PAC

(1) נוכיח שמתקיים (א) ⇐ (ב).

אם נכון $P_{S\sim D^m}[L_D(A(S))\le\epsilon]\ge 1-\delta$

אזי $P_{S\sim D^m}[L_D(A(S))>\epsilon]\le\delta$ קבוע $\lim_{m\to\infty} E_{S\sim D^m}[L_D(A(S))]$

שקול לכך ש:
$\forall\epsilon>0\ \exists m_0\in\mathbb{N}$
s.t. $\forall m\ge m_0,\ E_{S\sim D^m}[L_D(A(S))]<\epsilon$

יהא $\epsilon>0$ ונקח m_0 כך ש $m_0=m(\epsilon/2,\frac{\epsilon}{2})$.

מכיוון שאנו יודעים ש $L_D(A(S))\in[0,1]$ אזי לכל $m\ge m_0$

מתקיים

$$E_{S\sim D^m}[L_D(A(S))]=\int_{\hat\epsilon=0}^{1}\hat\epsilon\cdot P_{S\sim D^m}[L_D(A(S))=\hat\epsilon]\,d\hat\epsilon$$

$$\le\int_{\hat\epsilon=0}^{\frac{\epsilon}{2}}\frac{\epsilon}{2}P_{S\sim D^m}[L_D(A(S))=\hat\epsilon]\,d\hat\epsilon+\int_{\hat\epsilon=\frac{\epsilon}{2}}^{1}1\cdot P_{S\sim D^m}[L_D(A(S))=\hat\epsilon]\,d\hat\epsilon$$

$$=\frac{\epsilon}{2}\cdot P_{S\sim D^m}[L_D(A(S))\le\epsilon/2]+P_{S\sim D^m}[L_D(A(S))>\epsilon/2]<\frac{\epsilon}{2}\cdot 1+\frac{\epsilon}{2}=\epsilon.$$

בכיוון השני: יהא $\epsilon,\delta>0$ ואנו יודעים שקיים m_0 כך שמתקיים שלכל $m>m_0$ מתקיים: $E_{S\sim D^m}[L_D(A(S))]<\epsilon\delta$

ממרקוב מתקיים: $P_{S\sim D^m}[L_D(A(S))>\epsilon]\le\frac{E_{S\sim D^m}[L_D(A(S))]}{\epsilon}$
$<\frac{\epsilon\delta}{\epsilon}=\delta$

ולכן מתקיים (ב) ⇐ (א) ולכן מתקיים הנדרש ∎

2. נכון. נגדיר אלגוריתם ERM A שמחזיר בהינתן $S=\{(x_i,y_i)\}_{i=1}^{m}\subseteq(\mathbb{R}^2\times\{\pm1\})^m$ מחזיר את המעגל הקטן ביותר B_S אשר מכיל את כל הדוגמאות החיוביות.

נסמן את הרדיוס של B_S על ידי r_S.

מהנחת הריאליזביליות קיים h^* מעגל עם שמרכזו בראשית שמסווג נכון. נסמן את הרדיוס שלו כ-r^*.

יהי $\varepsilon\in(0,1)$. נגדיר $r^\varepsilon < r^*$ קטן כך ש

$$D_X(\{x: r^\varepsilon \le \|x\| \le r^*\})=\varepsilon$$

נגדיר $E=\{x\in\mathbb{R}^2 \mid r^\varepsilon\le\|x\|\le r^*\}$. ההסתברות ש $L_D(A(S))\ge\varepsilon$ (מעגל שנוצר ע"י A) קטנה מההסתברות שאין נקודה ב-S שנמצאת ב-E.

ההסתברות הזאת היא $(1-\varepsilon)^m\le e^{-m\varepsilon}$

הגודל הרצוי של m מספיק כדי שיתקיים $e^{-\varepsilon m}\le\delta$.

3. נניח ש-H מתכנסת אחידה עם הפונקציה m^{UC}_H.

