שאלה ד

· סט דגימות אימון $S = (z_1, \dots, z_m)$

$z_i = (x_i, y_i)$

· דגימה נוספת z_i'

· נגדיר $S^{(i)} = (z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_m)$

· ראינו כן $U(m) \sim U[1, m]$

· נסמן $f_S(w) = \frac{1}{m} \sum_i^m \ell(w, z_i) + \lambda \|w\|^2$

$$\hat{w} = \arg\min f_S(w)$$

נגדיר:

$$L_D(\hat{w}) = E_{(z) \sim D}\, \ell(\hat{w}, z)$$

$$L_S(\hat{w}) = \frac{1}{m} \sum \ell(\hat{w}, z_i)$$

נוכיח ש:

$$E_{S \sim D^m} \left[L_D(\hat{w}) - L_S(\hat{w}) \right] = E_{(S, z') \sim D^{m+1},\, i \sim U(m)} \left[\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i) \right]$$

* אלגוריתם יציב נקרא כך אם נגרם שינוי קטן בקלט, האלגוריתם לא ישנה את הפלט יותר מידי.
* כדי לבדוק יציבות המודל נבצע החלפה דגימה ה-i ב-S עם הפרמטר z' ונקרא למקבץ $S^{(i)}$ ונבדוק את השפעה על פונקציית ה-loss

הוכחה:

מכיוון שגם S וגם z' שניהם i.i.d מתוך D, ניתן לומר שעבור כל i, נקבל:

$$E_S[L_D(\hat{w})] = E_{S, z}[\ell(\hat{w}, z_i)] = E_{S, z'}[\ell(\hat{w}^{(i)}, z_i)] \quad (1)$$

ניתן לומר גם

$$E_S[L_S(\hat{w})] = E_{S, i}[\ell(\hat{w}, z_i)] \quad (2)$$

נחסר את שני המשוואות (1)-(2) ונקבל מה שרצינו להוכיח.

שאלה 3

נגדיר: $Y = f(x) + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

ונחפש $\hat{f}(x)$ שממזער את השגיאה $E\{(y - \hat{f}(x))^2\}$

נגדיר:

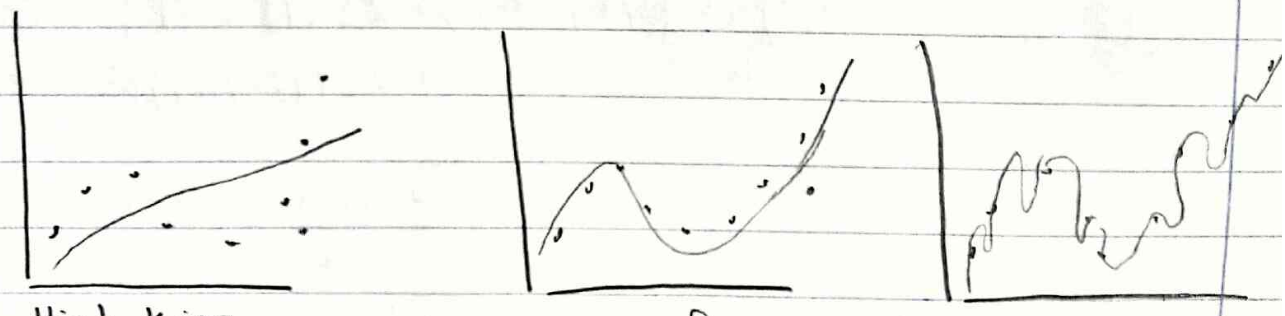
$$Bias\{\hat{f}(x)\} = E\{\hat{f}(x) - f(x)\}$$

$$Var\{f(x)\} = E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2$$

Bias = מודד את הטעות הפרדיקציה של המודל ביחס לערך שעבורו נצפה. Bias גבוה יכול לגרום לאלגוריתם הלמידה לפספס קשרים בין הפיצ'רים והתוצאות גם שגיאה גבוהה (underfitting)

Variance = מודד את השונות הפרדיקציה לנתוני האימון. Variance גבוה גורם למודל הלמידה להתאים לכל רעש בנתוני האימון גם אם הוא רעש רנדומלי. במקום לנסות להבין מגמה כללית. Variance גבוה יגרום לoverfitting

מבחינה ויזואלית:



High Bias
low varians

מודל עם התאמה כללית טובה

low Bias
High variance

• דוגמה של Bias גבוה, זה המצב שהנתונים לא לינאריים, ונצטרך להשתמש במודל לא לינארי במקום מודל לינארי. לדוגמא נשתמש ב- SVM פולינומי במקום SVM לינארי.

