<u>תרגיל בית 3 –Model Selection, Regularization and Stability</u>

יש להגיש קובץ ZIP ובו שני קבצים נפרדים: קובץ PDF ובו פתרון התרגיל כולל הפלטים של החלק המעשי וקובץ נוסף ובו הקוד שכתבתם. יש להקפיד על תשובות ברורות ומסודרות ועל קוד מסודר ומתועד היטב. רק אחד מבין חברי הזוג צריך להגיש את הפתרון.

שאלות על התרגיל יש לכתוב בפורום תרגילי הבית באתר הקורס . התרגיל מנוסח בלשון נקבה אך מתייחס לשני המינים. שאלות על התרגיל יש לכתוב בפורום תרגילי הבית באתר הקורס. התרגיל מנוסח בלשון נקבה אך מתייחס לשני המינים.

<u>שאלה 1</u>

יהי $S = (z_1, \dots, z_m)$ כאשר $z_i = (x_i, y_i)$ סט תצפיות אימון ותהי z_i' תצפית נוספת.

נגדיר $S^{(i)} = (z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_m)$, כמו כן $U(m) \sim U[1, m]$

נסמן $f_S(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i) + \lambda\|w\|^2$ ויהי $\hat{w} = \underset{w}{argmin}\, f_S(w)$

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

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

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

<u>שאלה 2</u>

למדנו בהרצאה כי כלל למידה יציב נמנע מהתאמת יתר לנתונים. נבחן למידה באמצעות regularized loss minimization:

יהי $S = (z_1, \dots, z_m)$ כאשר $z_i = (x_i, y_i)$ סט תצפיות אימון

נסמן $f_S(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i) + \lambda\|w\|^2$ ויהי $\hat{w} = \underset{w}{argmin}\, f_S(w)$

בהרצאה הוכחנו כי עבור ρ-Lipschitz loss function ($|\ell(w,z) - \ell(u,z)| \le \rho\|w - u\|$) מתקיים תנאי היציבות הבא:

$$\left|\frac{1}{m}\sum_{i=1}^{m}\ell(\widehat{w},z_i) - E_{(z)\sim D}\ell(\widehat{w},z)\right| \le \frac{2\rho^2}{\lambda|S|}$$

א. הסבירי במילים את תנאי היציבות.

ב. עבור פונקציית hinge loss - $loss(z,w) = max\{0,1-y\langle w,x\rangle\}$ אשר בה משתמשים בשיטת soft-SVM

a. הוכיחי כי הפונקציה היא $|x| - Lipschitz$, כאשר $x,w \in \mathbb{R}$

כלומר מתקיים: $|loss(z,w) - loss(z,u)| \le |x||w-u|$

b. מה ניתן להסיק לגבי soft-SVM?

<u>סעיף בונוס</u>

ג. עבור פונקציית log loss - $loss(z,w) = log\left(1+e^{-y\langle w,x\rangle}\right)$ כאשר $x,w \in \mathbb{R}$ אשר בה משתמשים ברגרסיה לוגיסטית

c. הראי כי הפונקציה הינה |x|-Lipschitz כאשר $x,w \in \mathbb{R}$,

הכוונה: ראשית הראי כי הפונקציה $f(t) = log(1+e^{-t})$ הינה 1-Lipschitz.

לפי משפט הערך הממוצע של לגראנז' (זוכרים?):

עבור פונקציה $f(t)$ רציפה בתחום $[a,b]$ וגזירה בתחום (a,b) קיימת נקודה $a<c<b$ עבורה:

$$f'(c) = \frac{f(b)-f(a)}{b-a}$$

מה ערך הנגזרת עבור הפונקציה $f(t) = log(1+e^{-t})$? האם ניתן לחסום את $|f(t)|$?

d. מה ניתן להסיק לגבי רגרסיה לוגיסטית?

<u>שאלה 3</u>

בשאלה זו נעסוק ב *Bias Variance trade-off* עליה דיברנו בכיתה.

נניח נתונים מן הסוג הבא: $Y = f(x)+\epsilon, \epsilon \sim N(0,\sigma^2)$

אנו מנסים למצוא $\hat{f}(x)$, כך שתמזער את השגיאה : $E[(Y-\hat{f}(x))^2]$.

הגדרות נוספות : $Bias[\hat{f}(x)] = E[\hat{f}(x) - f(x)], Var[f(x)] = E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2$

א. הסבירו מה מייצגים ה-*bias* וה-*variance* במודל? מה ניתן לעשות במקרה של *bias* גבוה? מה ניתן לעשות במקרה של *variance* גבוה?

ב. הראו כי מתקיים : $E[(Y-\hat{f}(x))^2] = (Bias[\hat{f}(x)])^2 + Var[\hat{f}(x)] + \sigma^2$

שאלה 4

1. כחלק מתהליך בחירת המודלים למדנו על מדד ה CV כשיטה לבחירת מודל.