לכל $\epsilon,\delta \in (0,1)$ אם S היא ε-נציגה של גודל $m \geq m^{UC}_H(\epsilon/2,\delta)$ אז עבור התפלגות D לפחות $1-\delta$, S היא $\epsilon/2$ מייצגת, מה שאומר שלכל $h \in H$ יש לנו

$$L_D(h_S) \leq L_D(h) + \frac{\epsilon}{2}$$

לכן לכל $h \in H$:

$$L_D(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_D(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_D(h) + \epsilon$$

כאשר האי שוויון הראשון והשלישי נובעים מההנחה ש-S היא $\epsilon/2$ מייצגת והאי שוויון השני נובע מהגדרה של h_S כפלט ERM. לכן עבור כל $h \in H$

$$L_D(h) \leq \min_{h' \in H} L_D(h') + \epsilon$$

ולכן הוכחנו ש-H היא אגנוסטית PAC ניתנת ללמידה עם

$$m_H(\epsilon,\delta) \leq m^{UC}_H(\epsilon/2,\delta)$$

VC - dimension

2.2

(1) המימד הוא לפחות n מכיוון שהקבוצה $\{e_1,\dots,e_n\}$ של וקטורי היחידה ב-$\mathbb{R}^n$ מנופצת, יכולים להוכיח:

לדוגמא: תהי y תיוג של הקבוצה ונגדיר $I=\{i \mid y_i=1\}$, ולכן כמובן $h_I(e_i)=y_i$.

מצד השני, נשים לב ש- $|H|=2^n$ ואז הטענה למחלקות היפותזות סופיות $VCDim(H)\le \log_2|H|$ שנראה בהרצאה גורר ש- $VCDim(H)=n$.

(2) מכיוון ש- $H_1 \subseteq H_2$ אז אם $C=\{c_1,\dots,c_m\}\subseteq X$ אם C מנופצת ע"י H_1 אז היא גם מנופצת ע"י H_2.

מכאן נקבל: $VC\text{-}Dim(H_1)\le VC\text{-}Dim(H_2)$.

Regularization

(1) (a)

$$\hat{w}_\lambda = (X^T X + \lambda I_d)^{-1} (X^T X) (X^T X)^{-1} X^T y$$
$$= (X^T X + \lambda I_d)^{-1} X^T y = \hat{w}_\lambda$$

(b)

$$E[\hat{w}_\lambda] = E[A_\lambda \hat{w}] = A_\lambda E[\hat{w}]$$
$$= (X^T X + \lambda I_d)^{-1} (X^T X) w$$

- $E[\hat{w}_\lambda] \neq w$ ולכן $\hat{w}_\lambda$ אינו אומד חסר הטיה

(c)

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

(d)

$$MSE(\hat{y}) = E[\|\hat{y} - y\|^2] = E\|\hat{y} - \bar{y}\|^2 + \|\bar{y} - y\|^2$$
$$= Var[\hat{y}] + bias^2[\hat{y}]$$

$\bar{y} = E[\hat{y}]$, $\hat{y} = \hat{w}_\lambda$, $\bar{y} = E[\hat{y}]$: כאשר נסמן

$$bias^2(\lambda) = \|E[\hat{w}_\lambda] - E[w]\|^2$$
$$= \|(A_\lambda - I) w\|^2$$

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

λ כפונקציה של MSE

$$MSE(\hat{w}_\lambda) = \sigma^2 Tr(A_\lambda (X^T X)^{-1} A_\lambda^T) + \|(A_\lambda - I) w\|^2$$

