon, \delta_2)$. By definition, $m_H(\varepsilon, \delta_2)$ is the minimal number of samples needed to achieve $D^m(\{S: L_{D,f}(h_S) \leq \varepsilon\}) \geq 1 - \delta_2$, but from out assumption we have $D^m(\{S: L_{D,f}(h_S) \leq \varepsilon\}) \geq 1 - \delta_1 \geq 1 - \delta_2$, using $m_H(\varepsilon, \delta_1) < m_H(\varepsilon, \delta_2)$. In other words, we have used less samples to achieve better accuracy, which is a contradiction.
10. Denote $VCdim(H_i) = d_i$ for $i \in \{1,2\}$. Let us assume by contradiction that $d_1 > d_2$. If so, there exists some $C = \{c_1, \ldots c_{d_1}\}$ that is shattered by H_1 and not by H_2, which in turns mean that there exists some $h_1 \in H_2$ that, given C, generates every label. Since $H_1 \subseteq H_2$, then $h_1 \in H_2$, therefore H_2 also shatters C, in contradiction to the claim that it's VC dimension is smaller than H_1's . From here we conclude $H_1 \subseteq H_2 \Rightarrow VCdim(H_1) \leq VCdim(H_2)$
11. Not for submission

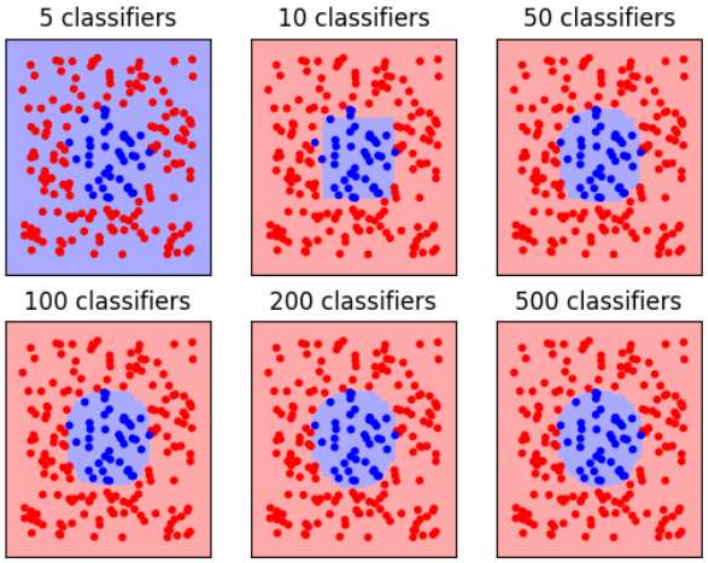
Practical Questions

Separate the Inseparable – AdaBoost

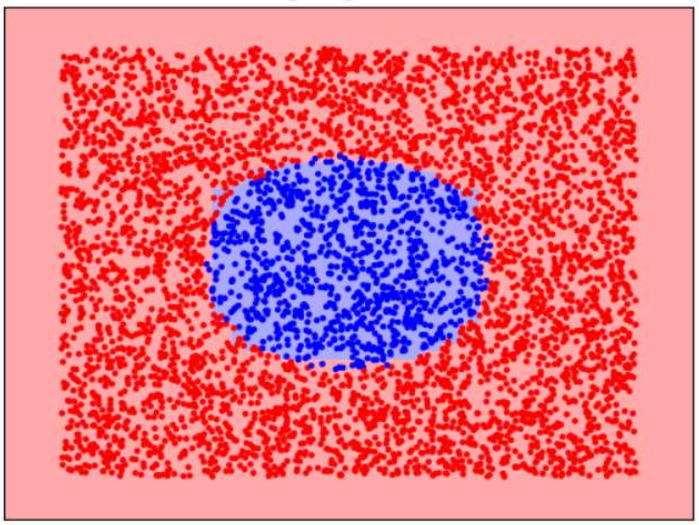
12. code
13. Ada-Boost Error graph:



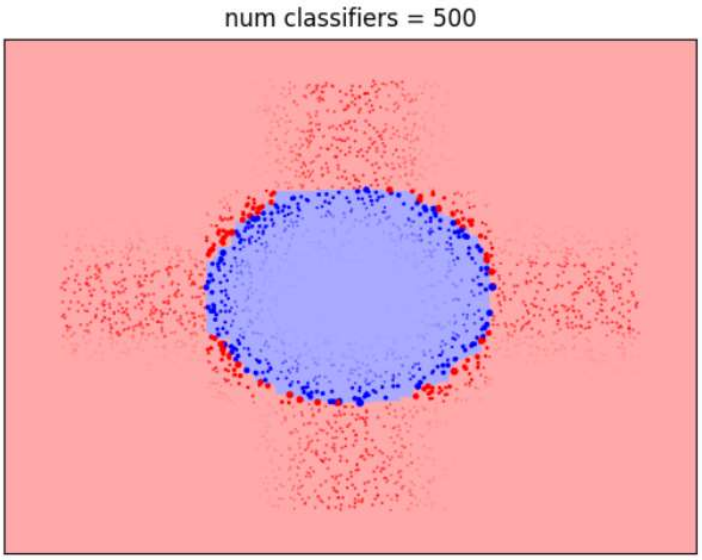
14. Decision boundaries of the learned classifiers:



15. $\hat{T} = 61$, and the given error is 0.005

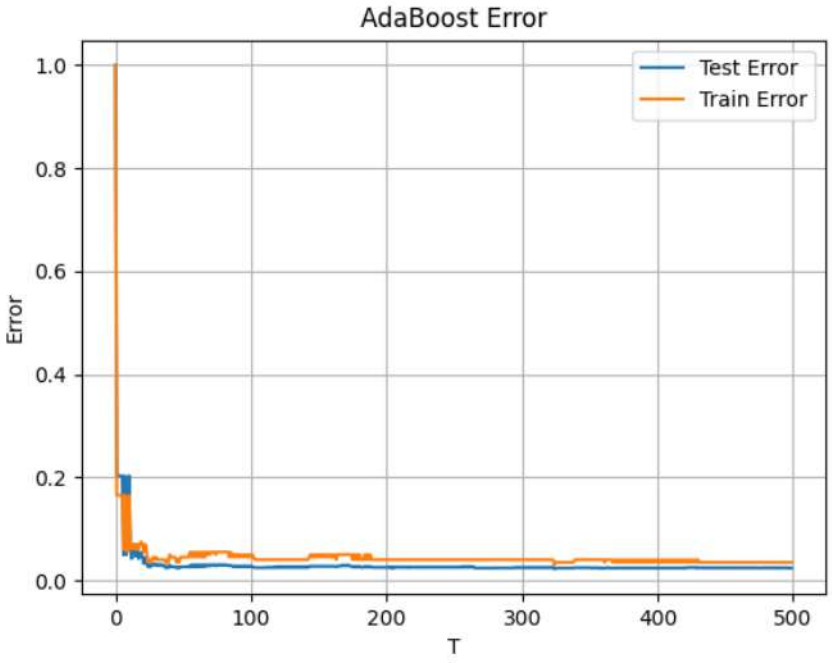


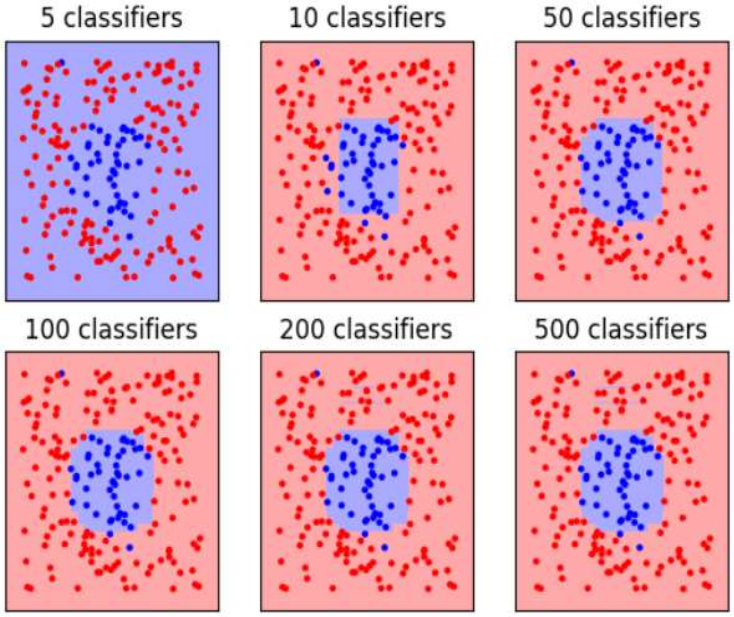
16. Training set with size proportional to D^T:

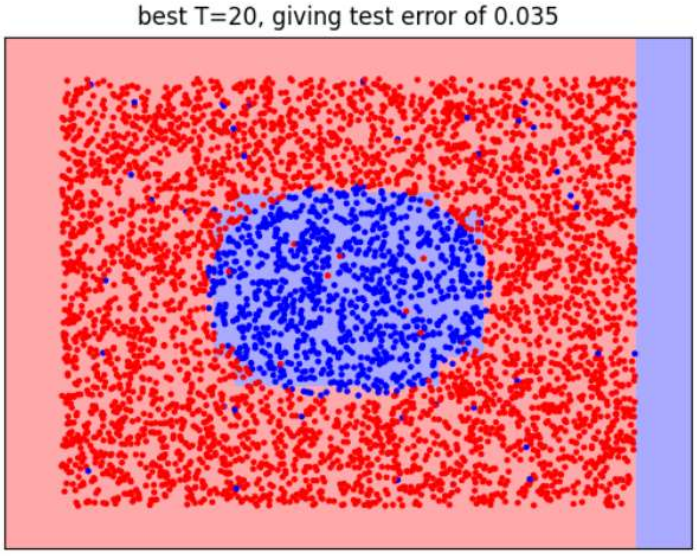


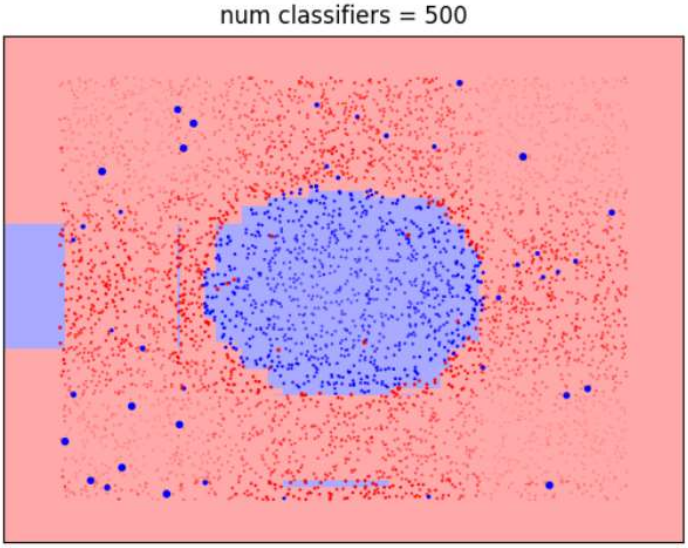
It is harder to classify points near the edges, therefore the classifier increments their respective weights many times (so they are represented with bigger points). Other points are easier to classify, which corresponds to smaller points with smaller weights.

17. Repeating the process with noise 0.01:

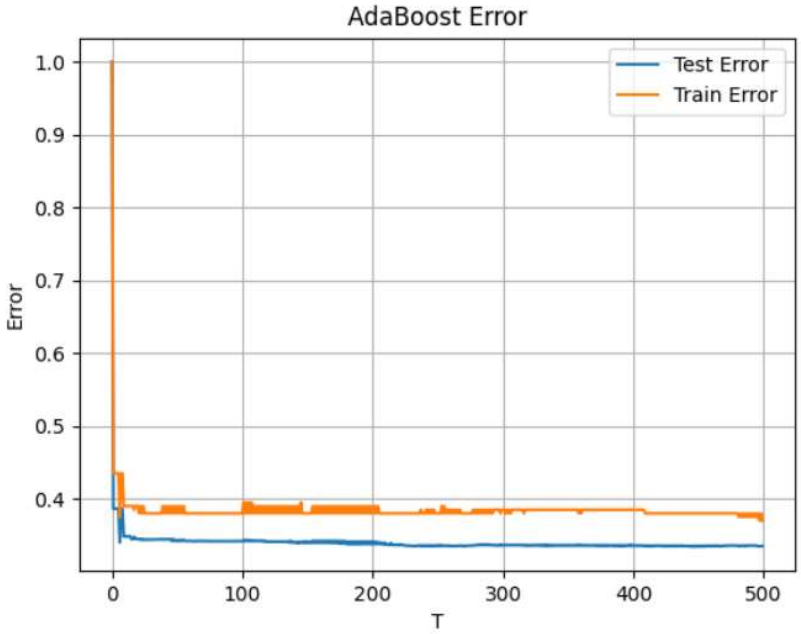

AdaBoost Error
Test Error
Train Error
Error
T
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500

