תרגיל בית 1 – רגרסיה לינארית ורגרסיה לוגיסטית

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

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

נתון מודל רגרסיה לינארית פשוטה עם שתי תצפיות: $Y_1 = w_1 + w_2X_1 + \epsilon_1,\ \ Y_2 = w_1 + w_2X_2 + \epsilon_2$

א. חשבי ומצאי אמד ריבועים פחותים מתאים $\widehat{w}_1$ ו- $\widehat{w}_2$ ורשמי אותם כפונקציה של $\bar{y}$ ו- $\bar{x}$ בין היתר כאשר $\bar{x} = \frac{1}{2}\sum_{i=1}^{2} x_i$ ו- $\bar{y} = \frac{1}{2}\sum_{i=1}^{2} y_i$

ב. הראי כי האמד $\widehat{w}_1$ ו- $\widehat{w}_2$ הוא אמד חסר הטיה

שאלה 2

באתר הקורס נמצא קובץ בשם "parkinsons_updrs_data.csv" ובו 5,875 רשומות. קובץ הנתונים מכיל מאפיינים של מדידות הקלטות דיבור של 42 חולי פרקינסון ותוצאות בדיקות רפואיות שנערכו להם. בתרגיל זה ננסה לחזות את מצבם הרפואי של חולי פרקינסון באמצעות נתוני הדיבור. קובץ בשם "parkinsons_updrs.names.txt" ובו הסבר על הנתונים נמצא גם כן באתר הקורס.

א. טעני את קובץ הנתונים לסביבת העבודה.

ב. הדפיסי סיכום של המשתנים שבקובץ הנתונים. תארי בקצרה את הפלט עבור שניים מהמשתנים.

ג. בחרי שישה משתנים מסבירים והדפיסי גרף פיזור שלהם ביחד עם משתנה התגובה motor_UPDRS (היעזרי בפונק' scatter_matrix())

כעת נבנה מודל רגרסיה לינארית מרובה על מנת לחזות את המשתנה motor_UPDRS באמצעות המשתנים שבחרת. ראשית, כדי לתרגל פעולות אריתמטיות בשפת Python וכן לחדד את ההבנה בנוגע לחישוב אמדי ריבועים פחותים,

ד. כתבי פונקציה אשר מקבלת *nparray* של תצפיות $\boldsymbol{X}$ ו *nparray* של משתנה התגובה $\boldsymbol{y}$ ומחזירה את ערך אמד הריבועים הפחותים $\widehat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

ה. השתמשי בפונ' מן הסעיף הקודם כדי לקבל את ערך האמד עבור התצפיות בקובץ הנתונים

ו. השתמשי בפונקציה המובנית ב-Python עבור מודל רגרסיה לינארית לחישוב האמד. האם קיבלת את אותו הערך? הציגי את סיכום המודל שקיבלת.

ז. עבור מי מהמשתנים המסבירים נוכל לדחות את השערת האפס לפיה $H_o: w_j = 0$ ברמת מובהקות $\alpha = 0.01$? באיזה מבחן נשתמש לשם כך? גבי את תשובתך בנתונים המתאימים מהפלט שהצגת. מה תהיה התשובה אם רמת המובהקות תהיה $\alpha = 0.001$?

שאלה 3

בתרגיל זה תפתחי פתרון לבעיית Logistic Regression בשיטה של פונ' הפסד (לסיווג בינארי עם תוויות $\mathbf{1}, -\mathbf{1}$). נזכיר כי הצגנו לבעיה פתרון בשיטת נראות מירבית, כאשר

$$\boldsymbol{Pr(y_i = 1|x_i) = \frac{e^{w^T x_i}}{1 + e^{w^T x_i}}}, \qquad \boldsymbol{Pr(y_i = -1|x_i) = 1 - p(y_i = 1|x_i)}$$

א. רשמי את לוג הנראות המירבית, $\boldsymbol{l(w)}$, במונחי ההסתברויות הנ"ל (בעיית מקסימיזציה).

ב. רינת מציעה את פונקצית הקנס $\boldsymbol{logloss(w, x_i, y_i) = \log(1 + e^{-y_i w^T x_i})}$.
וטוענת כי הבעיה שהצגת בסעיף א' שקולה לבעיה הבאה:

$$\boldsymbol{argmin_w \sum_{i=1}^{m} logloss(w, x_i, y_i)}$$

הראי כי טענתה של רינת נכונה.

שאלה 4

בשאלה זו תשתמשי בנתונים *iris dataset* מ- *sklearn* אשר מכילים 150 תצפיות של שלושה זנים של אירוסים. בשאלה תשתמשי במסווגים של רגרסיה לוגיסטית בכדי לבנות מסווג *multi-class* מסוג *one-versus-rest*.

א. טעני את הנתונים בעזרת:

from sklearn import datasets

iris = datasets.load_iris()

X = iris.data

y = iris.target

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.

ב. חלקי את הנתונים לנתוני אימון ומדגם בעזרת *train_test_split* השתמשי ב *random_state=1000*

ג. מסווג של רגרסיה לוגיסטית הינו מסווג בינארי. בנוסף, המסווג מספק רמת בטחון, כלומר ההסתברות להשתייך למחלקה אחת מבין השתיים. בסעיפים הבאים תבני מסווג *one-versus-rest* עבור נתוני האירוסים באופן שיפורט להלן. ראשית, בני מסווג של רגרסיה לוגיסטית לזן האירוס שמיוצג על ידי הספרה אפס. לשם כך עליך לייצר וקטור *(nparray)* חדש מווקטור משתנה התגובה שיהיה בו 1 עבור זן האירוס Setosa המיוצג על ידי אפס ו 1- עבור שני הזנים האחרים (המיוצגים על ידי 1 ו-2). התאימי את המסווג על נתוני האימון. שימי לב: בהמשך תשתמשי בהסתברות שמספק המסווג הזה.

ד. באופן דומה, בני שני מסווגים נוספים עבור זן האירוס Versicolour המיוצג על ידי הספרה 1 ו Virginicacv המיוצג על ידי הספרה 2, בהתאמה.

ה. בני פונקציה שמקבלת את כל המסווגים לעיל ואוסף תצפיות כמערך *np* ומחזירה ווקטור סיווג *one-versus-rest* עבורו, כך שסיווג כל תצפית הוא אחד משלושת זני האירוסים המיוצגים על ידי הספרות 0, 1 או 2.

ו. בסעיף זה תיצרי מטריצת בלבול עבור נתוני המבחן. השתמשי בפונקציה שכתבת לסיווג *one-versus-rest* עבור נתוני המבחן. השתמשי בפלט שלה ובסיווג הידוע של נתוני המבחן בכדי לייצר מטריצת בלבול בעזרת

from sklearn.metrics import confusion_matrix

confusion_matrix(y_true, y_pred)

ז. בחרי תצפית (דוגמה) אחת שאינה מסווגת נכון והסבירי מדוע לדעתך המסווג *one-versus-rest* סיווג אותה לא נכון