# 1. Theoretical Part

## 1.1 Regularization

1)

a) $\hat{w}_\lambda = (X^TX + \lambda I_d)^{-1}(X^Ty)$ as shown in class

$\hat{w} = (X^TX)^{-1}X^Ty \iff X^Ty = X^TX\hat{w}$

$A_\lambda = (X^TX + \lambda I_d)^{-1}X^TX$

$\hat{w}_\lambda = (X^TX + \lambda I_d)^{-1}(X^TX)\hat{w} = A_\lambda \hat{w}$ ∎

d) $E[\hat{w}_\lambda] \overset{a}{=} E[(X^TX + \lambda I_d)^{-1}(X^TX)\hat{w}] = A_\lambda E[\hat{w}] = A_\lambda \cdot w \neq w$ (constant)

c) $\hat{w}_\lambda = A_\lambda \hat{w} \iff Var(\hat{w}_\lambda) = Var(A_\lambda \hat{w})$

From the hint: where B is $A_\lambda$ and z is $\hat{w}$

$Var(A_\lambda \hat{w}) = A_\lambda \cdot Var(\hat{w}) \cdot A_\lambda^T \overset{(c}{=} A_\lambda \cdot \sigma^2 (X^TX)^{-1^T} A_\lambda$ ∎

given that $Var(\hat{w}) = \sigma^2(X^TX)^{-1}$

d) $Bias(\hat{w}_\lambda) = E[\hat{w}_\lambda] - w \overset{(d}{=} A_\lambda w - w = (A_\lambda - I)w$

$Var(\hat{w}_\lambda) \overset{(c}{=} A_\lambda \cdot \sigma^2 (X^TX)^{-1^T} A_\lambda$

$Squared\ Bias = \|(A_\lambda - I)w\|^2 = w^T(A_\lambda - I)^T(A_\lambda - I)w$

$E[\|\hat{w}_\lambda - w\|^2] = E[(\hat{w}_\lambda - w)^T(\hat{w}_\lambda - w)] = E\left[\sum_{i=1}^{d}(\hat{w}_{\lambda i} - w_i)^2\right]$

$= \sum_{i=1}^{d} E[(\hat{w}_{\lambda i} - w_i)^2] = \sum_{i=1}^{d}\left(E[(w_{\lambda i} - E(w_i))^2] + (E(w_i) - w_i)^2\right)$

$= E[\|w_\lambda - w\|^2] + \sum_{i=1}^{d}(E(w_i) - w_i)^2$

$= E[\|w_\lambda - w\|^2] + \|E(w) - w\|^2$

$\Downarrow$ $\qquad$ $\Downarrow$

$Var(\hat{w}_\lambda) = \sigma^2 A_\lambda (X^TX)^{-1} A_\lambda^T$ $\qquad$ $Bias(\hat{w}_\lambda) = E(\hat{w}_{\lambda i}) - \lambda = (A_\lambda - I_\lambda)w$

$\Rightarrow Var(\lambda_i) = Var(\hat{w}_\lambda)_{ii}$ $\qquad$ $\Rightarrow Bias^2(\lambda_i) = (A_\lambda - I)w)_i$

e) Because when $\lambda = 0$ then it's the normal ls so we search the best solution only relying on the data adding $\lambda$ increas variance and Bias but cause its negative we add more Bias but remove alot mor variance

2)

a) $z = w_L^{Tik}$

$$LS(w) = \|Xz - y\|^2 + \|Lz\|^2 = (Xz-y)^T(Xz-y) + (Lz)^T(Lz)$$

$$= (Xz)^T \cdot Xz - 2y^T Xz + y^T y + z^T L^T L z$$

$$= z^T (X^T X + L^T L) z + 2y^T Xz + y^T y$$

$$\nabla_z L(w) = 2(X^T X + L^T L) z - 2y^T X = 0$$

$$w_L^{Tik} = (X^T X + L^T L)^{-1} X^T y \quad \blacksquare$$

b) No it's not necessary because $X^T X$ is always psd so by adding $L^T L$ we can make $X^T X + L^T L$ invertable
Thus the new condition would be: $X^T X + L^T L$ invertable

c) i. $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^T \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix} = X^T X$

$$(X^T X)^{-1} = \frac{1}{10 \cdot 20 - 14^2} \begin{bmatrix} 20 & -14 \\ -14 & 10 \end{bmatrix} \qquad X^T y = \begin{bmatrix} 23 \\ 34 \end{bmatrix}$$

$$LS(w) = (X^T X)^{-1} X^T y = \begin{bmatrix} -4 \\ 9/2 \end{bmatrix} = \hat{w}$$

ii. $X^T X + I = \begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 14 \\ 14 & 21 \end{bmatrix}$

$$LS(w) = \frac{1}{11 \cdot 21 - 14^2} \begin{bmatrix} 21 & -14 \\ -14 & 10 \end{bmatrix} \cdot \begin{bmatrix} 23 \\ 34 \end{bmatrix} = \begin{bmatrix} 1/5 \\ \frac{52}{35} \end{bmatrix} = \hat{w}_{Ridge}$$

iii. $L^T L = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$

$$X^T X + L^T L = \begin{bmatrix} 10 & 14 \\ 14 & 24 \end{bmatrix}$$

$$LS(w) = \frac{1}{10 \cdot 24 - 14^2} \begin{bmatrix} 24 & -14 \\ -14 & 10 \end{bmatrix} \cdot \begin{bmatrix} 23 \\ 34 \end{bmatrix} = \begin{bmatrix} 19/11 \\ 9/22 \end{bmatrix} = \hat{w}_L^{Tik}$$

i. $\|\hat{w}\| = \sqrt{(-4 \;\; 9/2)\begin{pmatrix} -4 \\ 9/2 \end{pmatrix}} = \sqrt{\frac{4 \cdot 16}{4} + \frac{81}{4}} = \frac{\sqrt{145}}{2} \approx 6$

$\|\hat{w}_{Ridge}\| = \left(\frac{1}{5} \;\; \frac{32}{35}\right)\begin{pmatrix} 1/5 \\ 32/35 \end{pmatrix} = \frac{\sqrt{2753}}{35} \approx 1.5$

$\|\hat{w}_L^{Tilc}\| = \left(\frac{19}{11} \;\; \frac{9}{22}\right)\begin{pmatrix} 19/11 \\ 9/22 \end{pmatrix} = \frac{\sqrt{1525}}{22} \approx 1.8$

The largest is $\hat{w}$ because There is no penalize for large waights we just trying to fit the data
in Ridge its small becas we penalize both wights
and $w_L^{Tilc}$ is inthe middle because we punish just for one wight

ii. $\hat{w}_2 = 9/2 = 4.5$
$\hat{w}_{Ridge} = 52/35 \approx 1.5$
$\hat{w}_L^{Tilc} = 9/22 \approx 0.5$