(e) ראינו שעבור $\alpha=0$ מתקיים $\hat{w} = \hat{w}_{ה}$

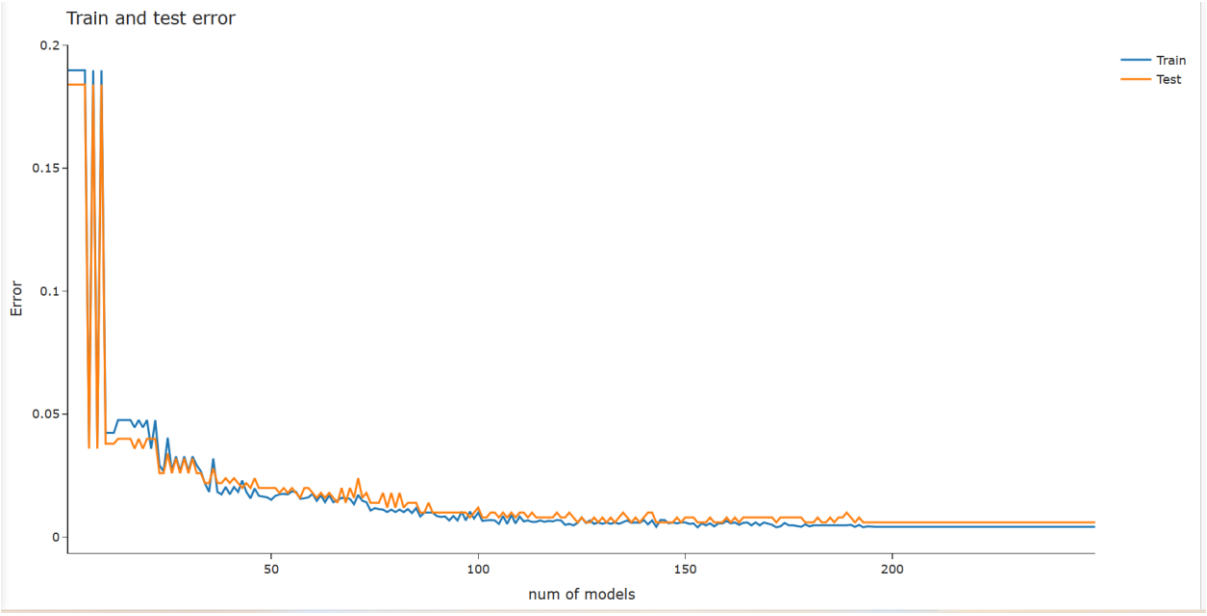
אם מסתכלים סביב לנקודה $\alpha=0$ אולי הפונקציה MSE יורדת בנקודה α אזי קיים $\alpha>0$ כך ש

$$MSE(\hat{w}_{ר"ה}) < MSE(\hat{w}_{\alpha=0}) = MSE(\hat{w})$$

[illegible]

