אויגמר ניני 207047150

$(b) \Rightarrow (a)$

נניח ש $\lim_{m\to\infty} E_{S\sim D^m}\left[L_D(A(S))\right]=0$

מא"ש מרקוב:

$$P(X \geq a) \leq \frac{E(X)}{a} \Rightarrow$$

$$P_{S\sim D^m}\left[L_D(A(S)) \geq \varepsilon\right] \leq \frac{E_{S\sim D^m}\left[L_D(A(S))\right]}{\varepsilon} \xrightarrow{m\to\infty} 0 \Rightarrow$$

$$P_{S\sim D^m}\left[L_D(A(S)) \leq \varepsilon\right] \geq 1-\frac{E\,L_D(A(S))}{\varepsilon}$$

$$\underbrace{\geq}_{\text{עבור } m \text{ גדול מספיק}} 1-\delta$$

כנדרש

$(a) \Rightarrow (b)$

נ״ח (a):

$$E\left(L_D(A(S))\right) = \int_0^1 x \, P_{S\sim D^m}[L_D(A(S))=x] \, dx$$

$$= \int_0^{\varepsilon} x \, P_{S\sim D^m}[L_D(A(S))=x] \, dx +$$

$$\int_{\varepsilon}^{1} x \, P_{S\sim D^m}[L_D(A(S))=x] \, dx \le$$

$$\le \varepsilon \cdot 1 + 1 \cdot (1-(1-\delta)) =$$

מההנחה:

לכל ε,δ קיים $\bar{m}$ כך שלכל m גדול יותר

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

$$\varepsilon + \delta$$

לכן עבור $\varepsilon,\delta \to 0 \Rightarrow \varepsilon+\delta\to 0$

ומכאן $\lim_{m\to\infty} P_{S\sim D^m}\{L_D(A(S))=x\} \overset{e}{=} 0$

2. צ"ל $\forall \delta, \varepsilon > 0, \forall D$

$$P_{S\sim D^m}(L_D(A(S)) \le \varepsilon) \ge 1-\delta$$

נסמן h^* את ההיפותזה עבורה $L_D(h^*)=0$.

נסמן ב- h_S את המעגל הקטן ביותר המכיל
את כל הנקודות שהתווית שלהם 1 ב-S

נסמן r^* הרדיוס של h^*.
נבחר D על X כך ש-
$$D(\{x : r < \|x\|_2 < r^*\}) = \varepsilon$$

עבור $0<r^*$, $r<r^*$
$$P_{S\sim D^m}(L_D(h_S)>\varepsilon)$$

חסומה ע"י כך שכל $s\in S$ לא
מקיימת את התנאי ב-D כאשר $r_S=r$.

לכן $P_{S\sim D^m}(L_D(h_S)>\varepsilon)\le(1-\varepsilon)^m\le e^{-\varepsilon m}$

כאשר לכל $e^{-\varepsilon m}<\delta$ התנאי מתקיים

ב. צ"ל:

$$D^m(\{S \mid L_D(h_S)\le \min L_D(h^*)+\varepsilon\})\ge 1-\delta$$

כאשר $h_S, h^*\in H$ עבור מס' דגימות

$$m_H(\varepsilon,\delta)\le m_H^{UC}(\varepsilon/2,\delta)$$

מכך ש H היא UC ידוע לנו

שלכל $\varepsilon,\delta>0$ קיימת $m^{UC}(\frac{\varepsilon}{2},\delta)$

עבור ה

$$D^m(\{S \mid |L_S(h)-L_D(h)|<\tfrac{\varepsilon}{2}\})\ge 1-\delta$$

(S מ"צג)

לכן: $\forall h\ L_D(h)\le L_S(h)+\frac{\varepsilon}{2}$

$$L_D(h) \le L_S(h) + \frac{\varepsilon}{2}) \le L_S(h) + \varepsilon$$

אחידה
הכללה

[illegible]
האמפירי

$\Leftarrow$ לכן H agnostic PAC

כנדרש.

