מבוא לבינה מלאכותית – תרגיל בית 3

מגישים

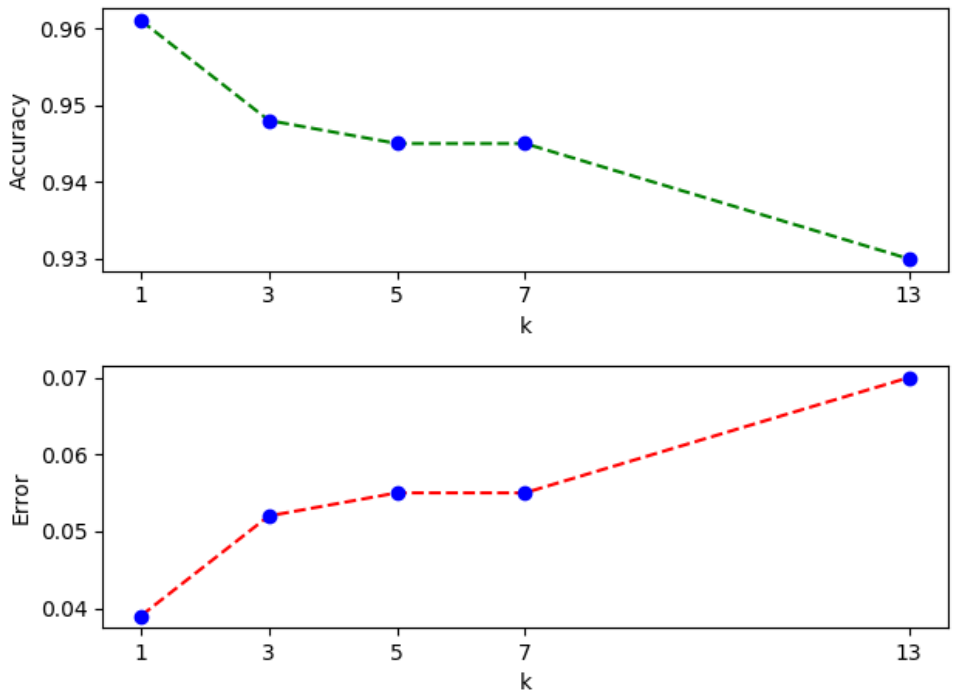
אורי קירשטיין – 311137095

פבל רסטופצ'ין - 321081016

חלק ראשון – הרצת ניסויים והשוואת מודלים.

נפצל את הdataset הנתון לשני folds ונשמור בשני קבצים. בניסויים הבאים נשתמש באותם קבצים בדיוק. חשוב לבצע את כל הניסויים על אותם קבצים כי ברצוננו להשוות בין 3 מודלים ואם נאמן כל מודל על dataset המכיל דגימות שונות, ההשוואה לא תהיה תקינה.

גרף תוצאות הניסויים על knn:



ניתן לראות שהתקבל דיוק הכי גבוה עבור k=1. המגמה היא ירידה של דיוק עם הגדלת k. הסיבה לכך היא שיותר ויותר שכנים משפיעים על הסיווג וישנן דגימות שמושפעות משכנים רחוקים ולא רלוונטיים ומקבלות סיווג שגוי. ערך המקסימום המתקבל עבור k=1 של הדיוק הוא 0.961. עבור קונפיגורציה זו כל דגימה מושפעת רק משכן אחד שהכי קרוב אליה. דיוק 1 לא מתקבל מפני שעדיין ישנן דגימות קרובות מאוד בעלות סיווג שונה.

נריץ ניסוי עם עץ החלטה ופרספטרון. הדיוק שהתקבל עבור עץ החלטה הוא 0.904 ועבור פרספטרון - 0.916. ניתן לראות בקלות שאלגוריתם ה-knn הוא בעל הביצועים הטובים ביותר עבור dataset נתון.

חלק שני – בניית מסווג חדש

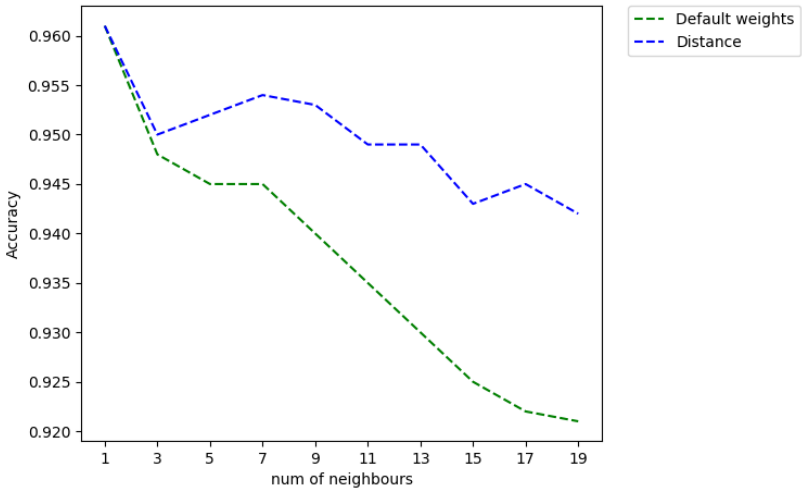
בהשראת מוטיב חוזר בTechnion Confessions, נגדיר מסווג בשם "מודל השלישייה" (͡° ͜ʖ ͡°) . המודל מורכב מ-3 מסווגים שונים:

- Knn עם מספר שכנים שווה ל-1
- Knn עם מספר שכנים שונה מ-1 עם מרחקים ממושקלים.
- עץ החלטות.

