דו"ח מסכם

תרגיל בית 3

יונתן בתן

302279138

Yonibettan@gmail.com

עמרי פרוינד

301695490

omrifro@gamil.com

חלק א'

מבוא

חלק ב'

סעיף 1:

בקוד

סעיף 2:

1. בקוד
2. בקוד

סעיף 3:

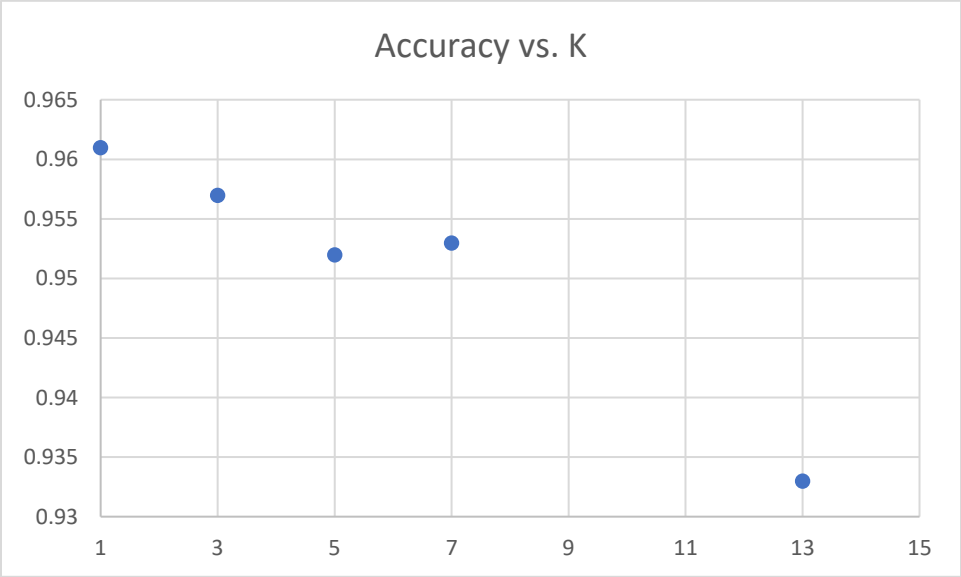
1. בקוד
2. יצרנו חלוקה ל-2_folds כפי שהתבקשנו, שמרנו אותם במדריך הראשי תחת השמות ecg_fold_<i>.data
כמו כן, כתבנו פונקציית עזר בשם load_k_fold_data הטוענת את הדוגמאות המתוייגות כמתבקש.
אנו מעוניינים להשוות בין מסווגים שונים. מכיוון שאין לנו תיוגים עבור סט המבחן, אנו מחלקים את דוגמאות האימון לדוגמאות עליהם 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:



3. ניתן לראות שהדיוק הגבוה ביותר Accuracy = ~0.96 מתקבל דווקא עבור שכן קרוב יחיד k=1

4. ניתן לראות מגמה ברור של ירידה בביצועיי המסווג ככל שמספר השכנים גדל. הערך המינימלי מתקבל כאמור עבור מספר השכנים הגדול ביותר – accuracy=0.932 כאשר k=13.
הסבר לכך יכול להיות השימוש במדד המרחק האוקלידי לסיווג השכנים הקרובים ביותר – מדד זה נותן משקל שווה לכל אחת מהתכונות. 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). הסבר נוסף הוא שייתכן וסט האימון שלנו מכיל כפילויות (למעט רעש) ולכן מתייג נכון על סמך שכן קרוב יחיד אך עבור סט המבחן זה לא יעבוד באותה צורה.

סעיף 7:

1. בקוד

2. בקוד

3. מצורף קובץ experiment12.csv הכולל את מספר הניסוי, הדיוק והשגיאה עבור שני הניסויים.

