

מבוא לבניה מלאכותית 236501

תרגיל בית 3

דני פריימק 307003434

יוני בן-צבי 203668900

חלק ב'

שאלה 3

סעיף 2

חשוב לשמור על עקביות זו משום שהחלוקה אקראית. אמנם מספר הדוגמאות המסווגות כחיוביות ושליליות קבוע בכל קבוצה עבור כל הרצה של הפונקציה (עם אותו `num_folds`) אבל הדוגמאות מחולקות בין הקבוצות באופן אקראי. לכן, על-מנת לשמור על אותו מסווג יש להשתמש באותה החלוקה, כלומר לקרוא לפונקציה פעם אחת בלבד.

שאלה 5

סעיף 2

איור 1: דיוק אלגוריתם `knn` על קבוצת המבחן עבור `num_folds=2` כפונקציה של k

סעיפים 3 ו-4

כפי שניתן לראות מהגרף בסעיף 2, הביצועים הטובים ביותר התקבלו עבור $k = 1$ והביצועים הגרועים ביותר התקבלו עבור $k = 7$. ערכי הדיוק הממוצע הם 0.953 ו-0.937 בהתאמה. ראשית, נציין שהבדלי הדיוק יחסית קטנים וכן החלוקה ל-`num_folds` אקראית, לכן לא ניתן להסיק מסקנות נחרצות מדי מהתוצאות. אם בכל-זאת מנתחים את התוצאות ניתן לשער שהביצועים היו מיטביים עבור $k = 1$ וגרועים ביותר עבור $k = 7$ משום שהגיוני שככל שאדם דומה יותר לאדם אחר (מבחינת נתונים גופניים), כך סביר יותר שיהיו לו בעיות רפואיות דומות לאותו אדם. לכן, במקרה שלנו, סביר שלאדם יהיו בעיות לב אם לאדם הדומה לו ביותר מבחינת נתונים גופניים יש בעיות לב, והתחשבות באנשים דומים פחות יכולה אף לפגוע בדיוק המסווג כפי שניתן לראות בתוצאותינו. בתוצאותינו ניתן כמו-כן להבחין במגמת ירידה בטיב הביצועים ככל שגדל ערכו של k , דבר שמחזק את הנימוק לעיל (הדיוקים הממוצעים עבור $k = 3, 5$ כמעט זהים וכנ"ל עבור $k = 7, 13$).

שאלה 7

סעיף 4

הניסוי שבו התקבלו התוצאות הטובות ביותר הוא הניסוי משאלה 3 סעיף 5 (`knn` עם $k = 1$).