מבוא לבינה מלאכותית – תרגיל בית 3

מגישים

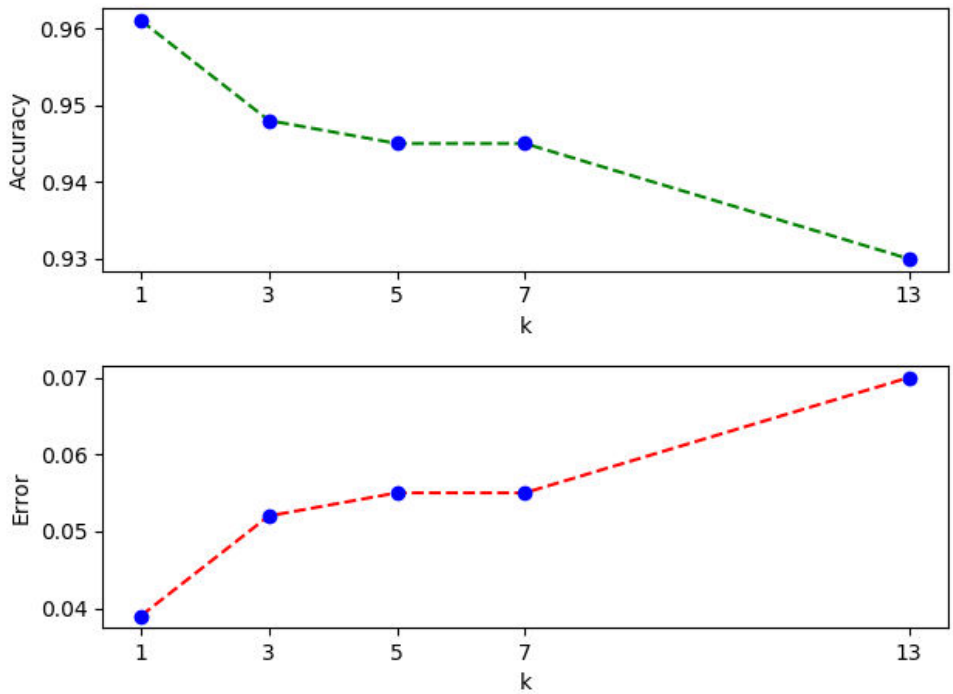
אורי קירשטיין – 311137095

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" . המודל מורכב מ-3 מסווגים שונים:

- רשת נוירונים.
- SVM – מכונת וקטורים תומכים.
- Knn.

בהינתן דגימה מנורמלת, שלושת המודלים נותנות סיווג כאשר הסיווג הסופי נבחר בהחלטת הרוב. לשם בחירת היפר-פרמטרים לכל מודל פיצלנו את הdataset ל-750 דוגמאות אימון ו-250 דוגמאות אימות ונירמלנו אותן. 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.

לאחר קריאה מעמיקה על SVM ורשת נוירונים באתר של sklearn, הגענו למסקנה שעלינו לדאוג לרכיב רגולריזציה גבוה ברשת נוירונים וב-SVM. רכיב הרגולריזציה הוא חלק מפונקציית ה-loss ותפקידו למנוע התאמת יתר של המסווג. כתוצאה מכך קיבלנו דיוק קצת פחות טוב של רשת נוירונים ושל SVM אך קיבלנו מסווגים בעלי יכולת הכללה יותר טובה. בנוסף, ניתן היה לראות שבמצב זה כל מסווג טועה בדגימה שונה משני המסווגים האחרים מה ששיפר משמעותית את הדיוק של "מודל השלישייה" - קיבלנו דיוק סיווג של קבוצת האימות 98%. לאחר קביעה סופית של היפר-פרמטרים הנ"ל נעשה אימון על כל ה-dataset הנתון.

למימוש המודל לא השתמשנו בקוד קיים אלא רק בדוגמאות מאתר של sklearn.