• דוגמה ל- variance גבוה, זה קשור לדוגמא הקודמת, נשחק עם הדרגה של הפולינום כי דרגה גבוהה נקבל variance גבוה ודרגה נמוכה נקבל Bias גבוה.

ג.

$$E\{(y-\hat{f}(x))^2\} = (Bias\{\hat{f}(x)\})^2 + Var\{\hat{f}(x)\} + \sigma^2$$

• נסמן $f = f(x),\ \hat{f} = \hat{f}(x)$

• מתוך הגדרת ה-var ←

1) $E(x^2) = var(x) + (E(x))^2$

• מכיוון ש-f היא דטרמיניסטית, ניתן לומר

$E(f) = f$

$\Rightarrow$ 2) $y = f + \varepsilon$, and $E(\varepsilon) \overset{\text{לפי הגדרה}}{=} 0 \Rightarrow E(y) = E(f+\varepsilon) = E(f) + \underbrace{E(\varepsilon)}_{0} = f$

• לפי הגדרה:

$var\{\varepsilon\} = \sigma^2$

• נרצה להוכיח ש-var:

3) $var\{y\} = E\{(y - E(y))^2\} = E\{(y-f)^2\} = E\{(f+\varepsilon-f)^2\} = E(\varepsilon^2)$

(מתוך 1) $= var(\varepsilon) + \underbrace{(E(x))^2}_{\text{0 לפי הגדרה}} = \sigma^2$

• נפתח את הביטוי הכללי.

$$E\{(y-\hat{f})^2\} = E\{(f+\varepsilon-\hat{f})^2\} = E\{(f+\varepsilon-\hat{f}+E(\hat{f})-E(\hat{f}))^2\}$$

$$= E\{(f-E(\hat{f}))^2\} + E(\varepsilon^2) + E((E(\hat{f})-\hat{f})^2) + 2E((f-E(\hat{f}))\cdot \underbrace{E(\varepsilon)}_{0}$$

$$+ 2E((f-E(\hat{f}))\cdot \underbrace{E((E(\hat{f})-\hat{f})}_{0}$$

$$+ \underbrace{E(\varepsilon)}_{0}\cdot \underbrace{E((E(\hat{f})-\hat{f}))}_{0}$$

$$= E\{(f-E(\hat{f}))^2\} + E(\varepsilon^2) + E((E(\hat{f})-\hat{f})^2)$$

$$= (f-E(\hat{f}))^2 + \sigma^2 + var(\hat{f})$$

$$= Bias(\hat{f}(x))^2 + \sigma^2 + var(\hat{f})$$

מ.ש.ל.

2 א. תנאי הליפשיץ אומר כי הפער בין ממוצע השגיאה על מדגם האימון של $\hat{w}$, המסווג הטוב ביותר על מדגם האימון, לבין תוחלת השגיאה ~~השגיאה האמיתית~~ של $\hat{w}$ על נקודה כלשהי "במציאות", חסום ע"י קבוע שהינו פרופורציונלי לקבוע הליפשיץ של פונקציית ההפסד בריבוע, ופרופורציונלי ביחס הפוך לגודל המדגם וקבוע הרגולריזציה.

⇐ ככל שהמדגם גדל, השגיאה על המדגם מקרבת טוב יותר את שגיאת האמת.

ב.

a. נחלק לשלושה מקרים:

(i) $loss(z,w)=loss(z,u)=0$

נקבל: $|loss(z,w)-loss(z,u)| = 0 \le |z|\,\|w-u\|$

(ii) $loss(z,w)=1-y\langle w,x\rangle$, $loss(z,u)=1-y\langle u,x\rangle$

$$|loss(z,w)-loss(z,u)| = |1-y\langle w,x\rangle-(1-y\langle u,x\rangle)|$$
$$= |y|\cdot|\langle w,x\rangle-\langle u,x\rangle| = |(w-u)\cdot x| \le |w-u|\cdot|x|$$

(iii) בה"כ ~~loss(...)~~

$$loss(z,w)=1-y\langle w,x\rangle>0$$
$$loss(z,u)=0>1-y\langle u,x\rangle$$

ונקבל

$$|loss(z,w)-loss(z,u)|$$
$$= |\underbrace{1-y\langle w,x\rangle}_{\ge 0}-0| \le |1-y\langle w,x\rangle-(\underbrace{1-y\langle u,x\rangle}_{<0})| \le |w-u|\cdot|x|$$

לפי מקרה (ii)

ג. ניתן להסיק כי כאשר כל ערכי x של הבעיה שייכים ל-domain החסום ע"י קבוע כלשהו C, soft-SVM תהיה C-ליפשיצית וקמורה ועונה על תנאי הלימוד.