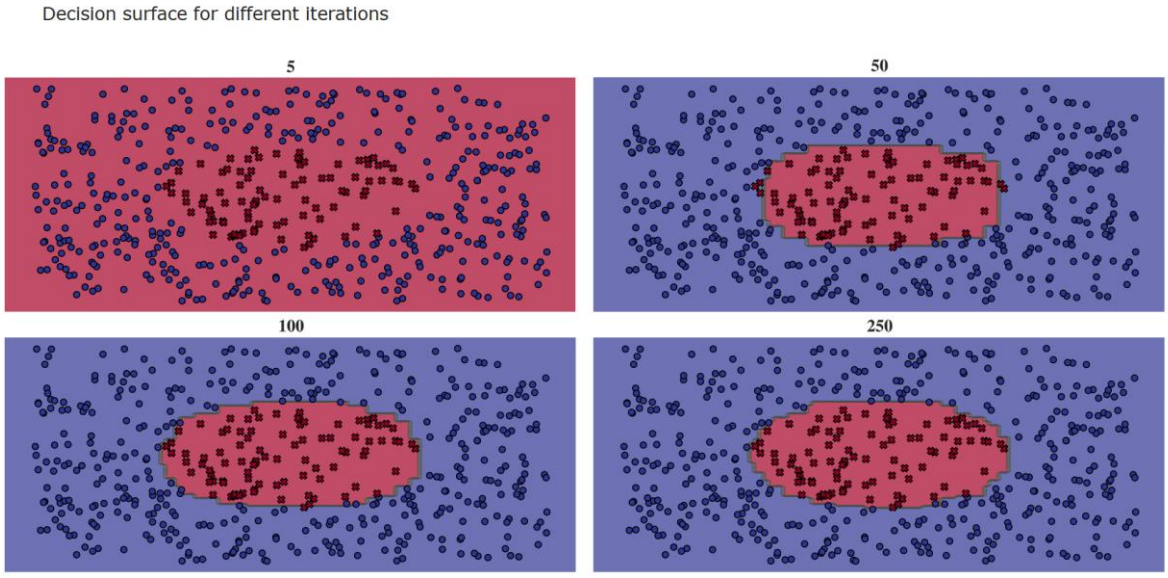
PRACTICAL PART-3.1.1

According to the graph, it can be seen that as the number of iterations increases, the error decreases. This demonstrates what we have learned, that boosting reduces bias more rapidly than it increases variance. Another interesting point observed in the graph is that overfitting does not occur. Although the test error is slightly higher than the training error, at no point do we see a significant increase in test error due to overfitting. (graph 1)

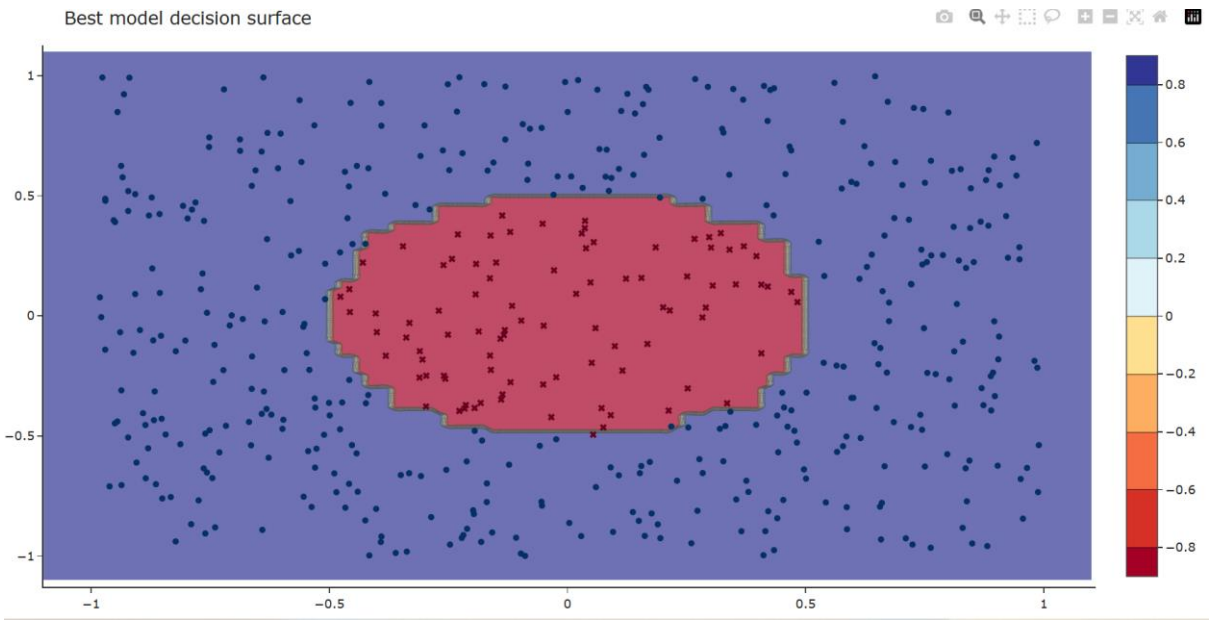


3.1.2

It can be seen that for a small number of iterations (5), the model fails to classify between the red and blue points (as seen in the graph from section 1 where the error is very high). However, when the number of iterations is slightly increased (to 50), it can be observed that the model performs very well in separating the points, as shown in the previous graph. Further increasing the number of iterations only slightly improves the result.