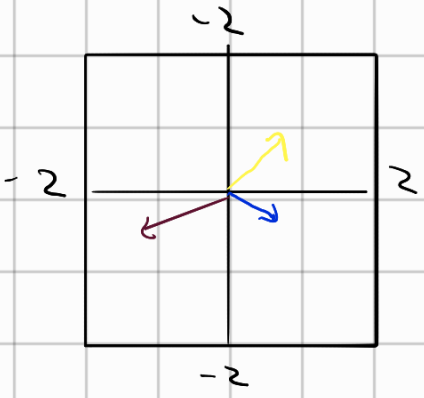
This happens for the same reasonse above and because in Tilc We punish more Than ridge
The second coordinate is smaller

## 1.2 PCA

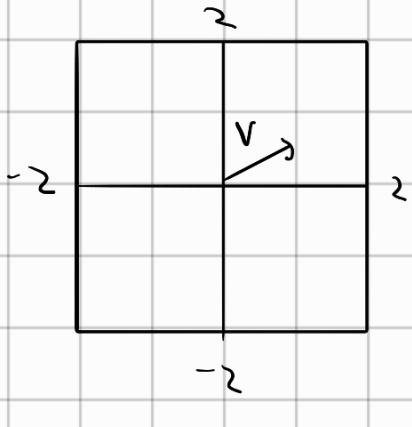
1.

a) $\bar{X} = \frac{1}{3}\sum_{i=1}^{3} X_i = \begin{bmatrix} \frac{1}{3} \\ \frac{4}{3} \end{bmatrix}$

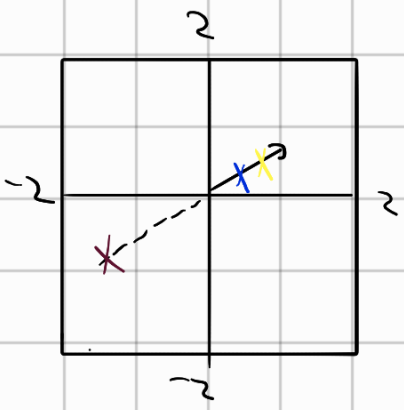
$z_1 = \begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$ $z_2 = \begin{bmatrix} -\frac{4}{3} \\ -\frac{1}{3} \end{bmatrix}$ $z_3 = \begin{bmatrix} \frac{2}{3} \\ -\frac{1}{3} \end{bmatrix}$




b) $\Sigma = \begin{bmatrix} 8/9 & 2/9 \\ 2/9 & 2/9 \end{bmatrix}$ $V \approx \begin{pmatrix} 0.95 \\ 0.29 \end{pmatrix}$



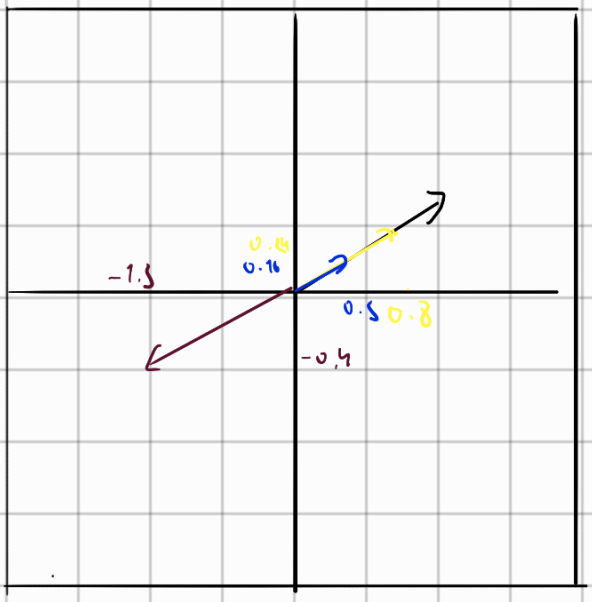

c) $V^T z_1 = 0.82$ X
$V^T z_2 = -1.37$ X
$V^T z_3 = 0.55$



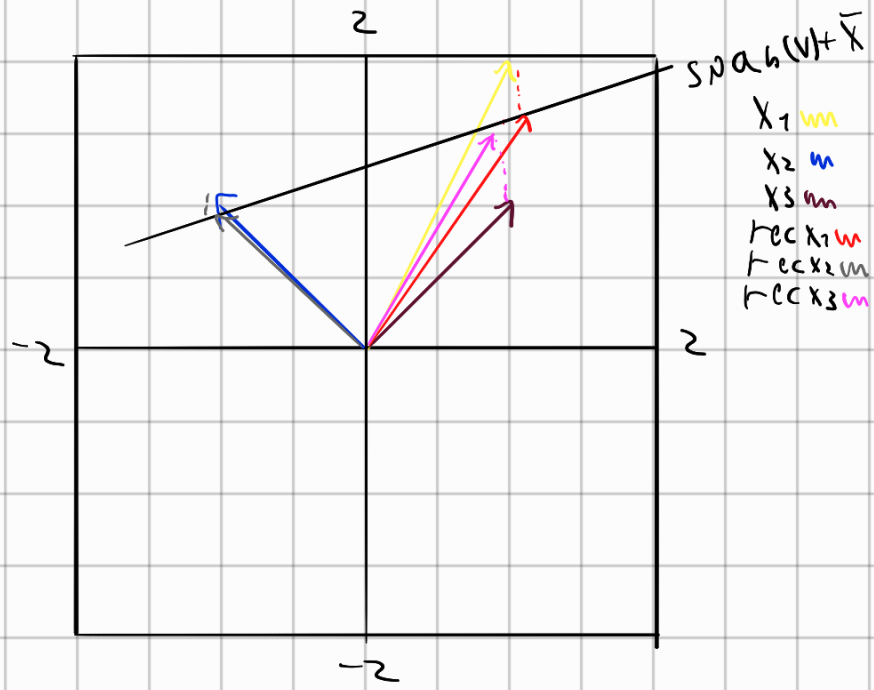

d) $proj_V(z_1) = \begin{bmatrix} 0.79 \\ 0.24 \end{bmatrix}$ $\hat{z_1}$

$proj_V(z_2) = \begin{bmatrix} -1.3 \\ -0.4 \end{bmatrix}$ $\hat{z_2}$

$proj_V(z_3) = \begin{bmatrix} 0.5 \\ 0.16 \end{bmatrix}$ $\hat{z_3}$




e) $rec X_1 = \begin{bmatrix} 1.1 \\ 1.6 \end{bmatrix}$ $rec X_2 = \begin{bmatrix} -0.98 \\ 0.936 \end{bmatrix}$ $rec X_3 = \begin{bmatrix} 0.852 \\ 1.49 \end{bmatrix}$