ג. ב. עבור $loss(z,w)=\log(1+e^{-y\langle x,w\rangle})$

נגדיר כ- $f(t)=\log(1+e^{-t})$

מכיוון ש- $0<e^{-t}$ לכל $t\in\mathbb{R}$, $f(t)$ רציפה ב- $\mathbb{R}$

וגזירה, לכן מקיימת את משפט הערך הממוצע

$$f'(c)=\frac{f(a)-f(b)}{a-b} \qquad \forall a,b\in\mathbb{R}$$

עבור c כלשהו $b<c<a$

$$f'(t)=\frac{1}{1+e^{-t}}\cdot(-e^{-t})=-\frac{e^{-t}}{1+e^{-t}}$$

זהו הסיגמואיד (עם סימן מינוס), לכן ערך הנגזרת:

$$|f'(t)|<1 \quad \forall t\in\mathbb{R}$$

ומקבלים

$$\mathrm{I} \quad f(a)-f(b)\le a-b$$

(1-ליפשיץ)

עבור $x,w\in\mathbb{R}$, $logloss(z,w)$ היא הרכבה של $f\circ g(z,w)$, כאשר $g(z,w)=yxw$

לכן מכלל השרשרת $\frac{\partial\, logloss}{\partial w}=\frac{\partial f}{\partial t}\frac{\partial g}{\partial w}\le yx$

ובשילוב I מקבלים כי

$$logloss(z,w)-logloss(z,u)\le|w-u||yx|=|w-u||x|$$

ג. בדומה לסעיף ב' - אם ערכי x מוגבלים ע"י קבוע ρ הפונקציה היא ליפשיצית עבור ρ ומתקיים תנאי [illegible]

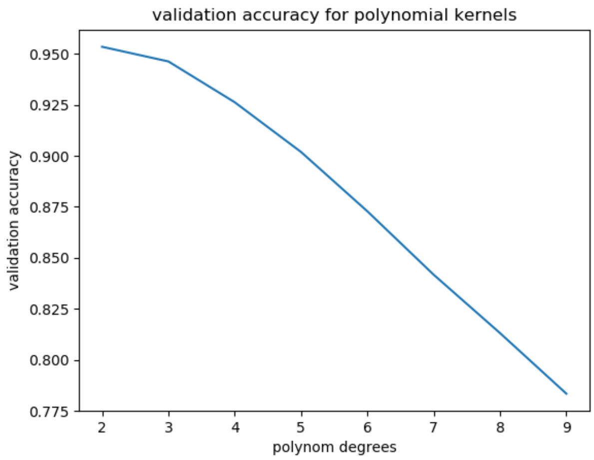
4)

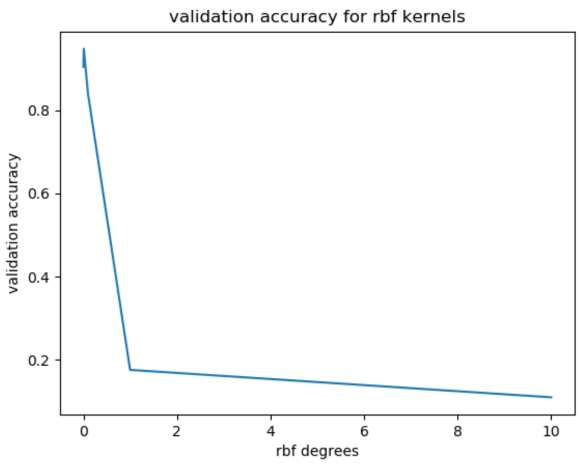
Results:

linear: [1.0, 0.90390625, 0.913125]
poly_2: [1.0, 0.9534374999999999, 0.960625]
poly_3: [1.0, 0.94625, 0.950625]
poly_4: [1.0, 0.9262499999999999, 0.94125]
poly_5: [1.0, 0.9018750000000001, 0.920625]
poly_6: [1.0, 0.8728125, 0.895625]
poly_7: [1.0, 0.8417187500000001, 0.86875]
poly_8: [1.0, 0.813125, 0.845625]
poly_9: [1.0, 0.7834375, 0.8125]
rbf_0.001: [0.9140625, 0.9040625, 0.903125]
rbf_0.01: [0.9776953124999999, 0.9475000000000001, 0.951875]
rbf_0.1: [1.0, 0.83890625, 0.85625]
rbf_1: [1.0, 0.17625000000000002, 0.165625]
rbf_10: [1.0, 0.11046875, 0.105]

the poly_2 model was best on cross validation with error 0.9534374999999999

the poly_2 model was best on test data with error 0.960625





שאלה 2

$z_i=(x_i,y_i)$, $S=(z_1,\dots z_m)$, נגדיר וקטור - z'

