ביה"ס למערכות מידע

פרויקט גמר בקורס יישומי למידת מכונה בפיתון

הנחיות כלליות:

- יש להגיש את העבודה בשלשות. מספיקה הגשה אחת לכל קבוצה, אין להגיש יותר מהגשה אחת לכל שלשה. ציינו את שמות כל המגישים בגוף קובץ ההגשה.
- בתיבת ההגשה יש להגיש קובץ מסוג py ואת טבלת הנתונים.
- אנא הקפידו על טוהר הבחינה ואל תתייעצו עם קבוצות אחרות לגבי תוכן ההגשה, באם יש צורך נוכל לקיים שעת קבלה בתאום מראש.

חלק א': אלגוריתמי סיווג (classification) 70 נק'

1. מצאו נתונים המתאימים לסיווג בינארי (שתי קטגוריות דוגמת חולה/בריא, עבר/נכשל, קניה מכירה וכו'...) ב-google dataset או ב-Kaggle, צרפו את הנתונים להגשה או הכניסו את כתובת ה-URL להורדה. על הנתונים להכיל לפחות 1000 שורות (תצפיות – data points) ולפחות 4 משתנים רציפים.
2. תחילה ודאו שאין ערכים חסרים ואם ישנם השלימו אותם וכתבו בהערה בגוף הקוד משפט המסביר את אופן ההשלמה.
3. הפכו את כל המשתנים הקטגוריאליים למשתני דמה (dummy variables) כפי שהודגם בכיתה בדוגמת ה-telecom churn.
4. חלקו את טבלת הנתונים לאימון (train) ומבחן (test) ביחס של 80%-20%, בהתאמה.
5. עתה כתבו לולאה שמקבלת שמות של ארבעה המודלים לסיווג שלמדנו בכיתה:
   I. Logistic regression
   II. Support Vector Machine (SVM/SVC)
   III. K- nearest neighbor (KNN) בחרו ב: k=5 לכל הפרויקט
   IV. XGBoost

   הלולאה תעבור על המודלים כפי שהודגם בכיתה, תאמן על סט אימון ותחזה על סט המבחן. לבסוף, הלולאה תשמור בשתי רשימות שונות את מדדי הדיוק הבאים:
   a. accuracy
   b. sensitivity (recall of category 1) או כל מדד אחר הרלוונטי לנתונים שלכם/ן. למשל: specificity, F1-score, precision

   כדי לשלוף את ה-recall או כל מדד אחר באמצעות הפונקציה classification_report, אני מצרף למטלה זו גרסה מעודכנת של הקוד המדגים לולאה על מודלים בו מודגם כיצד ניתן להפוך את פלט הפונקציה למילון פיתוני ולגשת לערכיו.

6. הסבירו בגוף הקוד ועד שני משפטים את תוצאות הדיוק ורגישות המודל (או כל מדד אחר שבחרתם/ן להציג). האם שני המדדים מסכימים? בהתבסס על עולם התוכן של הנתונים הספציפיים שלכם/ן, מהו המדד שנכון להסתמך עליו?

7. עתה צרו פונקציה בשם best_accuracy_classification אשר הקלט שלה הוא מחרוזת עם שם טבלת הנתונים (csv) עליה יבוצעו המודלים והפלט שלה הוא

רשימה / טאפל שבאינדקס ה-0 רשימה עם שמות המודלים ובאינדקס ה-1 דיוקי (accuracy) של כל מודל. בנוסף, הפונקציה תדפיס את המחרוזת הבאה:

"The best model is XXX and its accuracy is YYY"

כאשר XXX הוא שם המודל המוצלח ביותר מבין ה-4 ו-YYY היא תוצאת הדיוק של אותו מודל.

הנחיות לביצוע סעיף זה:

- בכדי להגדיר מודלים בתוך הפונקציה יש להגדירם כמשתנים לוקאליים לפונקציה ולא כמשתנים גלובאליים כפי שנעשה בסעיף 5.
- בכדי למצוא את המיקום שם המודל הטוב ביותר ניתן להיעזר במקור הבא: stackoverflow.com /which-min-equivalent-in-python