2
span(v)+ $\bar{x}$
$x_1$
$x_2$
$x_3$
rec $x_1$
rec $x_2$
rec $x_3$
-2
2
-2

2)

a) $\sum_{i=1}^{m} \| x_i - \bar{x} - w^T w (x_i - \bar{x}) \|^2 = \sum_{i=1}^{m} \| z_i - w^T w z_i \|^2$

$$= \sum_{i=1}^{m} z_i^T z_i - 2 z_i^T (w^T w z_i) + (w^T w z_i)^T (w^T w z_i) \quad *$$

$$= \sum_{i=1}^{m} \|z_i\|^2 - 2 z_i^T w^T w z_i + z_i^T w^T w z_i$$

$$= \sum_{i=1}^{m} \|z_i\|^2 - z_i^T w^T w z_i$$

$$* = z_i^T w^T \underbrace{w w^T}_{I_k} w z_i = z_i^T w^T w z_i$$

□ because $w$ rows is orthonormal vectors

b) $\sum_{i=1}^{m} \underbrace{z_i^T w^T w z_i}_{\text{scalar}} = = \sum_{i=1}^{m} Tr(w z_i z_i^T w^T) = Tr\left(w \left(\sum_{i=1}^{m} z_i z_i^T\right) w\right)$

$$= Tr(w(m\Sigma)w^T) = m \cdot Tr(w \Sigma w^T)$$

c) $Tr(w \Sigma w^T) = Tr\left(\left(\sum_{i=1}^{k} v_i^T\right) \Sigma \left(\sum_{i=1}^{k} v_i\right)\right) = \sum_{i=1}^{k} Tr(v_i \Sigma v_i^T)$

d) From a we know

$$\sum_{i=1}^{m} \| x_i - \bar{x} - w^T w (x_i - \bar{x}) \|^2 = \sum_{i=1}^{m} \|z_i\|^2 - z_i^T w^T w z_i$$

$$\min\left(\sum_{i=1}^{m} \|z_i\| - z_i^T w^T w z_i\right) \overset{a,b}{=} \min\left(\sum_{i=1}^{m} \|z_i\| - m \cdot Tr(w \Sigma w)\right)$$

$$\max \sum_{j=1}^{k} v_j^T \Sigma v_j \overset{c}{=} \max(Tr(w \Sigma w))$$

To minimize the LS error we need to maximize the orthogonal projection

3. We need to prove: $\forall v \in \mathbb{R}^d$ with $\|v\|_2 = 1 \Rightarrow Var(v^T X) \leq Var(v_1^T X)$

which is equivalent to $v^T \Sigma v \leq v_1^T \Sigma v_1$

PCA returns $\sum_{i=1}^{k} v_j^T \Sigma v_j$ in descending order that is the variance of $\langle v_i, X \rangle$ is the largest ∎

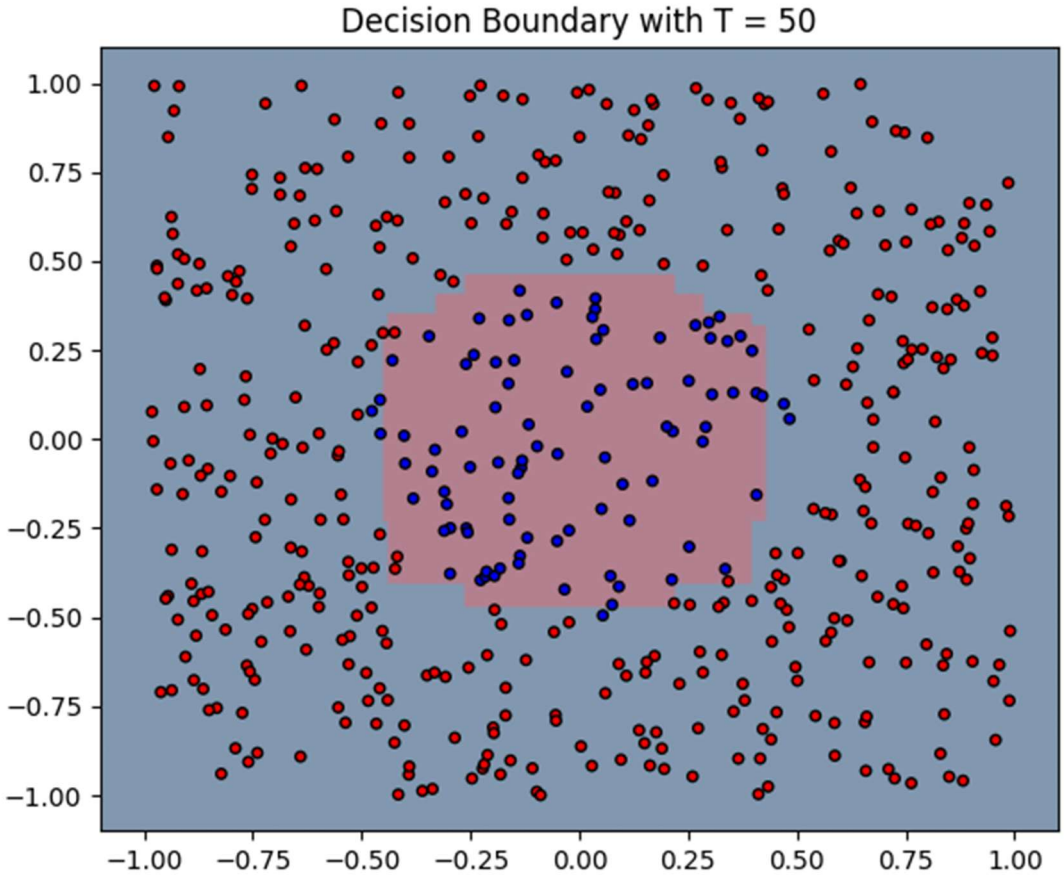
1.3 Clustring

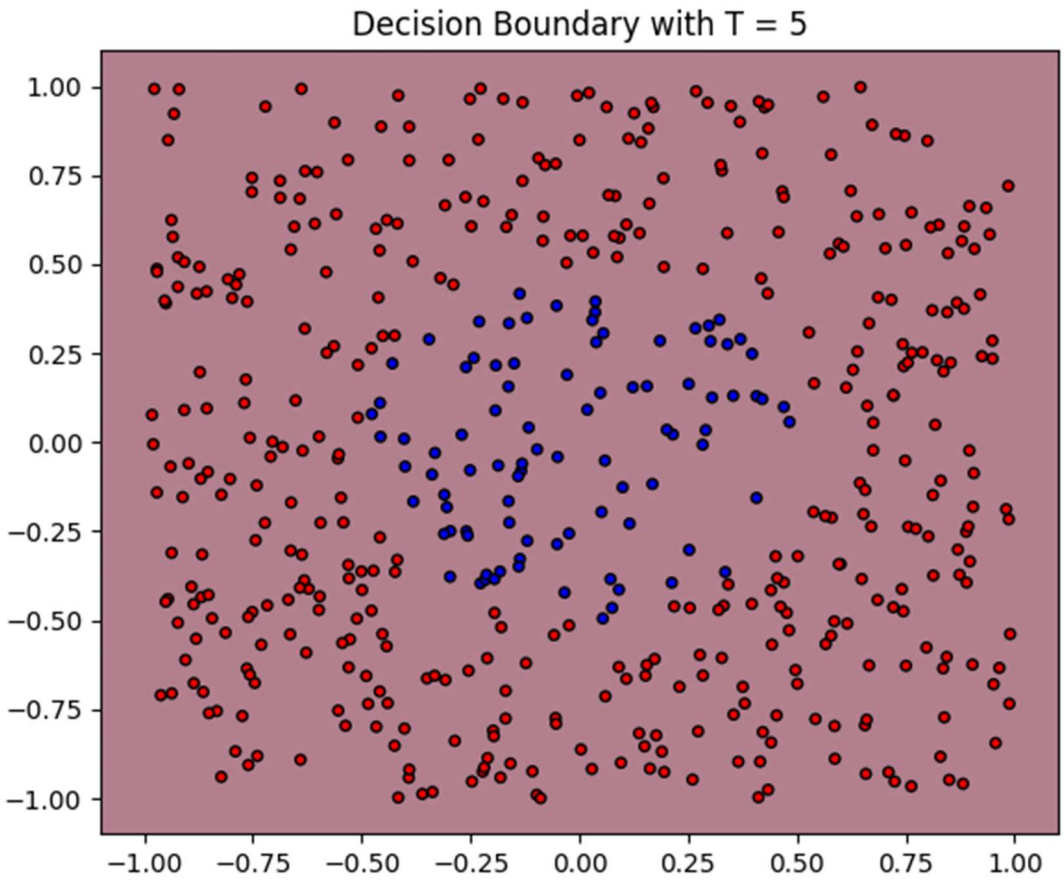
1. a) True if $m_{min}=1$ Then for cluster of size One can be pushed into final_C