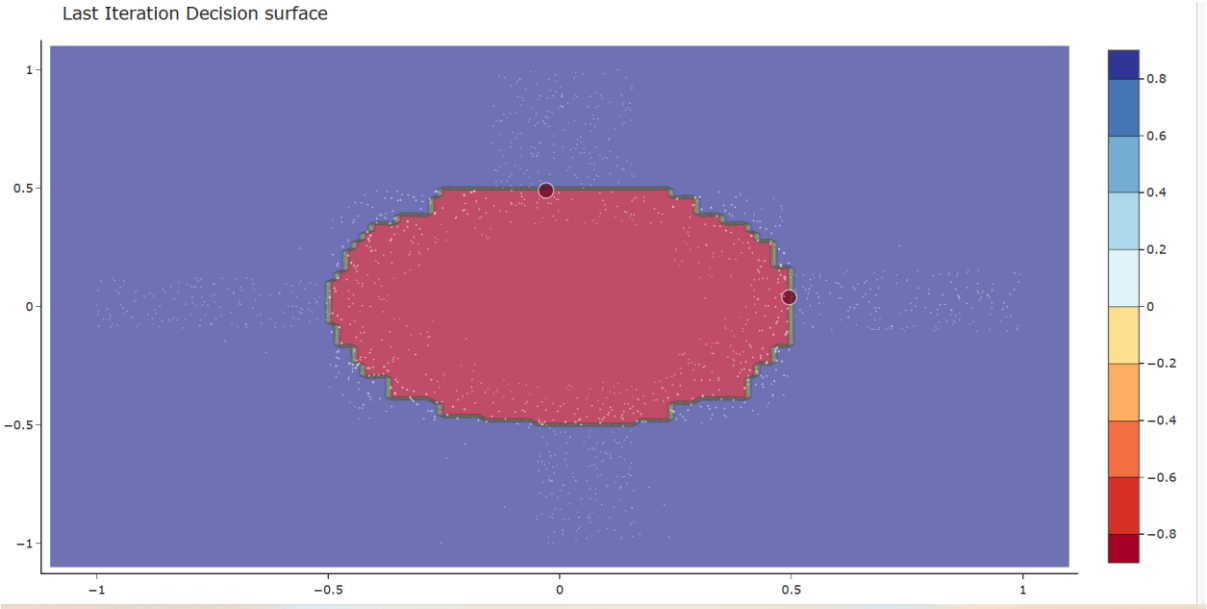


3.1.3


Best model decision surface
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8