כתבו פונקציה "cv" המקבלת את הפרמטרים הבאים (החתימה של הפונקציה מופיעה בקובץ הקוד המצורף לתרגיל):

X – Dataframe של המשתנים המסבירים

y- Dataframe של משתנה התגובה

Model – משתנה המכיל מודל מסויים (לדוג' רגרסיה לינארית)

Folds- מספר החלוקות אותו אנו מעוניינים לבצע

הפונקציה צריכה לחשב ולהחזיר את שגיאת האימון ושגיאת הוולידציה (ממוצעות על פני כל ה-folds) עבור המודל שהתקבל באמצעות שיטת CV בשימוש ב#Folds

השגיאה תחושב באמצעות מדד 1-accuracy (אחוז הדגימות בהן שגינו מתוך המדגם)

2. נשתמש שוב בסט הנתונים MNIST מש.ב 2.

טענו את קובץ הנתונים (בשימוש בפונקציה המוכנה) , וחלקו למדגם אימון ומבחן ביחס של 80/20.

כעת, השתמשו בפונקציה cv משאלה 4 סעיף 2 (folds=5) והריצו על מדגם האימון מודל svm עם ה-kernels הבאים:

1.לינארי

2.פולינומיאלי - עם דרגות בטווח [2,3,4,5,6,7,8,9,10]

3.רדיאלי עם פרמטרי gamma בטווח [0.001,0.01,0.1,1,10]

לאחר מכן התאימו כל מודל על מדגם האימון כולו וחשבו את השגיאה על מדגם המבחן.

החזירו dictionary כאשר ה-key הוא שם המודל (לדוג': svm_poly_d5) וה-value הוא list עם ערכי שגיאת האימון, שגיאת הוולידציה ושגיאת המבחן של המודל (לדוג' [0.3,0.4,0.5])

מהו המודל הטוב ביותר לפי שיטת cv? מהו המודל שביצע הטוב ביותר על מדגם המבחן?

בנוסף, הציגו גרפים של שגיאת הולידציה של הקרנלים : פולינומיאלי ורדיאלי כתלות בפרמטר שלהם.

הבהרה כללית- בשגיאת האימון הכוונה היא לשגיאה שהוחזרה מפונקציית ה-cv.

שאלה 5 -בונוס

בהראה למדנו על Regularized Loss Minimization (RLM) וראינו כי היא מספקת פתרון יציב שאינו מביא להתאמת יתר. בשאלה זו נעקוב אחר חלק הפיתוח שלא ראיתם בהרצאה כאשר המטרה תהיה להסביר כל שלב ולהצדיק אותו.

יהי $S = (z_1, \dots, z_m)$ כאשר $z_i = (x_i, y_i)$ סט תצפיות אימון ותהי z_i' תצפית נוספת.

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

א. הסבירי מהו $S^{(i)}$

נסמן $f_S(w) = L_S(w) + \lambda\|w\|^2$ ויהי $\hat{w} = \underset{w}{argmin}\, f_S(w)$

ב. הסבירי מיהו $\hat{w}$

לכל v, u ולכל i מתקיים:

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

$$L_{S^{(i)}}(v) + \lambda\|v\|^2 - \left(L_{S^{(i)}}(u) + \lambda\|u\|^2\right) + \frac{l(v, z_i) - l(u, z_i)}{m} + \frac{l(u, z_i') - l(v, z_i')}{m}$$

ג. הסבירי את המעבר האחרון

מהנ"ל נובע כי

$$f_S\left(\widehat{w^{(i)}}\right) - f_S(\hat{w}) \leq \frac{l\left(\widehat{w^{(i)}}, z_i\right) - l(\hat{w}, z_i)}{m} + \frac{l(\hat{w}, z_i') - l\left(\widehat{w^{(i)}}, z_i'\right)}{m}$$

ד. הצדיקי את הטענה

כמו כן נתון כי מתקיים לכל v:

$$f_S(v) - f_S(\hat{w}) \geq \lambda\|v - \hat{w}\|^2$$

ומכאן נובע:

$$\lambda\left\|\widehat{w^{(i)}} - \hat{w}\right\|^2 \leq \frac{l\left(\widehat{w^{(i)}}, z_i\right) - l(\hat{w}, z_i)}{m} + \frac{l(\hat{w}, z_i') - l\left(\widehat{w^{(i)}}, z_i'\right)}{m}$$

ה. הסבירי במילים את המסקנה הנ"ל

ו.

כעת, עבור ρ-Lipschitz loss function מתקיים כי:

$$l\left(\widehat{w^{(i)}}\right) - l(\hat{w}, z_i) \leq \rho\left\|\widehat{w^{(i)}} - \hat{w}\right\|$$

$$l(\hat{w}, z_i') - l\left(\widehat{w^{(i)}}, z_i'\right) \le \rho \left\| \widehat{w^{(i)}} - \hat{w} \right\|$$

ומכאן ש:

$$\lambda \left\| \widehat{w^{(i)}} - \hat{w} \right\|^2 \le \frac{2\rho \left\| \widehat{w^{(i)}} - \hat{w} \right\|}{m}$$

ולכן:

$$\left\| \widehat{w^{(i)}} - \hat{w} \right\| \le \frac{2\rho}{\lambda m}$$

ז. הסבירי את התוצאה הנ"ל

ובפרט מתקבל:

$$l\left(\widehat{w^{(i)}}, z_i\right) - l(\hat{w}, z_i) \le \frac{2\rho^2}{\lambda m}$$

ז. הסבירי את התוצאה ומהי המסקנה שניתן להסיק ממנה לגבי יציבות המודל שמתקבל באמצעות RLM