4. התוצאה הטובה ביותר מתקבלת עבור מסווג KNN עם שכן יחיד. ככל הנראה ניתן לקבל תוצאות טובות יותר בכל אחד מהמסווגים במידה ונעבד את המידע (עיבוד מקדים) לפני השימוש במסווגים – השערתנו היא שמסווג ה-KNN מתמודד עם סוג זה של המידע ללא עיבוד מקדים בצורה הטובה ביותר מבין שלושתם.

חלק ג'

<u>רקע</u>

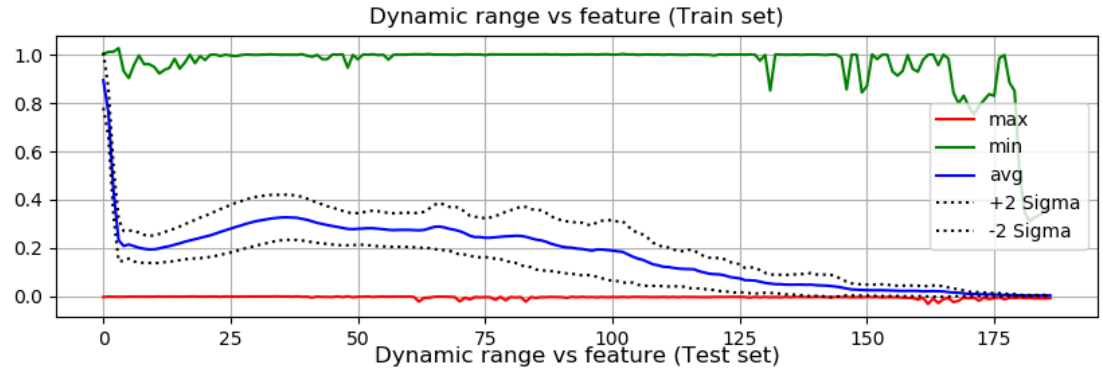
חילקנו את בעיית הסיווג ל-4 חלקים עפ"י הטבלה הבאה:

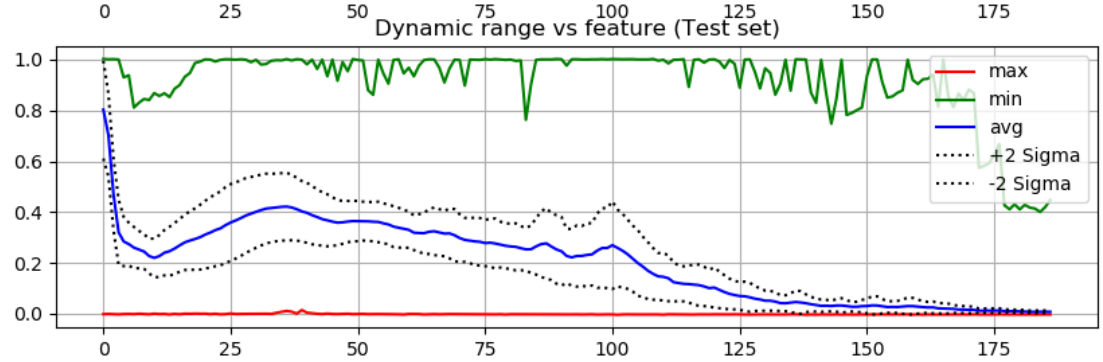
Pre-processing	Classifier type	Classifier's parameters **	Performance boosting
None	KNN	# neighbors	None
Feature selection	Linear (perceptron)	Distance method	Random forest
	Decision Tree	Inequality method	Majority vote
	Naïve Bayes	Kernel	
	SVM	Polynomial degree	
		Distribution	
		Weights	

