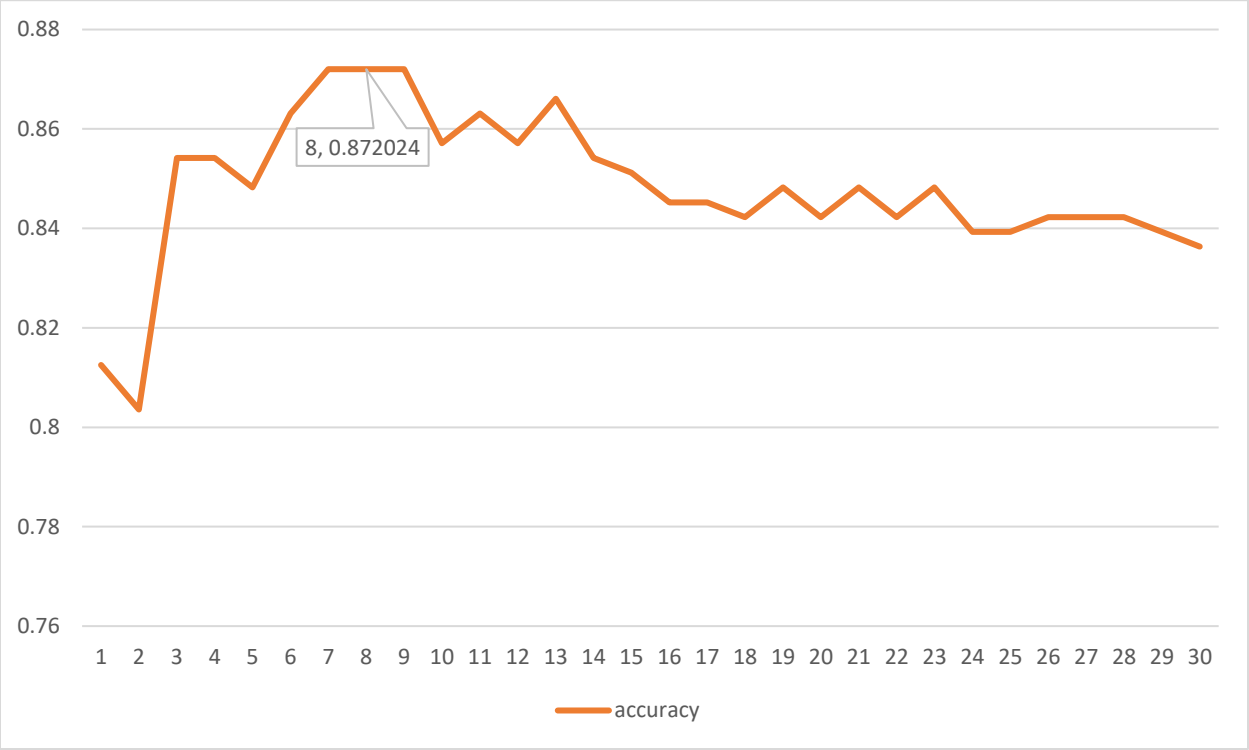
אלון יעקבי 308023910
עידו בוצר 312164569

## מעבדה 6 – KNN וולידציה

1. עבור מסווג 1NN, התקבל accuracy = 1.
אימון על כלל הנתונים הוא overfitting – כלומר, האלגוריתם מותאם בדיוק 1:1 לנתונים שעליהם נעשה test, אך רמת הדיוק לנתונים חדשים שיתקבלו תהיה נמוכה במיוחד (כיוון שיהיה רגיש לכל שינוי). מכיוון ש-K=1 כל נקודה תקבל את הערך העצמי של נקודה קרובה ביותר שגם היא בתוך הקלאסטר (למעט, *אולי*, נקודות קיצון בקלאסטר – אם נקודות הקצה של קלאסטרים שונים קרובות מאוד זו לזו, בתלות באופן יצירת הקלאסטרים), ולכן הערכת דיוק הסיווג תהיה הערכה עודפת (=1) ולמעשה לא תבוצע שום בדיקת נכונות הסיווג.