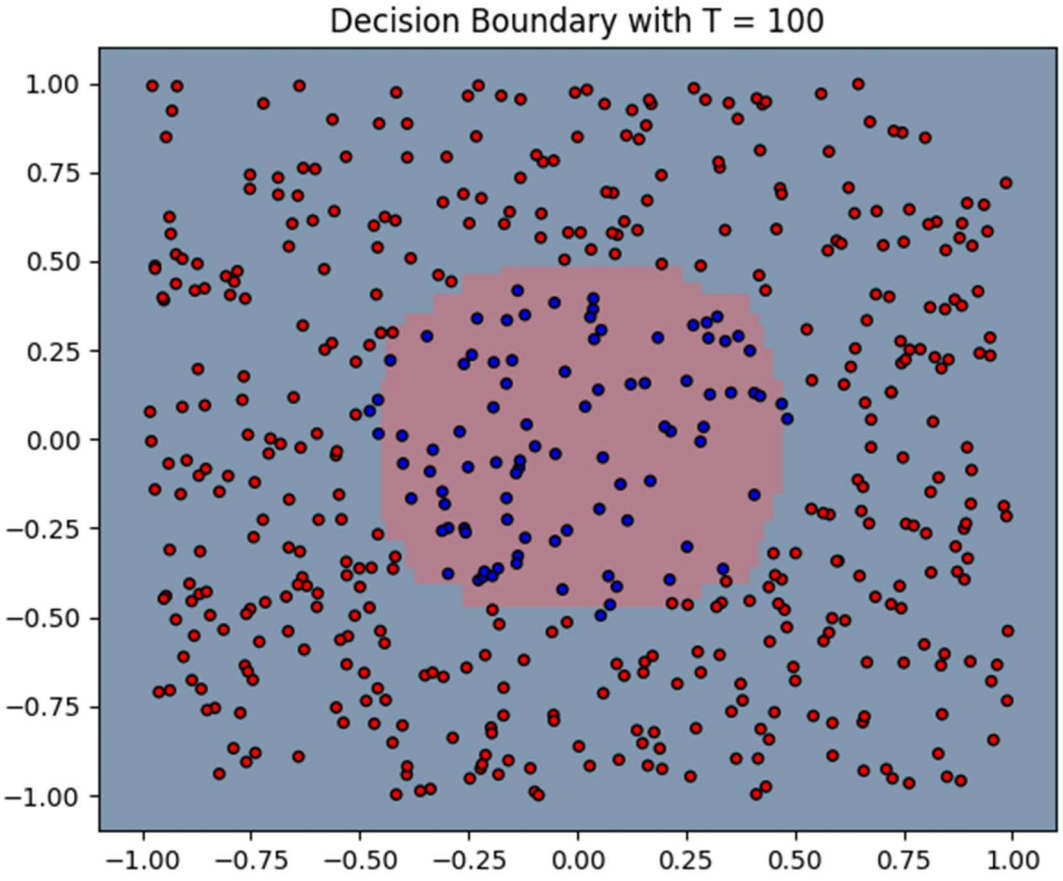
b) false its flying on K-means which is not optimal and can give different output

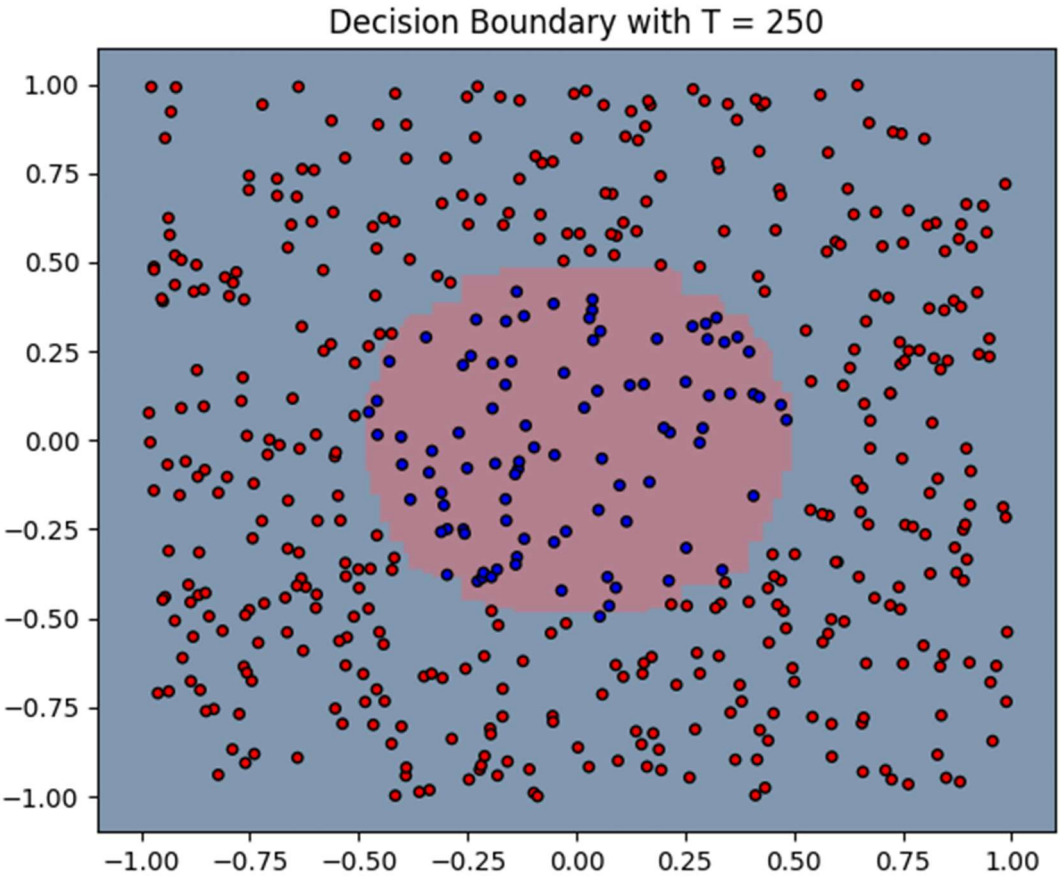
c) True because we making more room for noise