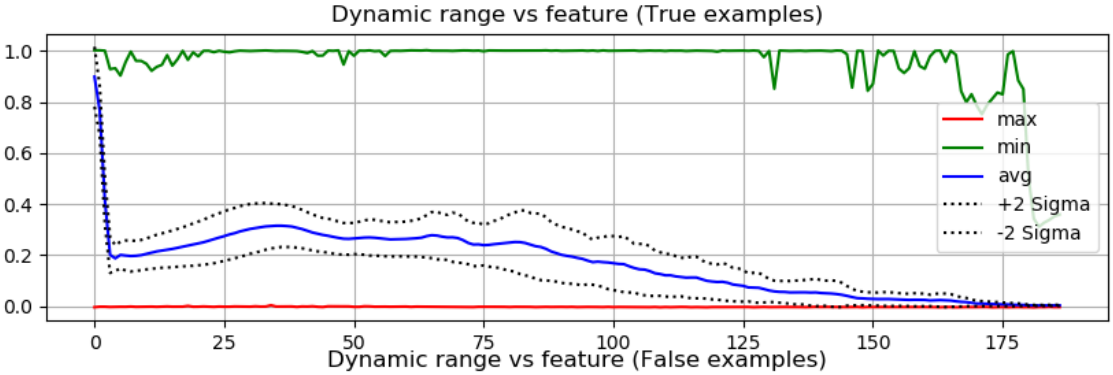
** הפרמטרים בטבלה הם רק קומץ דוגמאות מבין אלו שנבחנו והם אינדיבידואליים לכל מסווג.

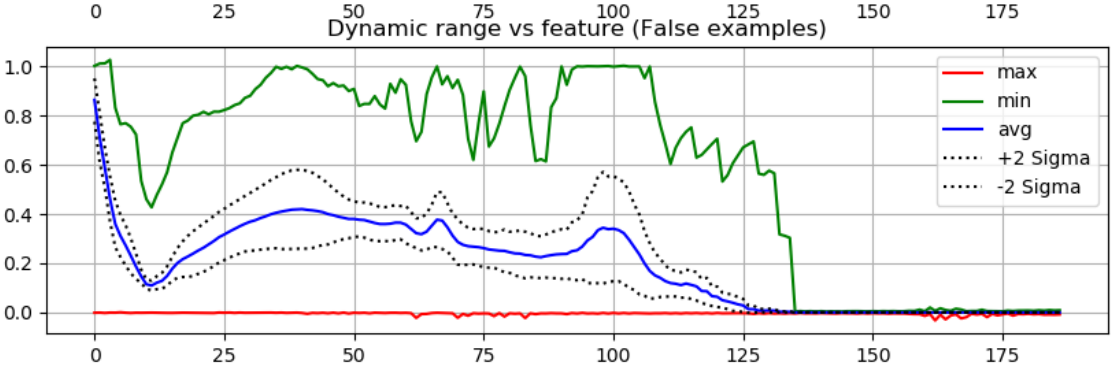
שלב א' – ניתוח המידע (data_analisys.py)

לצורך סיווג הדוגמאות לתחרות התחלנו בלמידת ה-data שברשותנו. קראנו עליו ככל שניתן, הצגנו גרפית מספר דוגמאות חיוביות ומספר שליליות, בדקנו רעש, תחום דינאמי, תוחלת ושונות. הבנו שהדוגמאות כולן בתחום דינאמי זהה 1-0 למעט רעש מפולג נורמלית עם תוחלת 0 ולכן אין צורך בנרמול. עוד ניתן היה להבחין שמרבית הדוגמאות (חיוביות ושליליות כאחד) כללו פיק במספר התכונות הראשונות וירידה לערך 0 לקראת האחרונות. מצורף גרף של 2 דוגמאות חיוביות ו-2 שליליות לצורך המחשה.