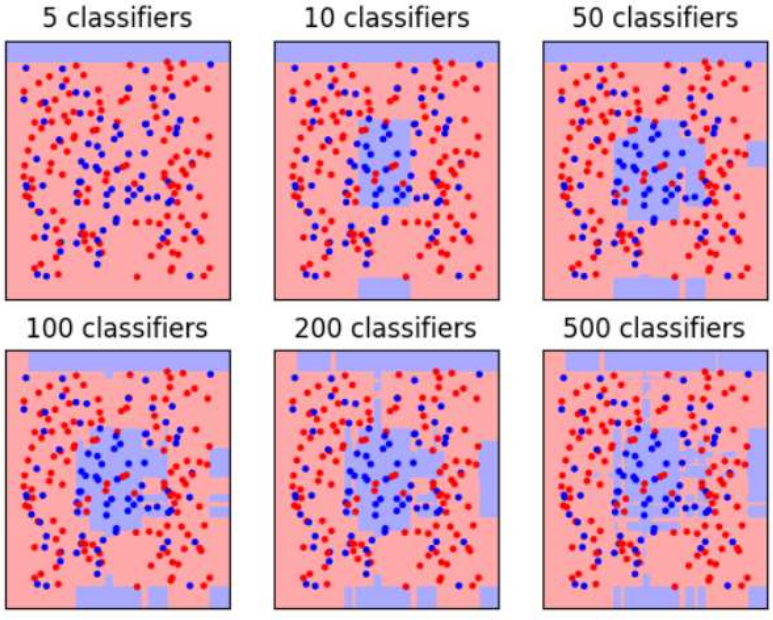

5 classifiers
10 classifiers
50 classifiers
100 classifiers
200 classifiers
500 classifiers

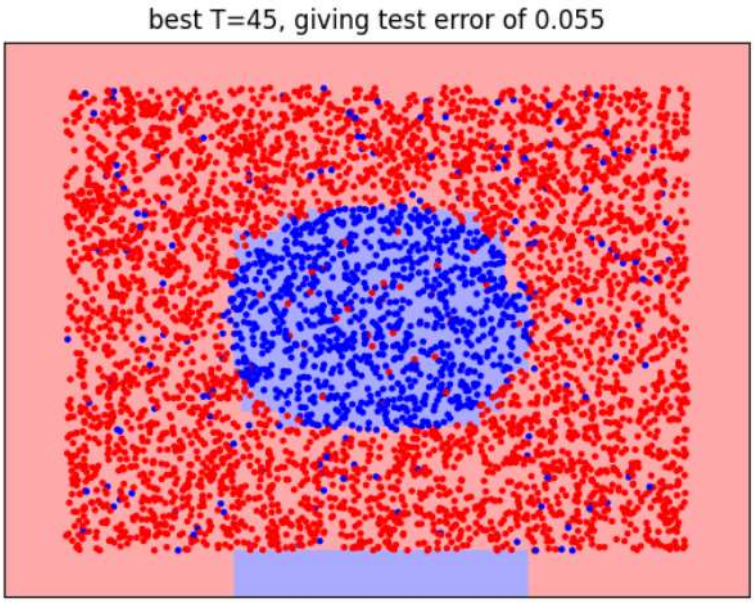

best T=20, giving test error of 0.035

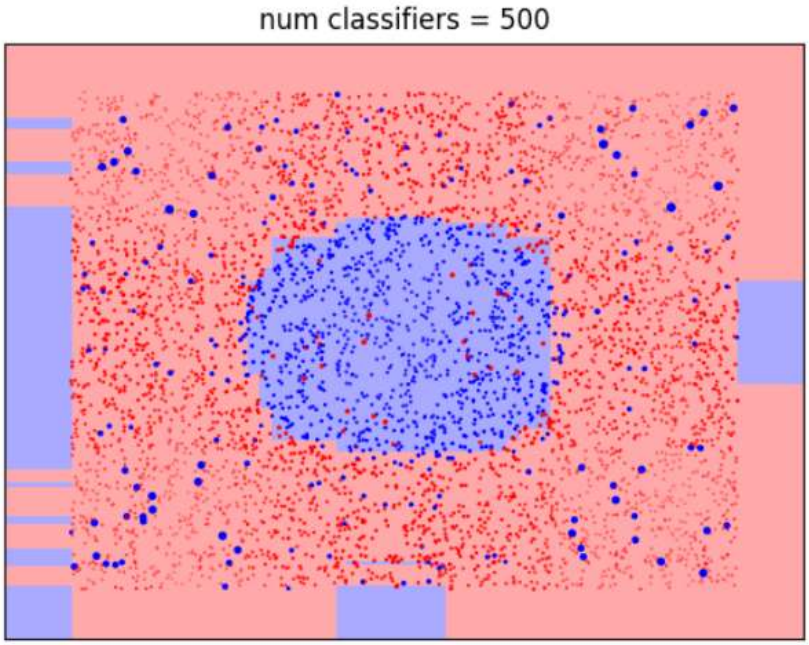


Repeating the process with noise 0.4:









Short description of the changes: We can see that for noisier data, the classification is not as good – the error and the mismatched samples higher, and the algorithm precision decreases.

Explain 13 in terms of bias-variance tradeoff: many decision stumps will cause overfit. This is because the addition of error creates a noised data which in turn means that the estimation error increases. Consequently, the generalization error increases

Explain the differences in 15: noised data causes over fitting for large T (many decision stumps). From here we understand that the best result is when the estimation error is still low, that is, when T is lower.