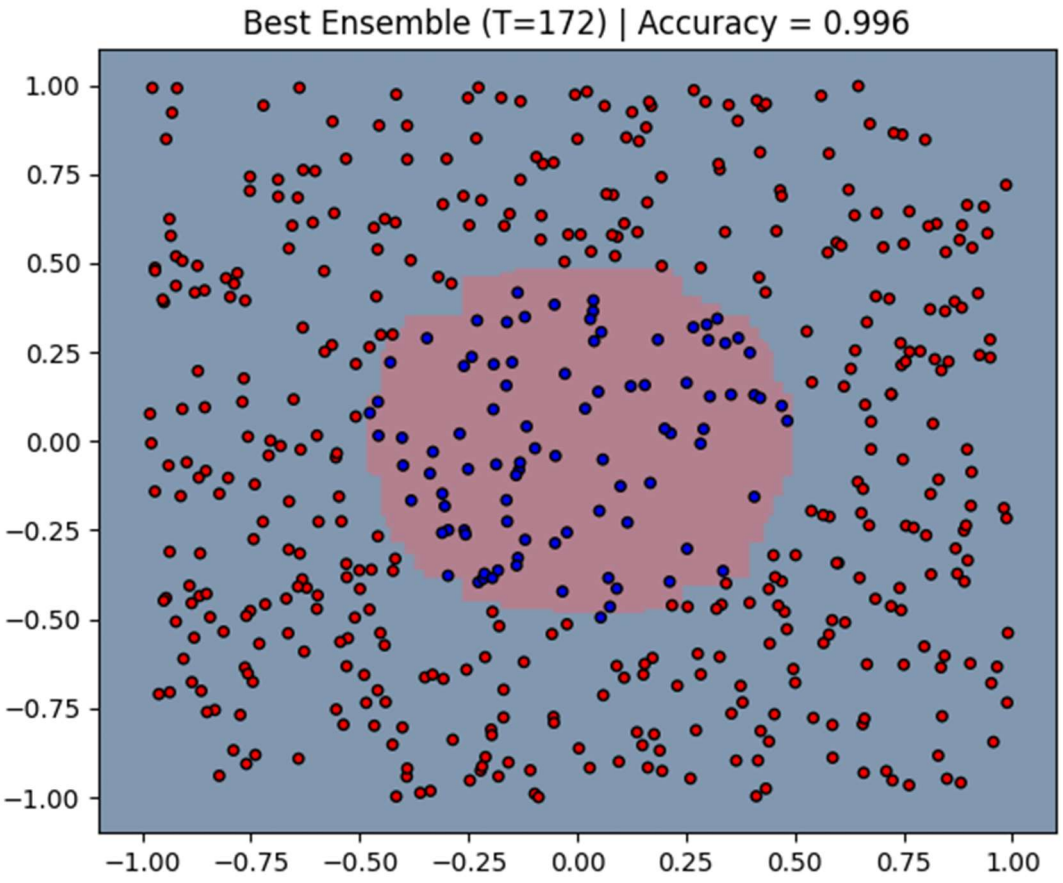
d) True we can get diffrent starting point and hands diffrch output

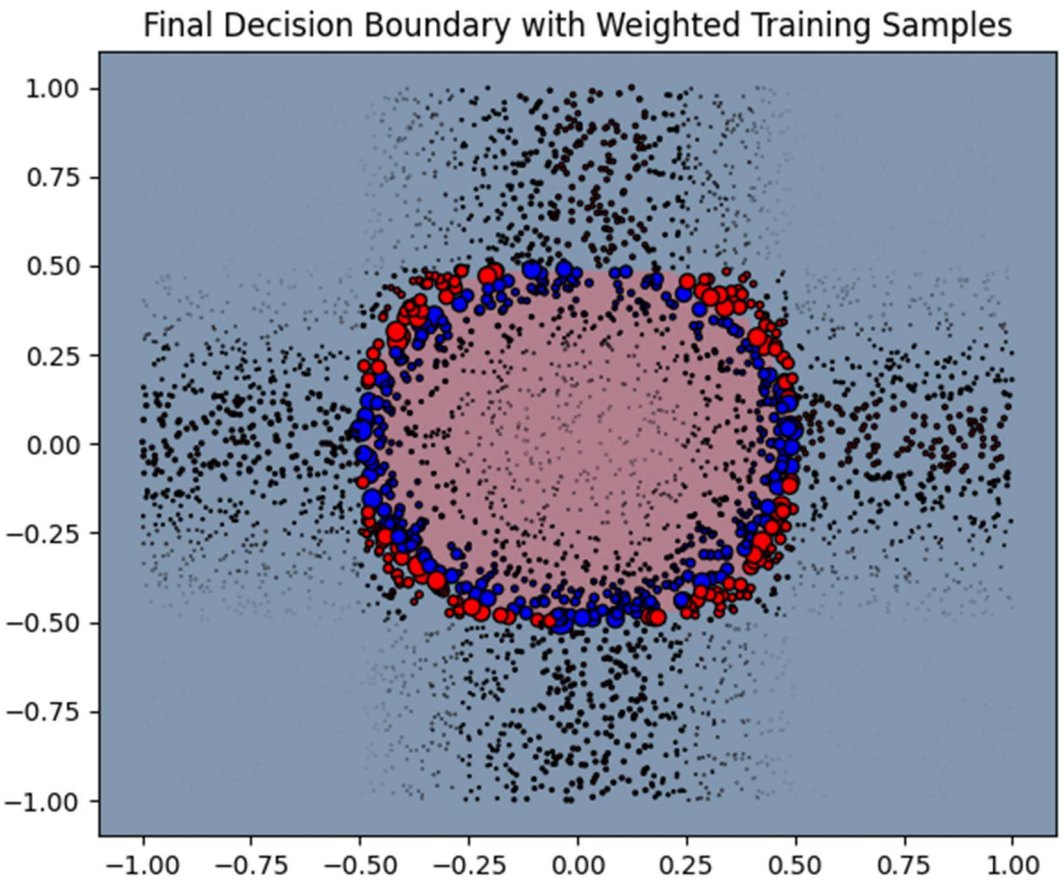

Decision Boundary with T = 50
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