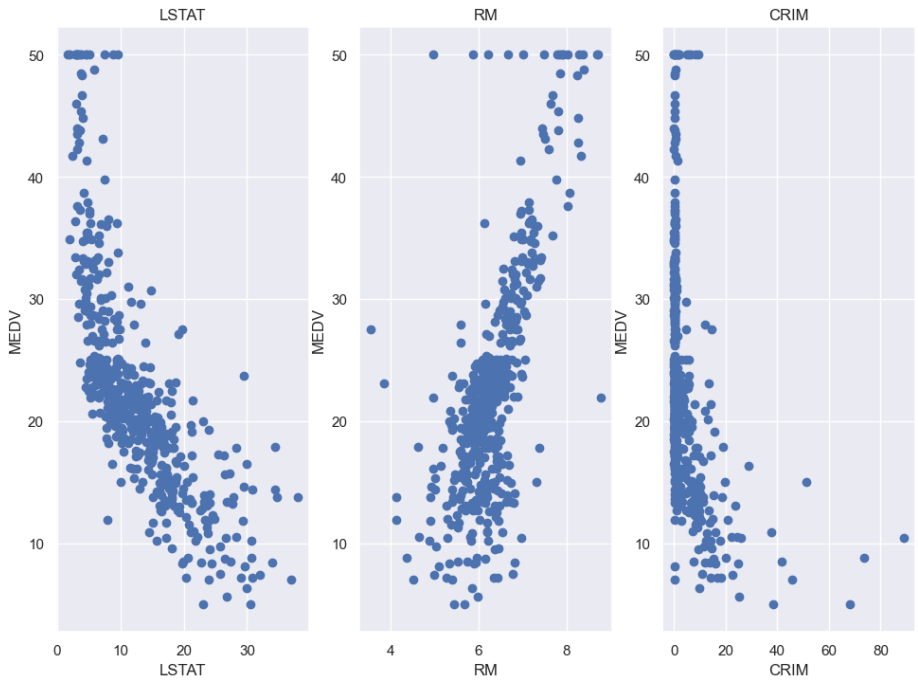
8. קראו לפונקציה מסעיף 7 עם הנתונים שלכם/ן ושמרו למשתנה את פלט הפונקציה. הדפיסו את המשתנה (רשימה או טאפל עם איבר שמות המודלים באינדקס ה-0 ואיבר דיוקי המודלים באינדקס 1.

חלק א': חיזוי משתנה רציף (regression analysis) 30 נק'

9. סננו 4 או יותר משתנים רציפים מתוך טבלת הנתונים איתה עבדתם/ן בחלק א' כך שלא יהיו משתנים קטגוריאליים בטבלה המסוננת (גם לא כאלה שהמרתם למשתנה דמה). בשלב זה ניתן לבחור גם טבלת נתונים אחרת מאותם מקורות ולצרפה להגשה לבחירתכם/ן.
10. בחרו משתנה רציף לחיזוי Y והגידור את שאר המשתנים כמשתנים מסבירים (X).
11. על הטבלה מהסעיף הקודם חלקו את טבלת הנתונים לאימון (train) ומבחן (test) ביחס של 80%-20%, בהתאמה.
12. לבחירתכם/ן הדפיסו גרף של מתאם משתנה Y לכל משתנה ב-X כפי שהודגם בקוד:

Boston_housing_linear_regression.py

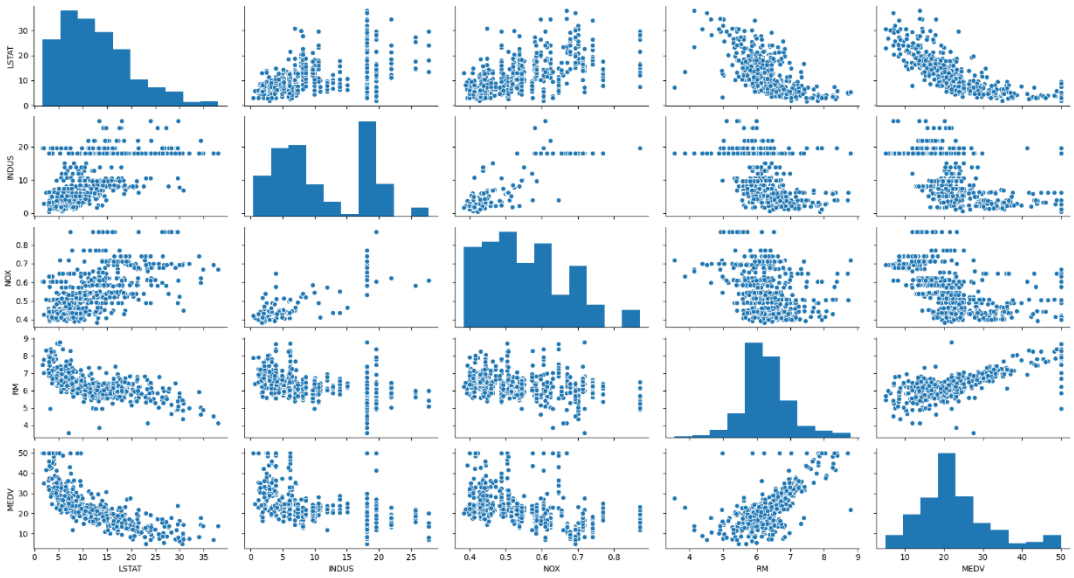
להלן הגרף:



או גרף של יחס משתנה למשתנה כפי שהודגם בקוד:

Non_linear_Bosoton_housing.py

להלן הגרף:



13. ענו במשפט בגוף הקוד לאור הגרף שהתקבל, האם היחס בין Y למשתנים X הוא יחס ישר (ליניארי)? אם לא, איזה משתנה איננו ביחס ישר?
14. עתה בצעו גם רגרסיה פולינומיאלית ממעלה 2 על הנתונים.
15. הציגו את מדדי הערכת המודלים הבאים לשני המודלים:
    a. R^2

b. Root Mean Square Error (RMSE)
c. Mean Absolut Error (MAE)
d. Mean Absolut Percent Error (MAPE)

16. בגוף הקוד בשלושה משפטים, קבעו מיהו המודל הטוב מבין השניים והסבירו בקצרה את תוצאות כל אחד מ- 4 המדדים לשגיאה (מהן היחידות, מה היחוד של כל מדד ומה המשמעות למקרה הספציפי).

בהצלחה!!!