VC-Dimension

נתון הבסיס הסטנדרטי בגודל n

נתון ($C = \{e_1, e_2, \dots, e_n\}$)

$I = \{i \mid y_i = 1\}$ לכן מתקיים $h_I(e_i) = y_i$

ולכן בעצם קיבלנו ש

$$\{h_I(e_1), h_I(e_2), \ldots h_I(e_n)\} = \{e_1, \ldots e_n\} \Rightarrow$$

לכן בסיס n $(0,1)^n$ נוכל להסביר כל label. מכאן $VC\text{-}dim(H_I) \geq n$

מצד שני, $|H| = 2^n$, וראינו בהרצאה ש $VCdim(H) \leq \log(|H|) = n$

$$\Rightarrow VC\text{-}dim(H) = n$$

2.

יהיו H_1, H_2 מחלקות

היפותזה כך ש $H_1 \subseteq H_2$.

יהי $n = VCdim(H_1)$

אז קיימת קבוצה $X \in \mathcal{X}$, $|X| = n$ כך

שלכל labeling אפשרי קיימת $h \in H_1$

שמסבירה אותו. אבל $h \in H_1 \subseteq H_2$

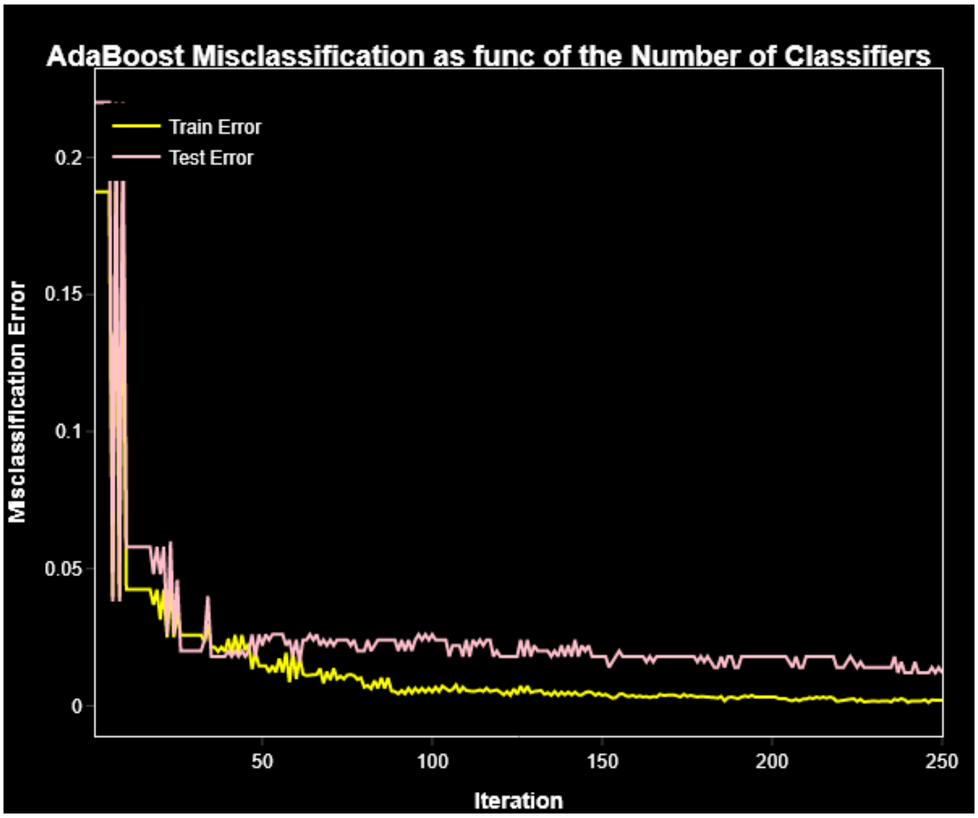