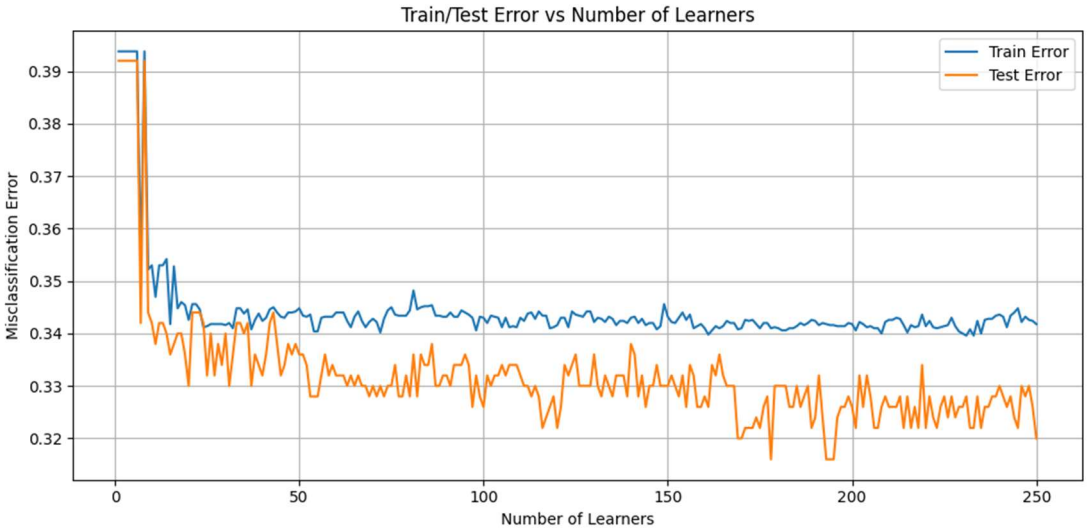

Decision Boundary with T = 5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

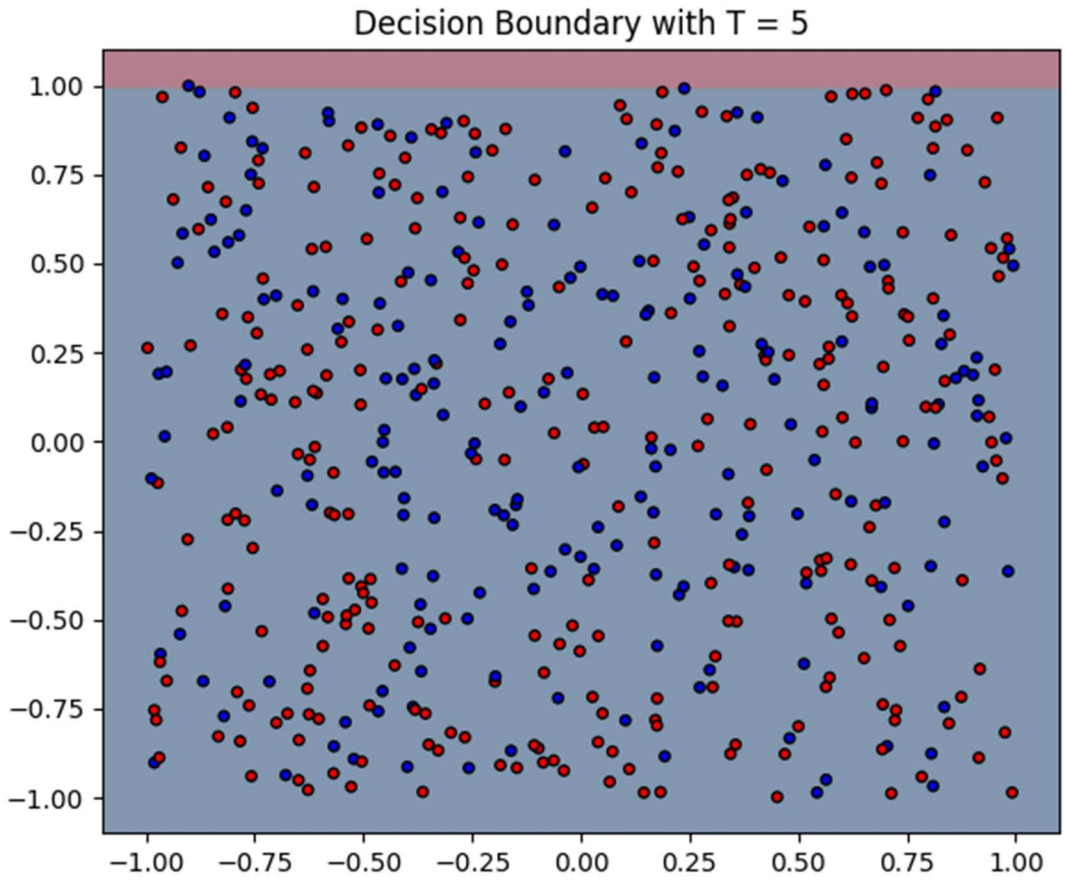

Decision Boundary with T = 100
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

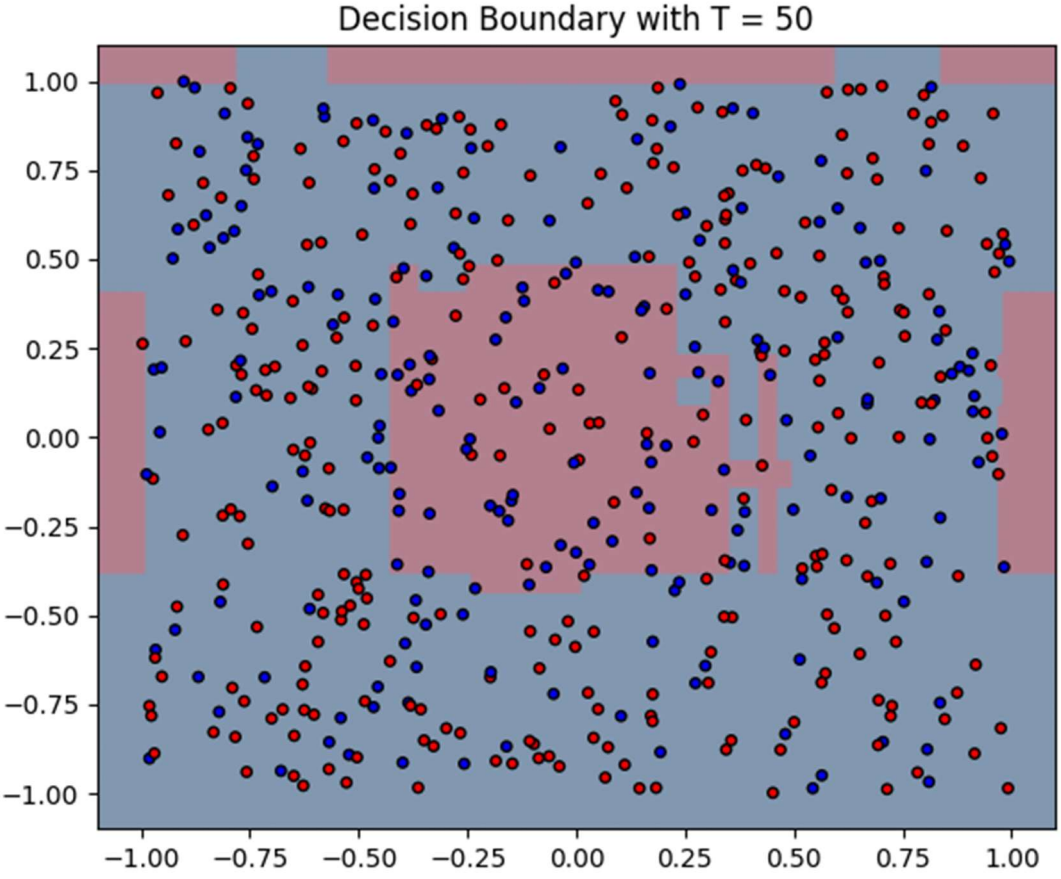

Decision Boundary with T = 250
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

