"דימות נתונים" – עבודת גמר, סמסטר ב תשפ״א

מחברת 1 - שיפור מחברת מסמטר א'

על מנת לשפר את המודל מסמסטר א' שיפרתי מספר דברים, הוספתי OneHotEncoder, scale
ואז השתמשתי ב-PCA להורדת המימדים ל-2d.
הדיוק בסמסטר שעבר הראה 0.7 ועכשיו השתפר ל-0.74 (וזה בנוסף להורדת המימדים ל-2d).
בדקתי גם עם VotingClassifier hard,
(RandomForest, XGBClassifier, KNN, DecisionTree, AdaBoostClassifier)
וקיבלתי דיוק של 0.72,
וגם עם XGBClassifier וקיבלתי דיוק של 0.725.
במחברת זו יצרתי פונקציה שבודקת עבור כל מודל את הפרמטר cv האופטימלי ב-cross validation.
המודל השתפר ב-0.04.

מחברת 2 - fmnist

במחברת זו ההתייחסות הייתה עבור הקובץ של האימון ככולו לאימון (כך שחילקתי את הפיצ'רים שצריך לאמן ל-X_train ואת מה שמנסים לחזות כ-y_train) ורק על אלו אימנתי את המודל.
כך בהקשר שלה מחברת של הטסט לא מכניסים פרמטרים משם עד לרגע שבודקים דיוק ואלו למעשה ה-X_test, y_test).
השתמשתי גם כאן ב-PCA.
השתמשתי כאן גם ב-pipeline וב- gridsearch על המודלים.
המודלים שהשתמשתי בהם:

1. Bagging XGB, accuracy: 0.5425
   (פה השתמשתי ב-pipeline עם kmeans ו- BaggingClassifier על XGBClassifier
   וה-gridsearch היה על מנת לבדוק את ה-n_clusters המיטבי עבור kmeans).
2. XGBClassifier with no bagging, accuracy: 0.5325

(פה ב-pipeline במקום להשתמש ב- BaggingClassifier על XGBClassifier בדקתי ישירות על XGBClassifier.)

3. VotingClassifier hard , accuracy: 0.8511
4. VotingClassifier soft, accuracy: 0.8405

   פה השתמשתי גם ב-hard וגם ב-soft כי סיקרן אותי לראות כמה הבדל יהיה בדיוק ביניהם.

5. Just XGBClassifier, accuracy: 0.8553

   לבסוף רציתי לבדוק על המודל XGBClassifier ישירות ללא pipeline וללא gridseach והוא למעשה הביא לי את הדיוק הרב ביותר.

מחברת 3 - dogs vs cats

מחברת זו דומה בעיקרה למחברת מספר 2 רק שכעת ישנה עבודה רבה יותר על יבוא הדאטה כי זה לא כמו במחברת מספר 2 בקובץ csv גם של 784 פיקסלים (28*28) ולכן היה צורך בעבודה מקדימה על הדאטה להעביר אותו לדאטה הפריים של 784.

1. BaggingClassifier with XGBClassifier, accuracy: 0.56048
2. VotingClassifier hard , accuracy: 0.63456
3. VotingClassifier soft, accuracy: 0.61552
4. XGBClassifier with no bagging, accuracy: 0.54768

(פה ב-pipeline במקום להשתמש ב- BaggingClassifier על XGBClassifier בדקתי ישירות על XGBClassifier.)

6. Just XGBClassifier, accuracy: 0.65772

מחברת מספר 4 - ידיים

במחברת זו עיקר העבודה היה עבודה על הכנת הדאטה.
לקחתי את כל הקבצים של הידיים עבור כל אחד מהמצבים, הורדתי את ה-7 שניות הראשונות, השתמשתי ב- get_dummies על העמודה של Hand_type, פיצלתי אותם ל-2 חלקים של הדאטה פריים יד שמאל יד ימין) כך שאוכל לאחד אותם כל שתהיה שורה שמסמלת יד ימין ושמאל ביחד, הורדתי עמודות עם ערכי NaN (לא היו הרבה כאלה ביחס לכל הדאטה ולכן החלטתי להוריד אותם).
בקשר לקובץ של ה-HadnRight איחדתי אותו עם כל קובץ של ה-Alone בהתאם ל-shape של alone.
כך לבסוף קיבלתי קובץ אחד גדול מאוחד עם כל המצבים ואחרי שעבדתי על הדאטה.
התייחסתי לקובץ הזה כאל ה-train וגם אימנתי את המודל על כל הדאטה פריים הזה ולבסוף בדקתי את ה-test על הקובץ של ה-validation. היה צורך גם להכין מכל הקבצים של ה-ה-validation דאטה פריים אחד גדול על מנת שיהיה אפשר לקחת את הטסט.
השתמשתי ב-PCA והמודלים שהשתמשתי בהם:

1. VotingClassifier hard , accuracy: 0.8314
2. VotingClassifier soft, accuracy: 0.8314
3. XGBClassifier, accuracy: 0.9095

נקודה שלא הצלחתי להבין במהלך הבדיקה - בדקתי classification_report אך הוא הראה לי דיוק לא טוב לעומת הדיוק שאני בדקתי בעזרת cross_val_score שהוא הראה לי דיוק טוב ואני הבנתי שהדיוק שאומר מה דיוק המודל הוא בעזרת cross validation אך עדיין לא הצלחתי להבין למה יש פער גדול.