בהינתן דגימה מנורמלת, שלושת המודלים נותנות סיווג כאשר הסיווג הסופי נבחר בהחלטת הרוב. לשם בחירת היפר-פרמטרים לכל מודל פיצלנו את הdataset בצורה זהה לניסויים הקודמים. בדומה לחלק הראשון הרצנו מספר איטרציות על כל מודל בשביל לבחור את הפרמטרים המתאימים.

שמנו לב שדיוק גבוה של כל מסווג בנפרד לא בהכרח מצביע על דיוק גבוה של "מודל השלישייה". ההסבר הוא פשוט: כאשר כל המסווגים הגיעו לדיוק המקסימלי שלהם, כולם טעו באותם הדגימות ולכן הדיוק הכולל לא השתפר. ניתן לומר שהגענו למצב של התאמת יתר של המודל.

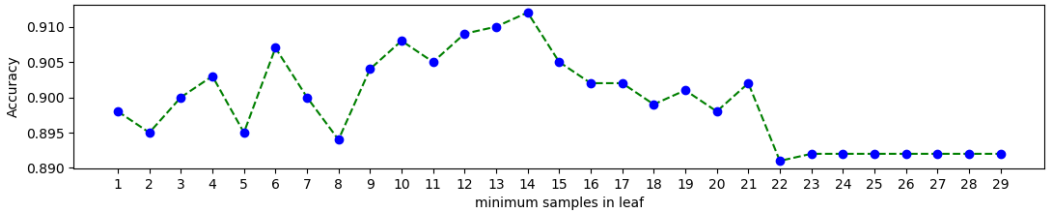
לשם כך בחרנו שני מסווגים להיות עם רגולריזציה גבוהה. במסווג ה-knn השני מספר השכנים הוא גדול, אך המרחקים ממושקלים ביחס הפוך למרחק. כלומר תרומה של השכן הכי קרוב לסיווג היא יותר גדולה מאשר תרומה של שכן יותר רחוק. נציג את ההשוואה של מסווג זה מול מסווג knn רגיל:



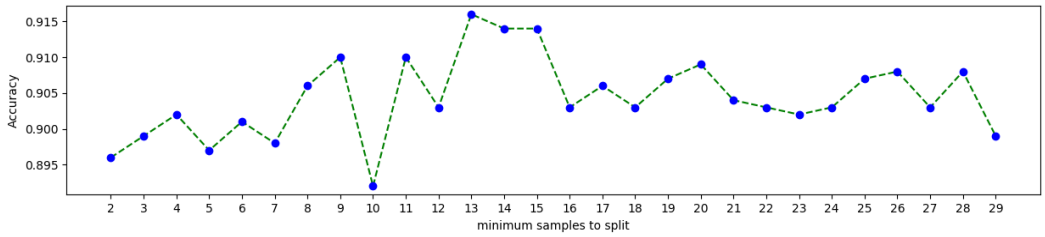
ניתן לראות שביצועי המסווג עם מרחקים ממושקלים יותר טוב מהמסווג הרגיל עבור כל מספר של שכנים. במקרה שלנו בחרנו k=7 על פי התוצאות שניתן לראות על הגרף.

המסווג השני הוא עץ החלטה. במקרה של עץ בחרנו בתור מקדם רגולריזציה להיות כמות עלים מינימלית בעלה על מנת למנוע התאמת יתר. כמו עם המסווגים האחרים בחרנו היפר-פרמטרים בעזרת הרצת ניסויים:

1) בחירת הערך של כמות מינימלית של דגימות בעלה:



2) בחירת הערך של מספר מינימלי של דגימות הניתנים לפיצול:



לאחר השוואת ביצועים של המודל הכללי בחרנו להשתמש רק בהיפר-פרמטר של כמות הדגימות המינימלית בעלה. לפרמטר השני לא הייתה השפעה מורגשת על דיוק המודל הכללי.

לאחר קביעה סופית של היפר-פרמטרים הנ"ל נעשה אימון על כל ה-dataset הנתון, ונעשה סיווג של הדגימות ללא תיוגים.

- למימוש המודל לא השתמשנו בקוד קיים אלא רק בדוגמאות מאתר של sklearn.
- בקבצי הקוד שלנו נעשה import למסווגים מתוך sklearn אך לא נעשה בהם שימוש בסעיפים של מימוש מסווג ה-knn.