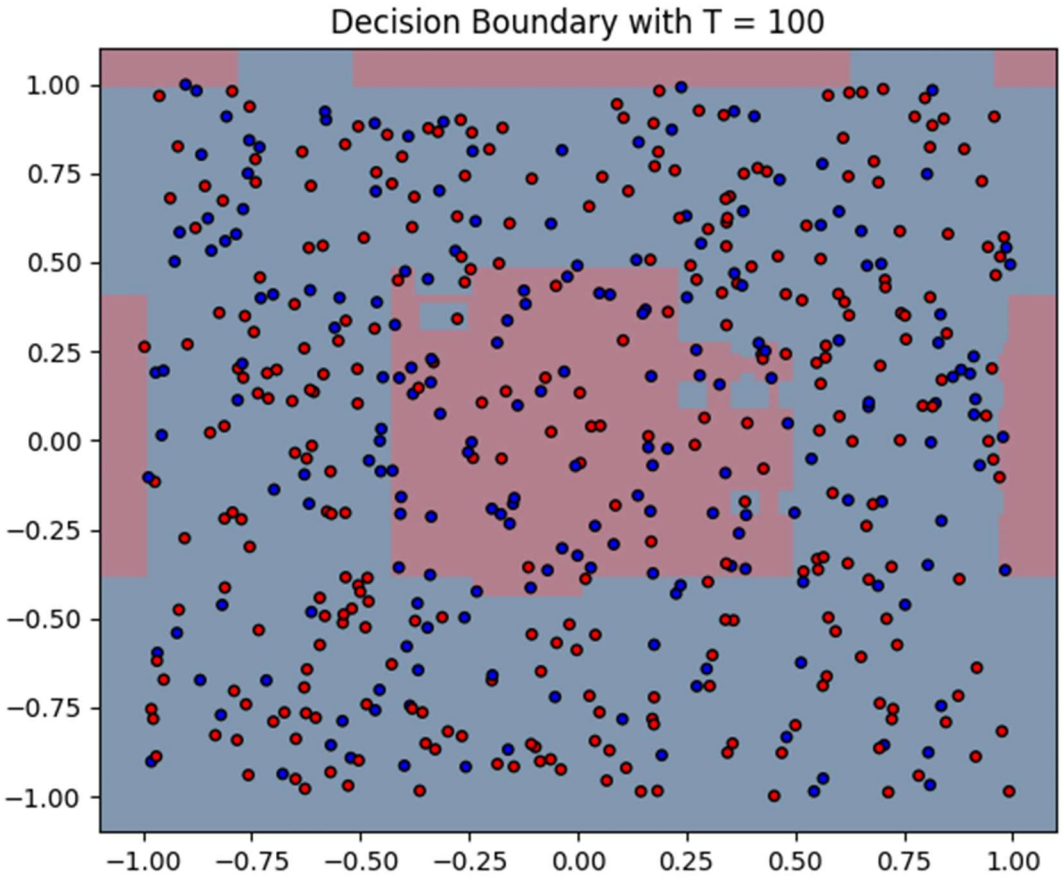

Best Ensemble (T=172) | Accuracy = 0.996
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

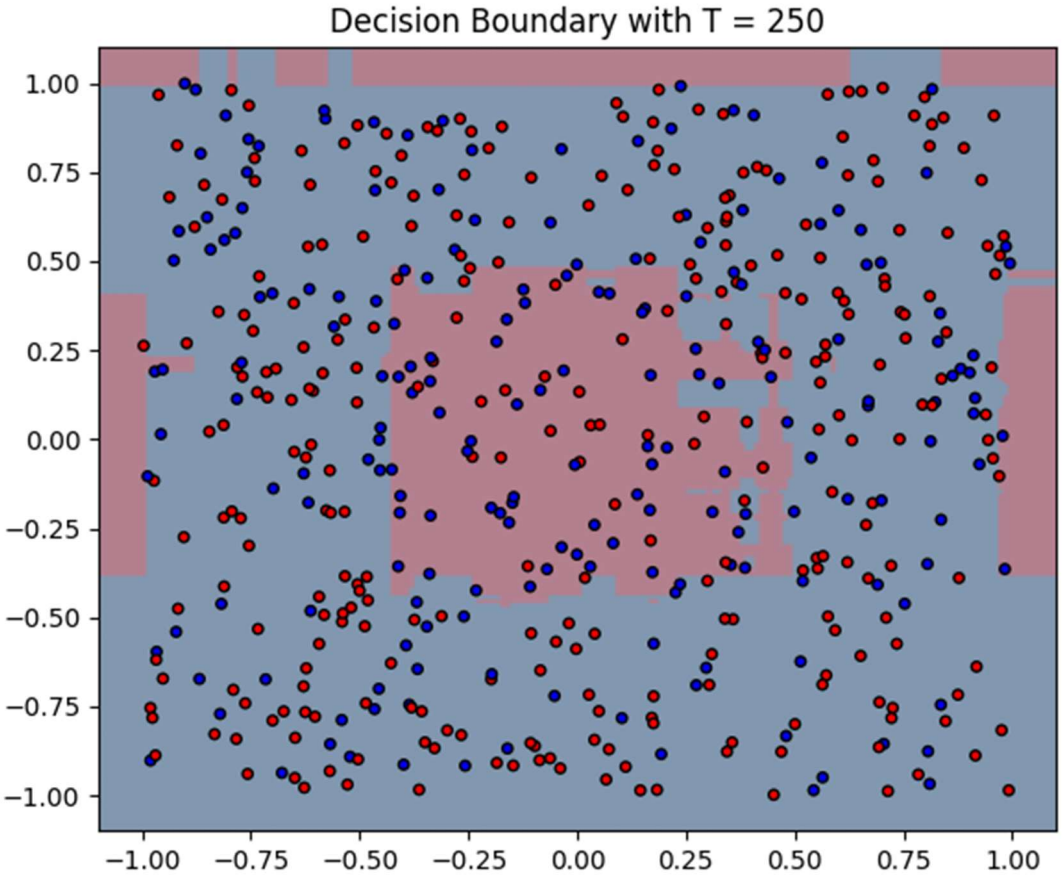

Final Decision Boundary with Weighted Training Samples
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

