דו"ח מסכם

תרגיל בית 3­

יונתן בתן

302279138

[Yonibettan@gmail.com](mailto:Yonibettan@gmail.com)

עמרי פרוינד

301695490

omrifro[@gamil.com](mailto:Neliahu@gamil.com)

# **חלק א'**

מבוא

# **חלק ב'**

# סעיף 1:

בקוד

# סעיף 2:

1. בקוד
2. בקוד

# סעיף 3:

1. בקוד
2. יצרנו חלוקה ל-2\_folds כפי שהתבקשנו, שמרנו אותם במדריך הראשי תחת השמות ecg\_fold\_<i>.data

כמו כן, כתבנו פונקציית עזר בשם load\_k\_fold\_data הטוענת את הדוגמאות המתוייגות כמתבקש.

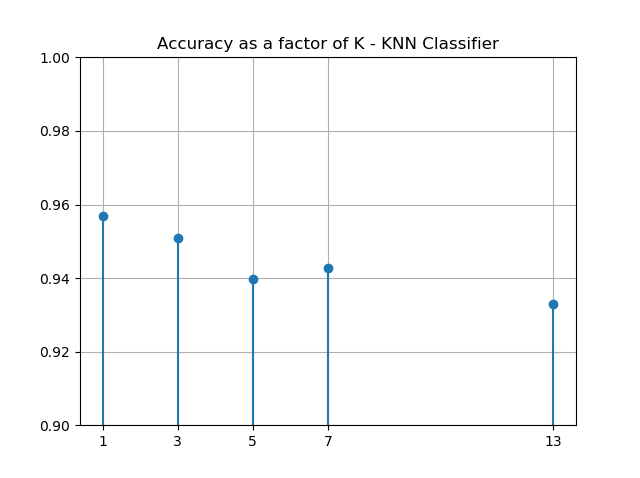
אנו מעוניינים להשוות בין מסווגים שונים. מכיוון שאין לנו תיוגים עבור סט המבחן, אנו מחלקים את דוגמאות האימון לדוגמאות עליהם נאמן בפועל את המסווג ודוגמאות עליהן נבחן אותו. חשוב מאוד לשמור על אותה החלוקה לאורך ההשוואה מהסיבה הפשוטה שבחירה שונה של דוגמאות תשפיע מאוד על תוצאות המסווג ואנו רוצים לבודד משתנים (במקרה זה המסווג) כך שלא יהיה תלוי בחלוקה כזו או אחרת.

# סעיף 4:

בקוד

# סעיף 5:

1. הקובץ experiment3.csv מצורף
2. מצורף גרף של הדיוק הממוצע של מסווג ה-Knn (Accuracy) עבור ערכי k שונים:



השתמשתי בקובץ נוסף כדי לצייר את הגרף (כדי לא לכלול את matplotlib) אפשר באקסל במקום – מה אומר?

1. ניתן לראות שהדיוק הגבוה ביותר accuracy = 0.957 מתקבל דווקא עבור שכן קרוב יחיד k=1
2. למעט עליה קלה עבור K=7 ביחס ל- k=5 ניתן לראות מגמה ברור של ירידה בביצועיי המסווג ככל שמספר השכנים גדל. הערך המינימלי מתקבל כאמור עבור מספר השכנים הגדול ביותר – accuracy=0.933 כאשר k=13.

הסבר לכך יכול להיות השימוש במדד המרחק האוקלידי לסיווג השכנים הקרובים ביותר – מדד זה נותן משקל שווה לכל אחת מהתכונות בין אם מפולגת בתחום גדול מאוד ובין אם היא תכונה בינארית המקבלת ערכי 0/1. כאשר מדובר בשכן בודד, עובדה זו פחות מורגשת ולכן התוצאות טובות יחסית אך כאשר מספר השכנים גדל אנו מתרחקים יותר ויותר מהסיווג הקרוב ומתקרבים לסיווג רוב כך שרוב הדוגמאות יתוייגו 1 במקרה שלנו. אם נגדיל את מספר השכנים עוד ועוד, בגבול נסווג תמיד 1 (זה תיוג רוב הדוגמאות) ולכן הדיוק ישאף לאחוז הדוגמאות המתוייגות 1 (כ-0.881 בסט האימון)

# סעיף 6:

השאלה נעלמה..

# סעיף 7:

1. בקוד

השתמשתי בעץ החלטה של sklearn עם מדד אנטרופיה (למרות שהוא לא באמת משתמש בID3) נדבר על זה

1. בקוד
2. מצורף קובץ experiment12.csv הכולל את מספר הניסוי, הדיוק והשגיאה עבור שני הניסויים.
3. התוצאה הטובה ביותר מתקבלת עבור מסווג KNN עם שכן יחיד. אני מניח שניתן לקבל תוצאות טובות יותר בכל אחד מהמסווגים במידה ונעבד את המידע (עיבוד מקדים) לפני השימוש במסווגים – ככל הנראה KNN מתמודד עם מידע לא אחיד בצורה הטובה ביותר מבין שלושתם.

לא באמת יודע להסביר את זה, בדרך כלל knn זה דיי חרא (לפחות בלומדות הוא נתן תוצאות על הפרצוף)

# **חלק ג'**

בינתיים משאיר את זה כאן אם ייזכרו אבל נראה כאילו הורידו את זה.. ☹

מדד המרחק האוקלידי מתייחס לכל אחת מהתכונות בווקטור התכונות כעל ממד במרחב. כלומר, ההשפעה של כל אחת מהתכונות הללו שוות ערך לגודלה ביחס לאחרות גם אם בפועל היחס הפוך. למשל, בעיית סיווג קשה תיווצר אם הנתונים מכילים שתי תכונות: "מעשן" / "לא מעשן" – {0,1} (לצורך הדוגמה תכונה מכרעת הקובעת כמעט לבדה את הסיווג), ו-"לחץ דם" של האדם {80-180} (לצורך הדוגמה בעל השפעה זניחה). במקרה זה מדד המרחק האוקלידי יסווג את האנשים כמעט אך ורק על פי לחץ הדם מה שבפועל יביא לסיווג שגוי. בעיה זו ממחישה את החשיבות העליונה של ביצוע עיבוד מקדים לכל dataset עליו עובדים.