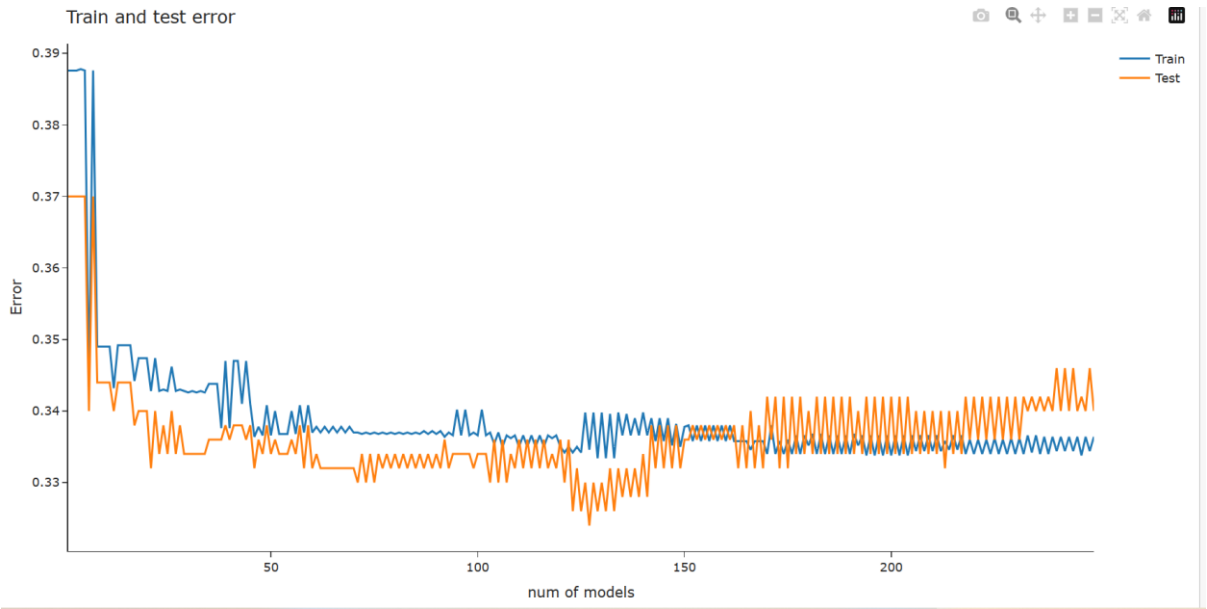
3.1.4

It can be observed that the points on the boundary receive higher weights (the points that are not close to the boundary are hardly visible). In other words, it is challenging for the model to classify .them



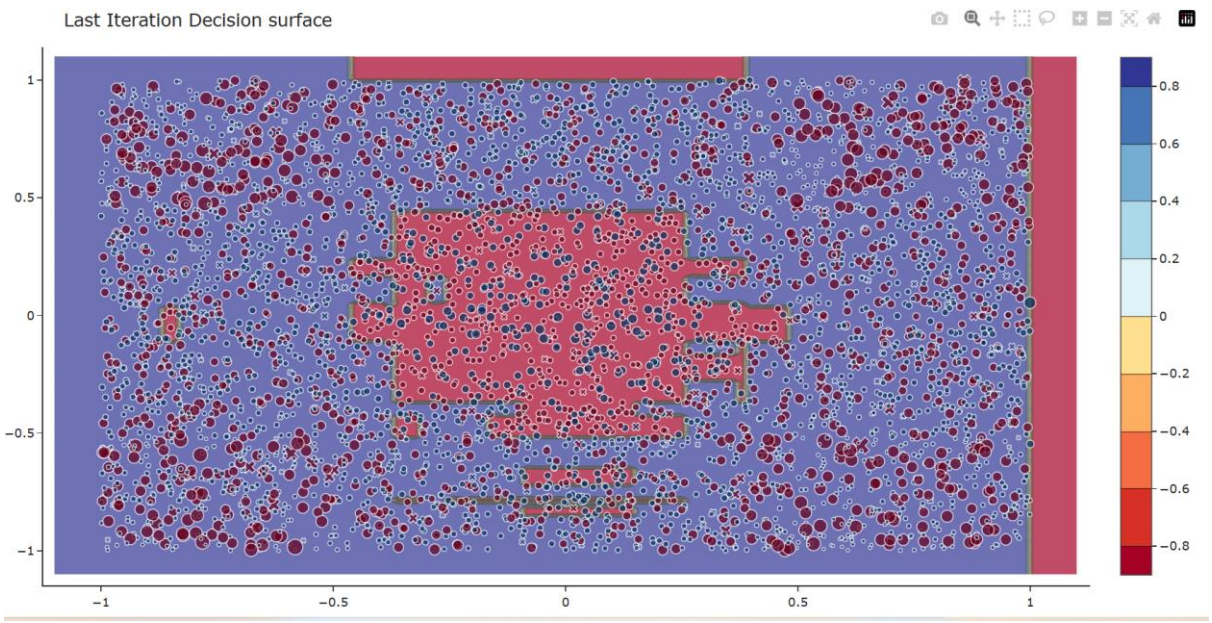
3.1.5

Due to the addition of a large amount of noise, we expect each sample to be significantly different from the other. However, despite this, in the graph, we can see that although the error has increased for both the training set and the test set, it is reasonable because, indeed, due to the high noise, many of the red points are now outside the red circle. Nevertheless, it can be observed that the error for the test set has not increased much compared to the error for the training set. This means that the generalization error has not increased significantly. This is further evidence that .AdaBoost reduces bias significantly more than it increases variance