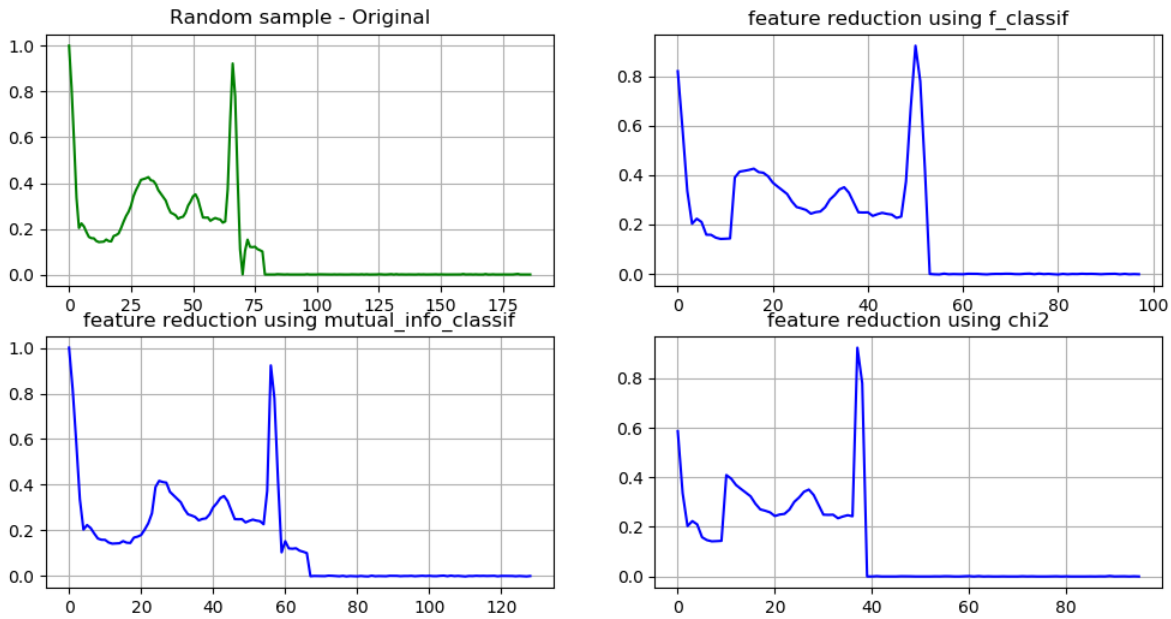




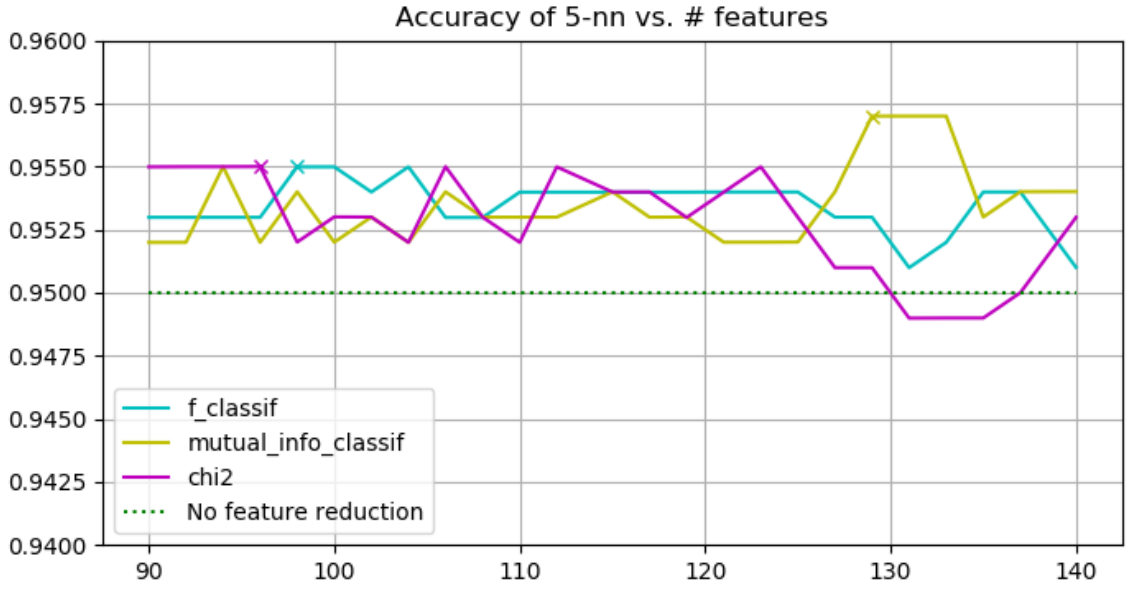
שלב ב' – עיבוד מקדים (pre_processing.py)