לכן $h \in H_2$ ומכאן שגם H_2

מסבירה כל label עבור הקבוצה

בגודל n ולכן $n \le VCdim(H_2)$

כנדרש

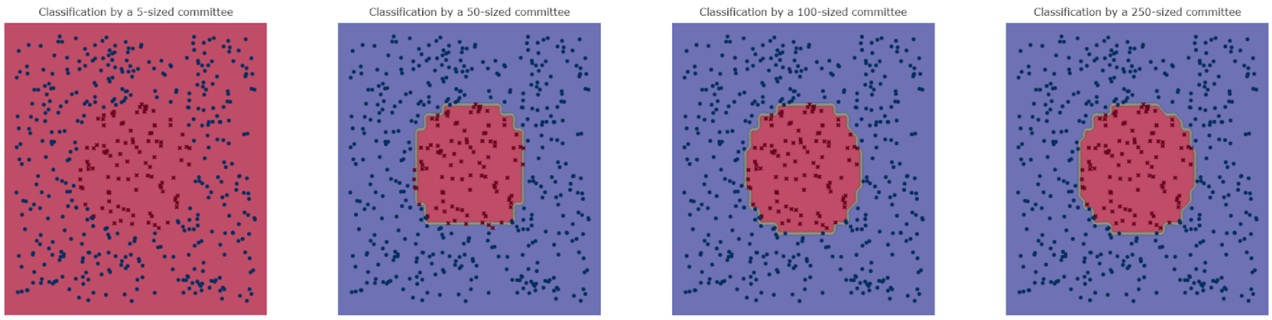
פרקטיקה



לפי גרף השגיאה האמפירית ושל
שגיאות ההכללה ירדו כפונקציה של
גודל הועדה.

כיוון שאין רעש לא נצפה אובר
פיט בכלל, אבל נראה שיכולנו להסתפק
בועדה בגודל קטן מ ~ אומבים.

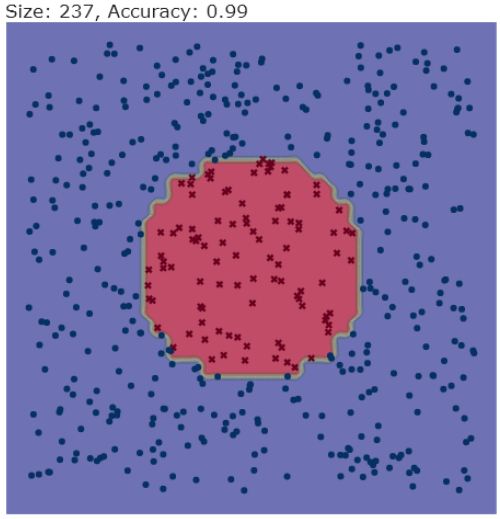
2ב.



ככל שהגדלנו את גודל הועדה קיבלנו הפרדה
טובה יותר. עבור 5 weak learners לא
הצלחנו לקבל הפרדה כלל.
עבור 50 קלסיפיירים קיבלנו כבר הפרדה
טובה שהולכה והפכה מדויקת יותר.
נשים לב שמדובר על רעש פ-

3.

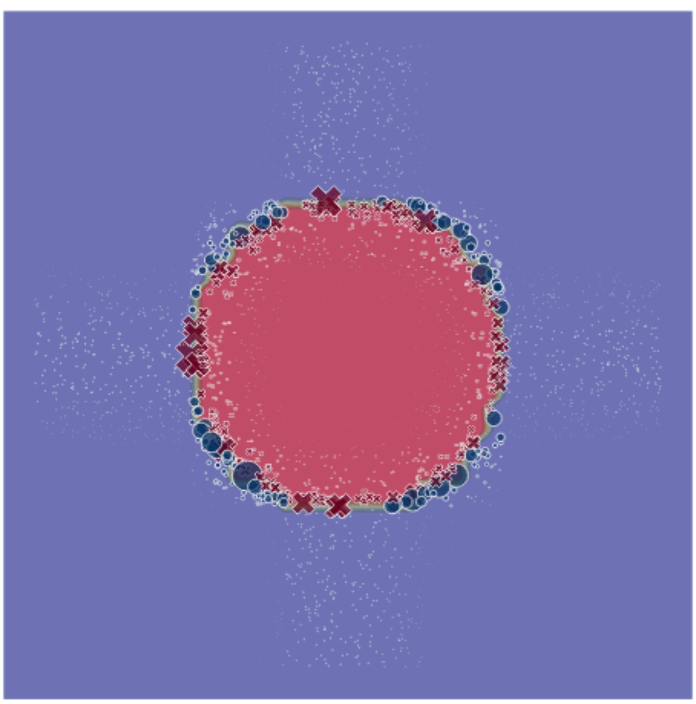
Minimal generalization error committee
Size: 237, Accuracy: 0.99