נגדיר: $S^{(i)}=(z_1,\dots z'\dots z_m)$

א. המטרה להראות [illegible] נקודת i [illegible] הוספת [illegible] input של [illegible] loss - . $S^{(i)}$ [illegible] הוספת z' ל- S

ב. נסמן $f_S(w)=L_S(w)+\lambda\|w\|^2$ וכך: $\hat{w}=\arg\min f_S(w)$

$\hat{w}$ - [illegible] $f(S)$ [illegible] S.

ג. לכל v,u מתקיים:

$$f_S(v)-f_S(u)=L_S(v)+\lambda\|v\|^2-(L_S(u)+\lambda\|u\|^2)$$

$$=L_{S^{(i)}}(v)+\lambda\|v\|^2-(L_{S^{(i)}}(u)+\lambda\|u\|^2)+\frac{\ell(v,z_i)-\ell(u,z_i)}{m}+\frac{\ell(u,z'_i)-\ell(v,z'_i)}{m}$$

- המעבר האחרון [illegible] z' [illegible] z_i

$$L_S(w)=\frac{1}{m}\sum_{i=1}^{m}\ell(w,z)=L_{S^{(i)}}+\frac{\ell(w,z_i)}{m}-\frac{\ell(u,z'_i)}{m}$$

- [illegible] u,v [illegible].

3. [illegible] $V=\hat{w}^{(i)}$, $U=\hat{w}$. [illegible]

$L_{S^{(i)}}w+\lambda\|w\|^2$ [illegible]

[illegible]:

$$f_S(\hat{w}^{(i)})-f_S(\hat{w})\le\frac{\ell(\hat{w}^{(i)},z_i)-\ell(\hat{w},z_i)}{m}+\frac{\ell(\hat{w},z'_i)-\ell(\hat{w}^{(i)},z'_i)}{m}$$

ב.

$$\frac{\ell(\hat{w},z_i')-\ell(\hat{w}^{(i)},z_i')}{m} + \frac{\ell(\hat{w}^{(i)},z_i)-\ell(\hat{w},z_i)}{m} \ge \lambda\|\hat{w}^{(i)}-\hat{w}\|^2$$

שכן, המשוואה הנ"ל נותנת חסם תחתון להפרש השגיאות כאשר אנו משווים פרדיקציה של מודל על נקודה שלא ראה לעומת נקודה שכבר התאמן עליה. אנו רוצים שיהיה כמה שני מיוצבים (פעם על z_i ופעם על z_i'). משמע החסם הוא למעשה overfit של המודל למדגם האימון.

1. היות שפונקציית ה-loss היא ליפשיץ, לכן לפי הגדרה:

$$1)\quad \ell(\hat{w}^{(i)},z_i)-\ell(\hat{w},z_i) \le \rho|\hat{w}^{(i)}-\hat{w}|$$

ובאופן דומה:

$$2)\quad \ell(\hat{w},z_i')-\ell(\hat{w}^{(i)},z_i') \le \rho|\hat{w}^{(i)}-\hat{w}|$$

נציב את משוואה 1+2 למשוואה המקורית. ונקבל

$$\frac{2\rho|\hat{w}^{(i)}-\hat{w}|}{m} = \frac{\rho|\hat{w}^{(i)}-\hat{w}|}{m} + \frac{\rho|\hat{w}^{(i)}-\hat{w}|}{m} \ge \lambda\|\hat{w}^{(i)}-\hat{w}\|^2$$

$$\Rightarrow \|\hat{w}^{(i)}-\hat{w}\| \le \frac{2\rho}{\lambda m} \Rightarrow \ell(\hat{w}^{(i)},z_i)-\ell(\hat{w},z_i) \le \frac{2\rho^2}{\lambda m}$$

$\Leftarrow$ אנו קיבלנו מהצד השני חסם לערך השגיאות. פונקציית ה-loss היא Θ ליפשיץ ותלוי בקבוע כאשר הוא ρ. ניתן לראות שככל שנגדיל את המדגם, גם ערך החסם ירד. אם נניח שהמדגם מספיק גדול, ניתן לומר שהמודל שלנו לא משתפר יותר מדי. משתנים קטנים והוא מוכלל יציב.