תרגיל בית 4 –רגולריזציה ועקומת ROC

יש להגיש 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.

שאלה 1

נתון מודל רגרסיה פשוטה: $Y_i = w_0 + w_1 X_i + \varepsilon_i \;\; i = 1, \dots, n$ כאשר מתקיים $\sum_{i=1}^{n} X_i = 0$

א. הראי שאומד *Ridge* ל-$\widehat{w}_1$ עבור פרמטר λ מסוים הוא:

$$\widehat{w}_1 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2 + \lambda}$$

ב. הראי שהתוחלת והשונות של האומד מסעיף א' הינן:

$$E(\widehat{w}_1) = w_1 \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2 + \lambda} \qquad Var(\widehat{w}_1) = \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^2}$$

ג. האם ההטיה בריבוע של האומד מסעיף א', $(E[\widehat{w}_1 - w_1])^2$ עולה או יורדת ב-λ? האם השונות של האומד עולה או יורדת ב-λ? הסבירי את התוצאה תוך התייחסות למטרה של השימוש ב-*Ridge*.

שאלה 2

(רק תשובה הכוללת נימוקים מתמטיים ומילוליים מלאים תקבל את מלוא הנקודות)

נתון מודל רגרסיה לינארית פשוטה עם שתי תצפיות:

$$Y_1 = w_1 + w_2 X_1 + \epsilon_1, \;\; Y_2 = w_1 + w_2 X_2 + \epsilon_2$$

נתון כי התצפיות ממורכזות, כלומר $X_1 + X_2 = Y_1 + Y_2 = 0$.

א. הראו כי אמד ריבועים פחותים הינו $\widehat{w}_1 = 0$ וכן $\widehat{w}_2 = \frac{Y_1}{X_1}$.

ב. הראו כי אמד *Ridge* הינו $\widehat{w}_1^R = 0$, וכן $\widehat{w}_2^R = \frac{Y_1 X_1}{X_1^2 + \frac{\lambda}{2}}$.

ג. הראו כי עבור *Lasso* מתקיים $\widehat{w}_1^L = 0$, וכן $\widehat{w}_2^L$ פותר את בעיית המינימיזציה הבאה:

$$\underset{\widetilde{w}}{\operatorname{argmin}}\, L(\widetilde{w}) \text{ , where } L(\widetilde{w}) = 2(Y_1 - \widetilde{w} X_1)^2 + \lambda |\widetilde{w}|$$

ד. הראו כי אם בנוסף $4|X_1 Y_1| < \lambda$, מתקיים $\widehat{w}_2^L = 0$

(רמז: הראו כי $L(0) = 2Y_1^2$ וכן שכאשר $4|X_1 Y_1| < \lambda$ אז לכל $\widetilde{w}$ מתקיים $L(\widetilde{w}) \geq 2Y_1^2$).

שאלה 3

בהתאמת מודל רגרסיה לוגיסטית על נתונים בעלי שני משתנים מסבירים - x_1, x_2 התקבלו המקדמים הבאים:

$w_0 = 0.1, w_1 = -0.1, w_2 = 0.3$.

א. כתוב את הנוסחה עבור $\hat{P}(Y = 1|X = x)$

כמו כן נתונות 5 התצפיות הבאות:

i	1	2	3	4	5
x_{i1}	10	2	15	2	8
x_{i2}	2	3	1	1	1
Y_i	0	1	0	0	1

ב. חשב את $\hat{P}(Y = 1|X = x)$ עבור חמש התצפיות

ג. צייר את עקום *ROC* עבור חמש התצפיות והמודל הנתון

שאלה 4

בקוד המצורף לתרגיל בקובץ *main.py* ישנן חתימות לפונקציות שתממשי בשאלה 4 כולל הסבר. את רשאית לכתוב פונקציות עזר, אך את חתימות פונקציות אלו אסור לשנות מכיוון שבדיקת התרגילים מתבצעת באופן אוטומטי ומסתמכת על שמות חתימות אלו.

בשאלה זו תשתמשי בנתונים *iris dataset* מ- *sklearn* אשר מכילים 150 תצפיות של שלושה זנים של אירוסים.