ט. ככל שגודל הסימון של הדגימה גדול יותר המשקולת שלה גדולה יותר, כלומר המשקולת גדלה מעידה על כך שטעינו בקשר לדגימה באיטרציות קודמות. כך מכאן שדגימות שנמצאות באיזור ה-borderline קבלו משקל גדול יותר זאת על מנת לבדוק אותן כל הזמן.

ככל שדגימה רחוקה יותר מהגבול היא קלה יותר - הסיווג שלה ברור, והדגימות הקשות מתרכזות באיזור הגבול.

Sample Distribution of Final AdaBoost

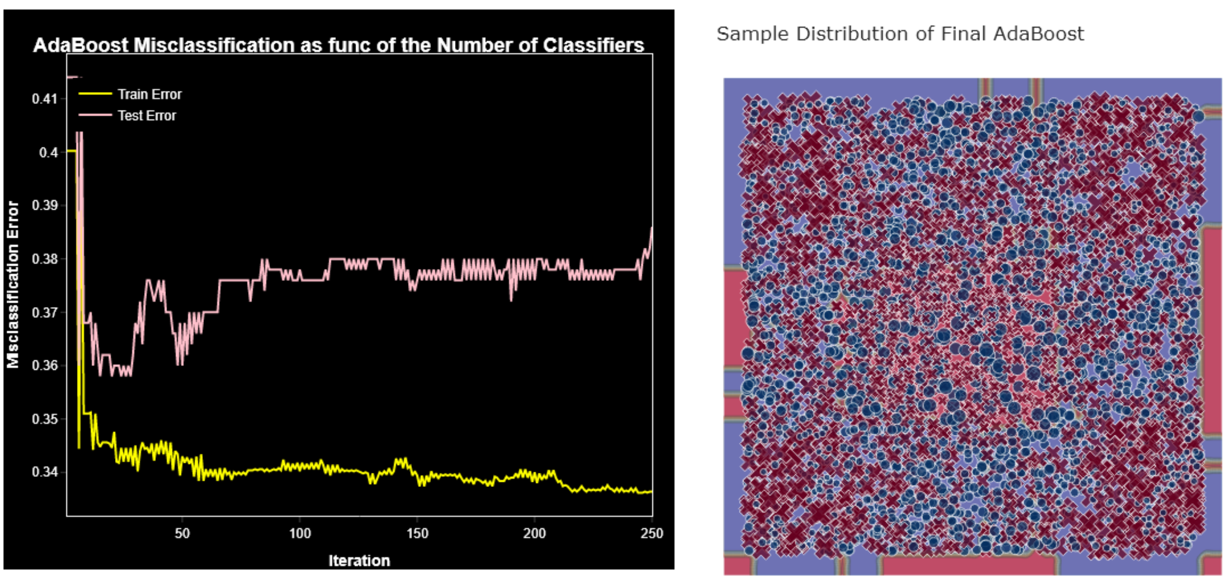


ב.

לגרפים יש שגיאות ל ! ב-ל

עבור רעש מוגבל ט-פ:

ניתן לראות אפילו אובר פיט.



a.

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

LS Solution

$\hat{w} = (X^TX)Xy$

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

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

b.

$$\hat{w}_\lambda = A_\lambda \hat{w} \Rightarrow$$

$$E(\hat{w}_\lambda) = E(A_\lambda \hat{w}) = A_\lambda E(\hat{w})$$

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

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

$E(\hat{w}_\lambda) \neq w$ $\quad \lambda > 0$ מוטה אמדן

.c

$$\mathrm{Var}(\hat{w}_\lambda) = \mathrm{Var}(A_\lambda \hat{w}) = A_\lambda \mathrm{Var}(\hat{w}) A_\lambda^T =$$

$$\sigma^2 A_\lambda (X^T X)^{-1} A_\lambda^T$$

.d

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

$$+ \|\bar{y}-y\|^2 = \mathrm{Var}[\hat{y}] + \mathrm{bias}^2[\hat{y}]$$

$$y = E[\hat{w}],\ \hat{y} = \hat{w}_\lambda,\ \bar{y} = E[\hat{w}_\lambda] \Rightarrow$$