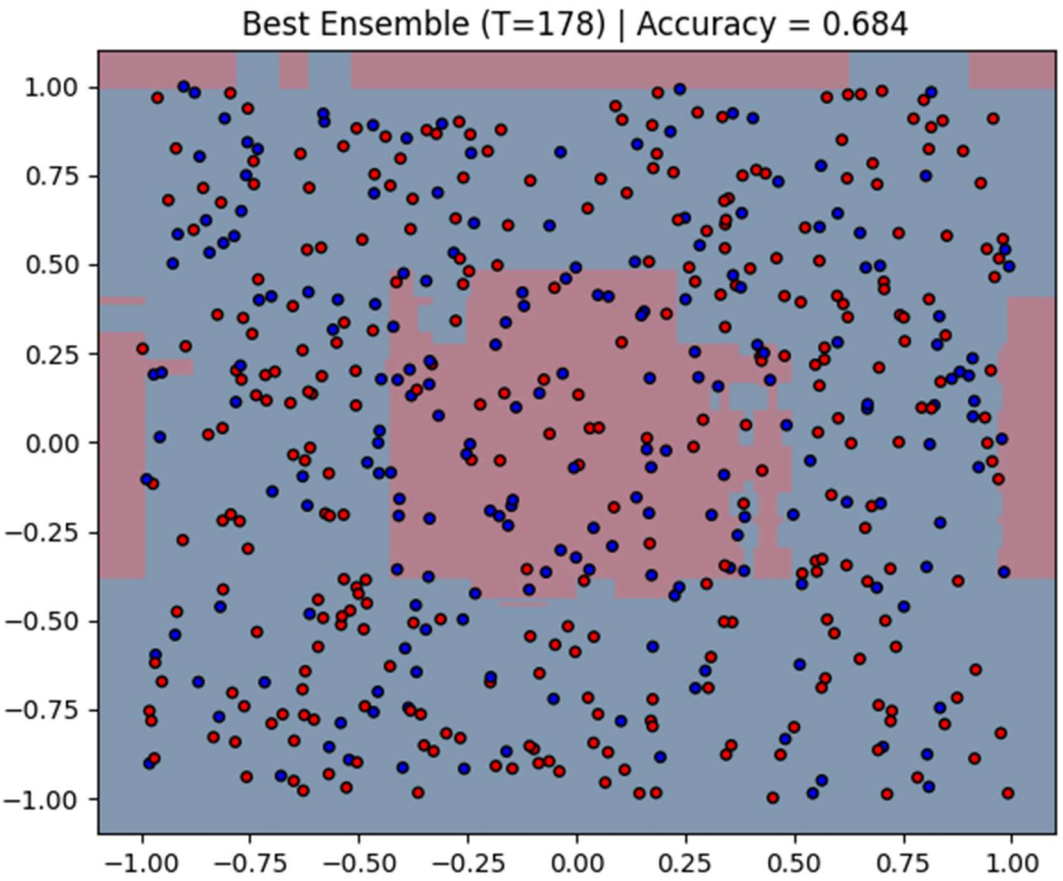

Train/Test Error vs Number of Learners
Train Error
Test Error
Misclassification Error
0.39
0.38
0.37
0.36
0.35
0.34
0.33
0.32
0
50
100
150
200
250
Number of Learners

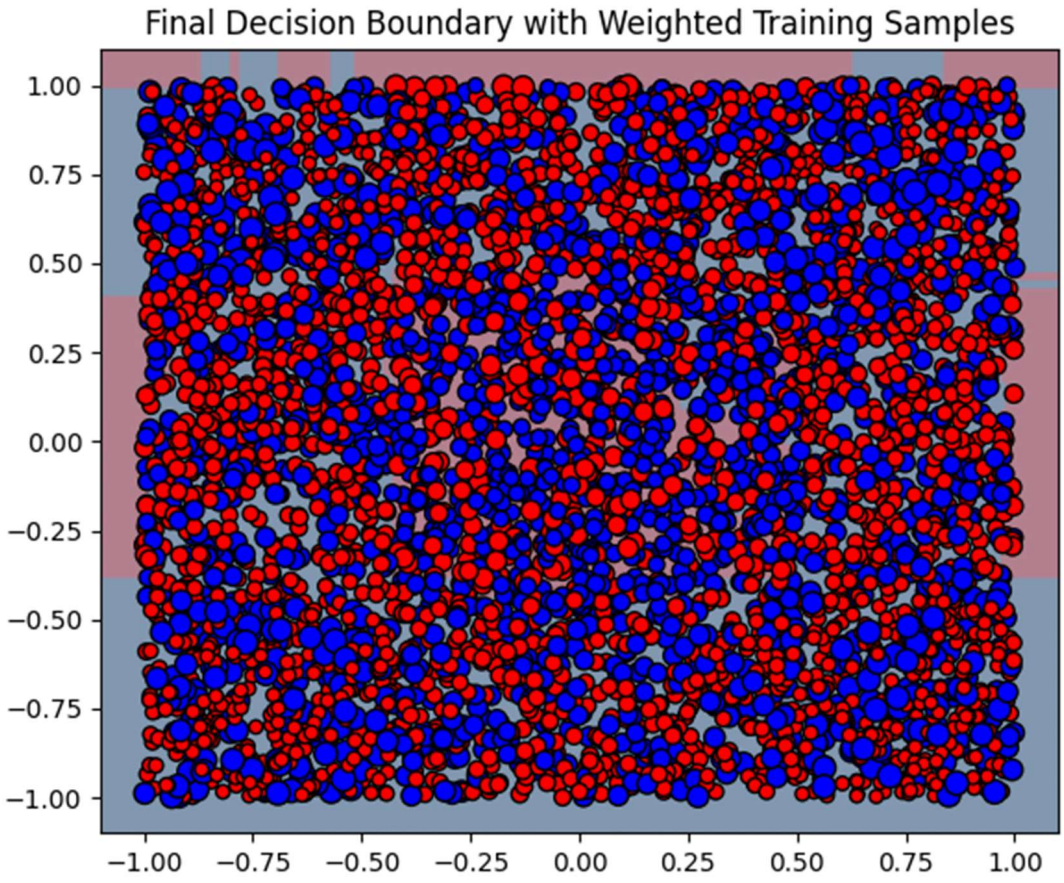

Decision Boundary with T = 5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Decision Boundary with T = 50
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Decision Boundary with T = 100
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Decision Boundary with T = 250
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Best Ensemble (T=178) | Accuracy = 0.684
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Final Decision Boundary with Weighted Training Samples
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

2 Practical Part

2.1 Boosting

1. מכיוון שאין כלל רעש המודל לומד את הנתונים לגמרי עד שאין לו שום טעות ולכן השכיאה מגיעה לכמעט 0 מכיוון שהנתונים נלמדים לגמרי

4.נראה שהמודל מתקשה להחליט עבור נקדודות שנמצאות על התפר בין הנתונים מכיוון שמם הגבול לא ברור ולכן זה קשה עבורו

לא ממש הצלחתי להמשיך מפה לצערי לא היה לי את הזמן ):