X היא *nparray* בעלת 150 שורות ו4 עמודות: Sepal Length, Sepal Width, Petal Length and Petal Width

y הוא *nparray* בעל 150 שורות ועמודה אחת בעל הערכים האפשריים: 0, 1, או 2 המייצגים 3 זנים של אירוסים Setosa, Versicolour, and Virginicacv בהתאמה.

בשאלה תשתמשי במסווג רגרסיה לוגיסטית בכדי לממש את המדדים:

micro average precision, micro average recall, micro average false positive rate, f_β

א. ממשי את הפונקציה *adjust_labels_to_binary(y_train, target_class_value)* המקבלת את *y_train* כמערך *np* ואת *target_class_value* כמחרוזת המייצגת את אחת המחלקות ומחזירה מערך *np* זהה מבחינת מימדים ל *y_train* שבו ערכים 1 עבור המחלקה *target_class_value* ו -1 אחרת.

ב. ממשי את הפונקציה *one_vs_rest(x_train, y_train, target_class_value)* המקבלת את *y_train* ו- *x_train* כמערכי *np* ואת *target_class_value* כמחרוזת המייצגת את אחת המחלקות. הפונקציה משתמשת בפונקציה מסעיף א בכדי ליצור *y_train_binarized* ומחזירה מודל רגרסיה לוגיסטית שאומן על *x_train* ו- *y_train_binarized*.

ג. ממשי את הפונקציה

binarized_confusion_matrix(X, y_binarized, one_vs_rest_model, prob_threshold)

המקבלת את *X, y_binarized* כמערכי *np* כאשר *X* קבוצת נתונים ו- *y_binarized* משתנה התגובה המתאים אחרי פעולת בינאריזציה בהתאם למחלקה המדוברת, את *one_vs_rest_model* המתאים ל *y_binarized* כאובייקט מודל, ואת *prob_threshold* ערך סף להסתברות שאם ההסתברות החזויה על ידי המודל גדולה או שווה לה הוא יחזה 1 ואחרת -1.

הפונקציה מחזירה מערך *np* של מטריצת הבלבול המתאימה למדגם *X, y_binarized* באופן הבא:

[TP, FN

FP, TN]

ד. הציגי את מטריצות הבלבול עבור נתוני האימון ונתוני המבחן עבור ערך סף להסתברות 0.5 לכל אחת מהמחלקות 0, 1 ו- 2

ה. ממוצע משוקלל של מדד הדיוק *micro average precision* מחושב באופן הבא:

$$\frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i}$$

כאשר לכל מחלקה *i* *TPi, FPi, FNi, TNi* סופרים את מספר התצפיות בהתאם לחיזוי של מודל *one_vs_rest* עבור *prob_threshold*.

ממשי את הפונקציה

micro_avg_precision(X, y, all_target_class_dict, prob_threshold)

המקבלת את *X, y* כמערכי *np* כאשר *X* קבוצת נתונים ו- *y* משתנה התגובה המתאים, את *all_target_class_dict* מילון לכל המחלקות עם מפתח מחרוזת המייצגת מחלקה ואובייקט מודל *one_vs_rest* המתאים למחלקה במפתח, ואת *prob_threshold* ערך סף להסתברות ומחזירה את ערך *micro average precision* של המדגם *X, y*.

ו. ממוצע משוקלל של מדד ה*micro average recall - recall* מחושב באופן הבא:

$$\frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i}$$

באופן דומה לסעיף ה, ממשי את הפונקציה

micro_avg_recall(X, y, all_target_class_dict, prob_threshold)

ז. באופן דומה לסעיף ה, ממשי את הפונקציה

micro_avg_false_positve_rate(X, y, all_target_class_dict, prob_threshold) בהתאם לנוסחה הבאה:

$$\frac{\sum_i FP_i}{\sum_i TN_i + \sum_i FP_i}$$

ח. הציגי גרף *ROC* של *micro average recall* כפונקציה של *micro_avg_false_positve_rate* עבור ערכי סף להסתברות *[0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75]* של מדגם האימון

ט. מדד f_β מחושב באופן הבא:

$$f_\beta = (1+\beta^2)\cdot\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

ממשי את הפונקציה

f_beta(precision, recall, beta)

י. הציגי גרף של f_β כפונקציה של $\beta \in [0,10]$ עבור *micro average recall* ו- *micro average precision* המחושבים על מדגם האימון עם ערכי הסף 0.3, 0.5 ו- 0.7