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

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

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

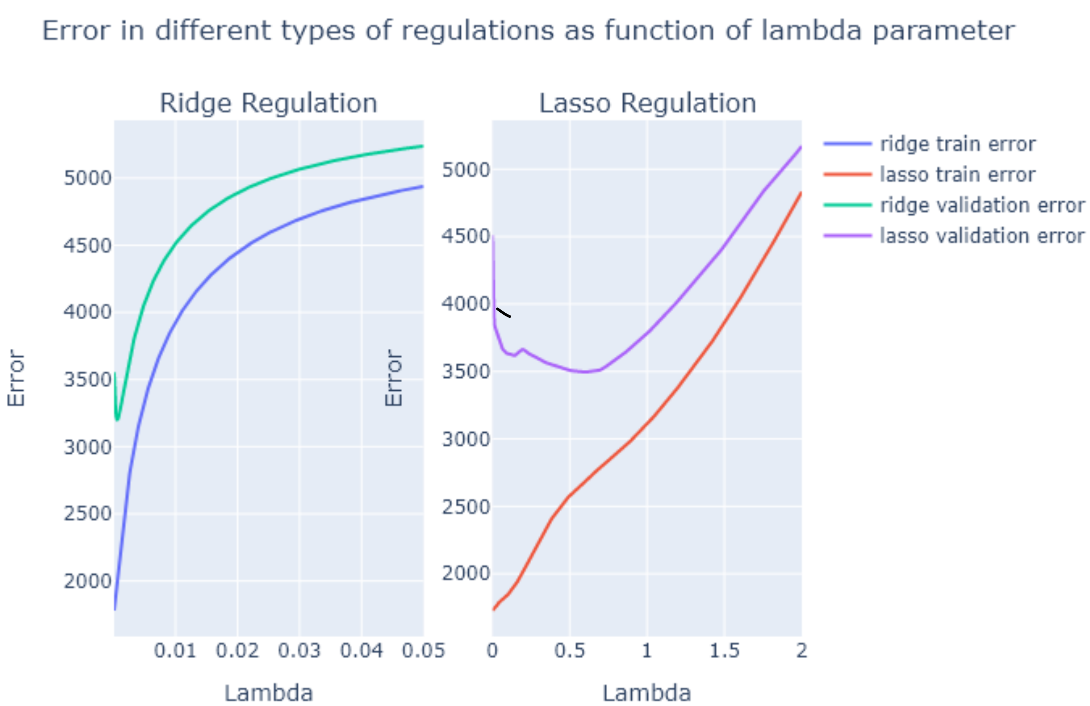
ט)

מהטענה הנתונה שהנגזרת שלילית עבור $\lambda=0$
נסיק שמהגדרת קיים $\lambda>0$ כך ש
$MSE(\lambda) < MSE(0) \Leftarrow$ כמו שאמרנו,

קיימת $\lambda>0$ כך שהסיכון של
רידג' עדיף על פני סיכון ללא כלל
רגולריזציה.

Regularization

2. עבור Ridge הערכים המשמעותיים של λ הם $0-0.01$.
עבור lasso הערכים המשמעותיים הם $0-1$.



עבור Ridge: אפשר לשוות קורלציה בין הטעות על הסט אימון לסט הוולידציה -
כאשר הערך הוולידציה של λ הוא 0.0006
עבור lasso: עד $\lambda \approx 0.6$ האלגוריתם
בוחר משקל מורכב יותר, פשוט, כ- 0.59

נבחר המודל האידיאלי.

3.

```
The regularization parameter values that achieved the best validation errors are:
Ridge lambda: 0.0006000000000000001
Lasso lambda: 0.5972645290581162
The test errors of each of the fitted models are:
Ridge model: 3216.41840256549
Lasso model: 3641.2103701315996
Least squares model: 6103.390440816327
```