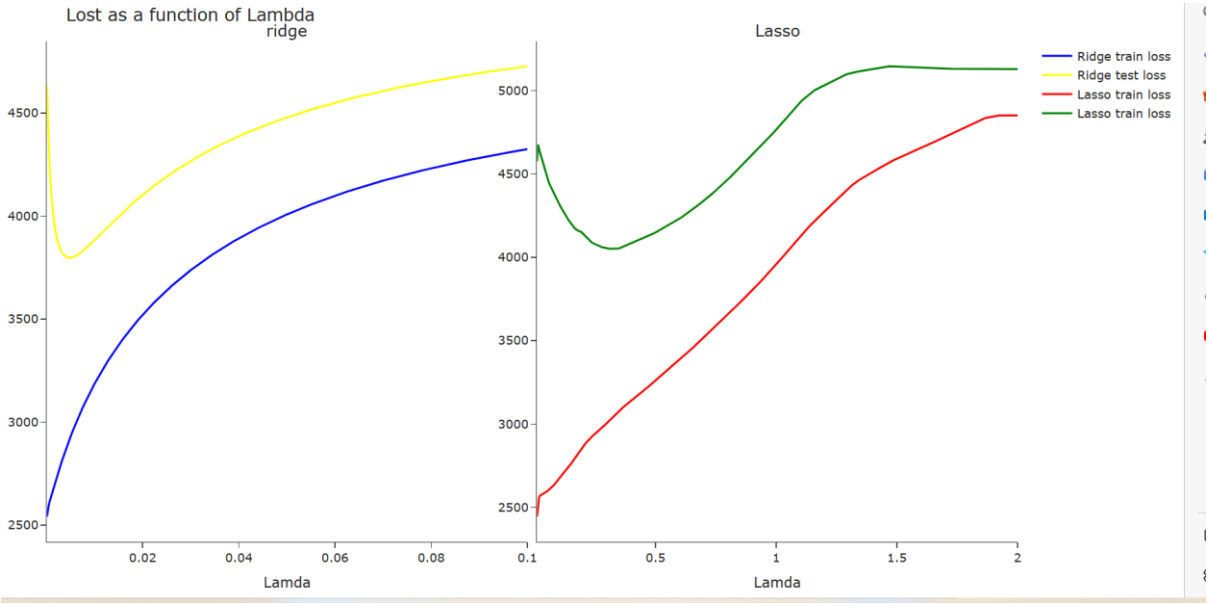
In this graph, it can be observed that despite the noise, the model manages to provide responses that are close to the truth, even though many of the red points are located outside the red circle and many of the blue points are inside it.

It can also be seen that the points in the corners receive the highest weight, which is because they are outside the range of the circle in both coordinates. Therefore, it is logical that most of the weak learners made mistakes regarding these points.



3.2

It can be observed that in both graphs, the train_loss is minimal when λ (lambda) is minimal. This is because when λ=0, we obtain the solution of least squares, which minimizes the empirical error. When λ is increased, we get a simpler model, but its empirical error and generalization error also .increase



3.2.3

```
Best lambda for ridge is:  0.004904809619238477
Best lambda for lasso is:  0.3170741482965932
Best ridge loss is:  4330.988920435104
Best lasso loss is:  4372.179666085873
Linear Regression loss is:  8146.267790957025

>>>
```