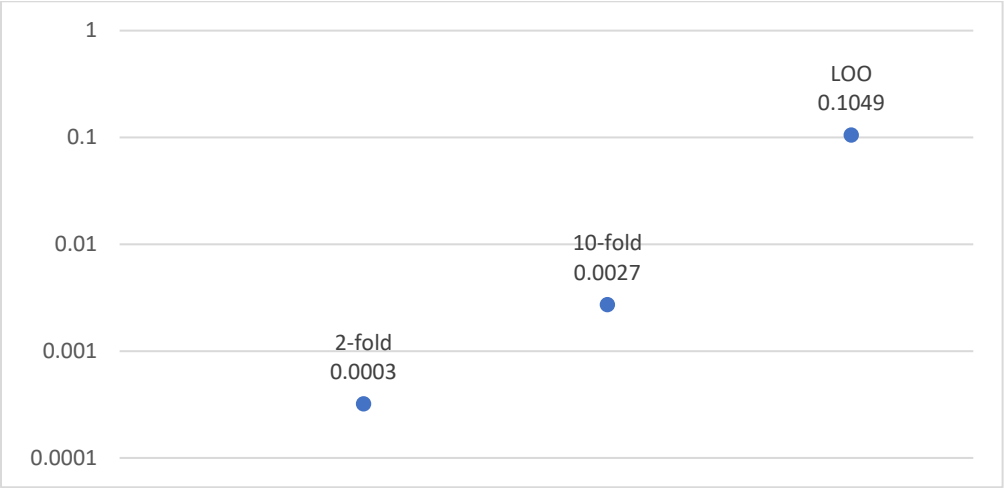
2. נרשום את ה-accuracy המתקבל מסיווג באמצעות K שכנים (ציר X), והדיוק המתקבל מ-LOO-CV על פני גרף:




ניתן לראות כי ה-accuracy הגבוה ביותר התקבל ב-K={7,8,9}, ועל כן אלו המסווגים הטובים ביותר.

אלון יעקבי 308023910
עידו בוצר 312164569

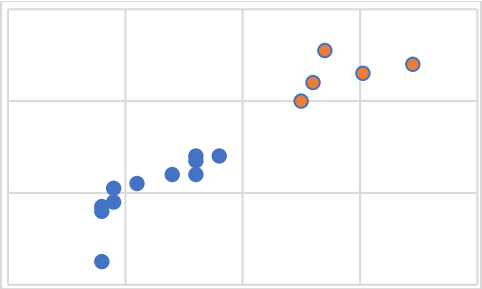
3. נראה את ה-accuracy של מס' ה-folds המשתנה על פני סקאלה לוגריתמית (למען הנוחות):



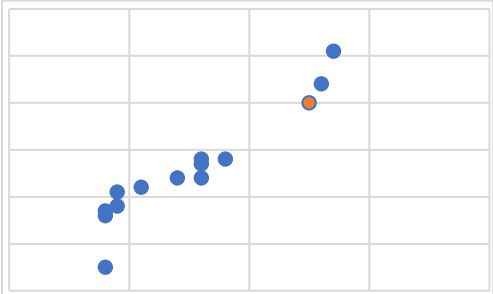

ניתן לראות ש-2-fold הוא בעל שונות נמוכה (בסדר גודל!) מהאפשרויות האחרות. ניתן להסביר זאת על ידי כך שכמות הנתונים בכל fold גבוהה משמעותית, כך שנוצרת 'החלקה' של שגיאות בסיווג בין כל fold לאחר. ככל שה-fold קטן יותר, כך ההשפעה של כל שגיאה על הממוצע גבוהה יותר.
לאור האמור, אם השגיאות 'מוסוות' באופן הנ"ל, נעדיף דווקא folds קטנים ככל הניתן על מנת לוודא את איכות המסווג. בהינתן כמות נתונים קטנה (בסט הנוכחי מדובר על 336 נקודות בלבד), סיבוכיות זמן ומקום אינם משמעותיים ולכן נעדיף אפילו את ה-LOOCV שמאפשר דיוק גבוה ככל הניתן.

אלון יעקבי 308023910
עידו בוצר 312164569

4. הצורך בערבוב רנדומלי של הנקודות לפני חלוקה ל-folds נובע מכך שאם לא נעשה כן, ייתכן ויהיו נקודות בודדות ב-fold שהן רחוקות מאוד משאר הנקודות, אך על פי ה-K המוגדר הפרדיקציה תהיה דווקא שהן שייכות לקלאסטר של אותן נקודות רחוקות.
המחשה:
אם ה-fold היה מורכב מהנק' הבאות, היה קל לראות את השיוך לקלאסטרים (לפי צבעים):



לעומת זאת, אם ה-fold היה מורכב רק מהנק' הבאות, הנקודה המסומנת בכתום הייתה מקבלת חיזוי שגוי (כאשר K>2):



כלומר, כאשר לא נעשה ערבול של הנתונים לפני CV, עלולה להתקבל הטיה של הסיווג לטובת קלאסטר מסוים.