בדקנו יתרונות וחסרונות של מספר שיטות להפחתת ממד (מספר הפיטצ'רים) של כל דוגמא ע"י select_K_best.

תחילה ניתן לראות שצורתן הכללית של הדוגמאות נשארת כפי שהייתה (מצורף גרף עבור דוגמא אקראית)



לאחר מכן ביצענו השוואה של ביצועי מסווג 5-NN (נבחר אקראית) ללא עיבוד מקדים ועם עיבוד עבור ערכי K שונים



שלב ג' – השוואת מסווגים (clf_comparison.py)

לצורך אימון ובחינת כל המסווגים השתמשנו בספרייה sklearn וביצענו השוואה בין המסווגים השונים באמצעות Cross-validation בעל 3-folds (תחת ההנחה שישנן 300 דוגמאות מבחן ו1000 אימון). ההשוואה בוצעה גם ללא עיבוד / עם, כאשר גם כמות התכונות וגם פונקציית הבחירה שלהן עברו השוואה. לבסוף, בדקנו האם השוואה בין מסווגים (ועדה) או בחירה רנדומלית של עצי החלטה ובחירה בין סיווג הרוב (Random Forest) יביאו לשיפור הסיווג.

את התוצאות שכללו את המסווג, הפרמטרים השונים, האם עבר עיבוד מקדים (ולכמה תכונות הופחת), הדיוק הממוצע והשונות בין הפולדים השונים שמרנו לקובץ CSV יחיד שבהמשך עבר מיון עפ"י הדיוק על מנת לקבל תמונה ברורה ככל הניתן של האפשרות הטובה ביותר (מצורף).

שלב ד' - סיכום (competition.py)

נבחר המסווג הבא:

1. **עיבוד מקדים**
   a. כדי להקטין את הרעש ולשמור על ערכים בתחום קבוע ותקין 1-0 ביצענו Trim (כל מה שגדול מ-1 קבענו 1 וכל מה שקטן מ-0 קבענו להיות 0.
   b. בחרנו 70 תכונות מתוך 187 לכל דוגמא ע"י שימוש בפונקציה select K best מהספריה Sklearn. בחירת התכונות נעשתה באמצעות f_classif לאחר שזו הראתה את התוצאות הטובות ביותר.
2. **מסווגים**
   a. המסווג בעל התוצאות הטובות ביותר היה KNN כאשר הדיוק היה דומה עבור 1,3,5 שכנים קרובים. מדד המרחק (אוקלידי/מנהטן) נבחר כתלות במספר השכנים עפ"י התוצאות הטובות שהושגו.
   b. מסווג שני שנתן תוצאות טובות מאוד היה SVM עם גרעין פולינומיאלי ממעלה 11 ואיבר חופשי 2.
   c. המסווג השלישי שתוצאותיו טובות מעט פחות מהקודמים היה למעשה אוסף מסווגים – Random Forest בעל מדד אי-שיוויון אנטרופיה ו-200 עצים רנדומליים.
3. **הכרעה**
   a. ההחלטה הסופית לגבי סיווג הדוגמא נלקח ע"י החלטה רוב מבין חמשת המסווגים הנ"ל ביחס שווה.
   b. תוצאת הדיוק הממוצעת שהתקבלה למסווג הסופי על 3 פולדים היה: Accuracy = 0.971