**מערכות לומדות – תרגיל 2 – מגיש נתאי יחזקאלי (ת.ז. 021976808)**

1.

|  |  |  |  |
| --- | --- | --- | --- |
| תשובת המסווג | |  | |
| Negative | Positive |
| 1 | 0 | Positive =1 | סיווג האמת |
| 999 | 0 | Negative =999 |

2. האלגוריתם עשה שגיאות.

3. וזהו דיוק המסווג.

4.

|  |  |  |  |
| --- | --- | --- | --- |
| תשובת המסווג | |  | |
| Negative | Positive |
|  | 0 | Positive | סיווג האמת |
|  | 0 | Negative |

בתשובת המסווג שהאיש חולה תמיד יש 0 במונה ולכן התשובה היא 0.

5.

|  |  |  |  |
| --- | --- | --- | --- |
| תשובת המסווג | |  | |
| Negative | Positive |
| 0 | 1 | Positive | סיווג האמת |
| 0 | 999 | Negative |

6. האלגוריתם עשה שגיאות.

7. וזהו דיוק המסווג.

8.

|  |  |  |  |
| --- | --- | --- | --- |
| Negative | Positive |  | |
|  |  | Positive | סיווג האמת |
|  |  | Negative |

בתשובת המסווג שהיא בריא התשובה היא תמיד 0 כי במונה יש תמיד 0.

9.

|  |  |  |  |
| --- | --- | --- | --- |
| תשובת המסווג | |  | |
| Negative | Positive |
|  |  | Positive =2 | סיווג האמת |
|  |  | Negative =1998 |

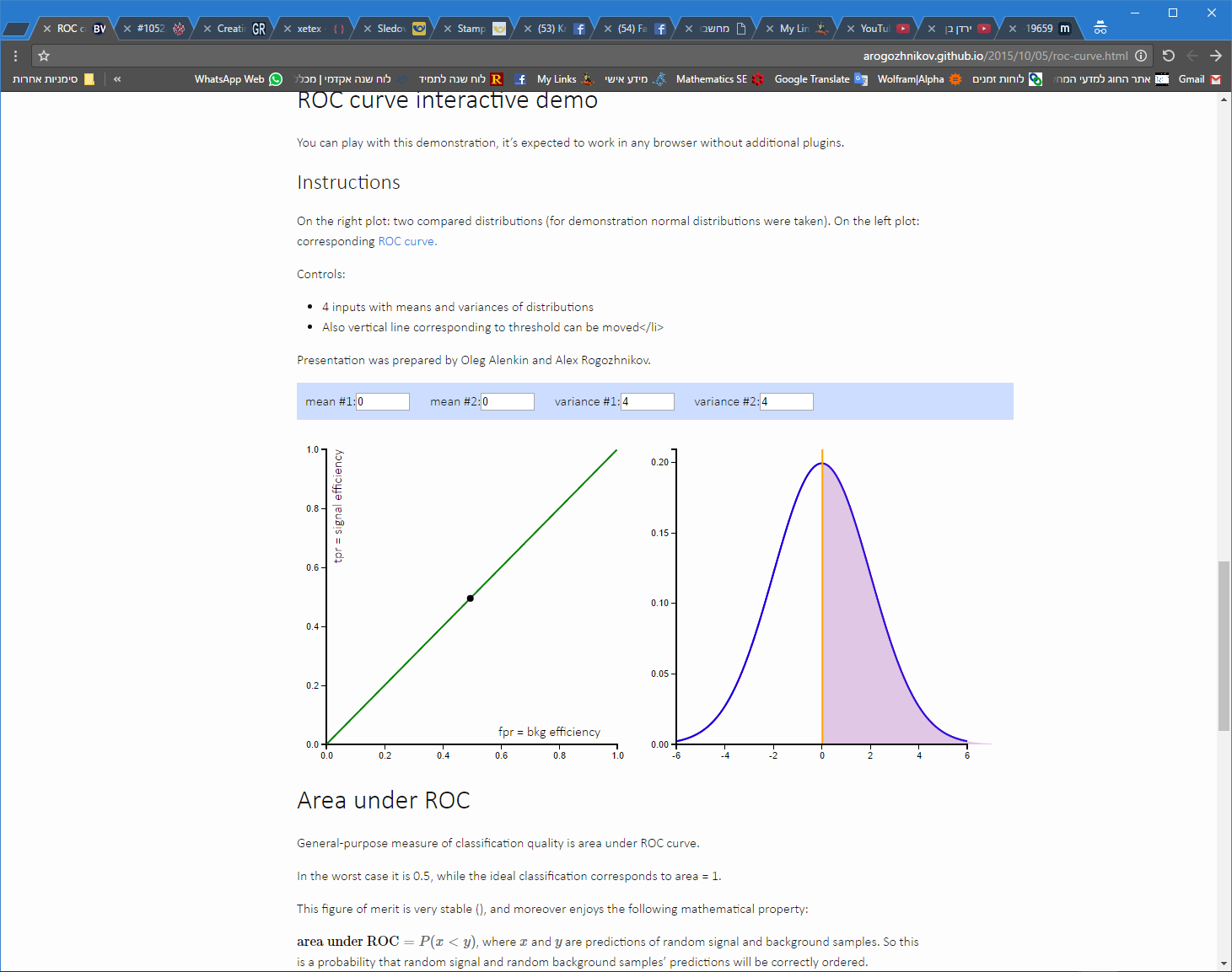
10. האלגוריתם עשה: שגיאות.

11. וזהו דיוק המסווג.

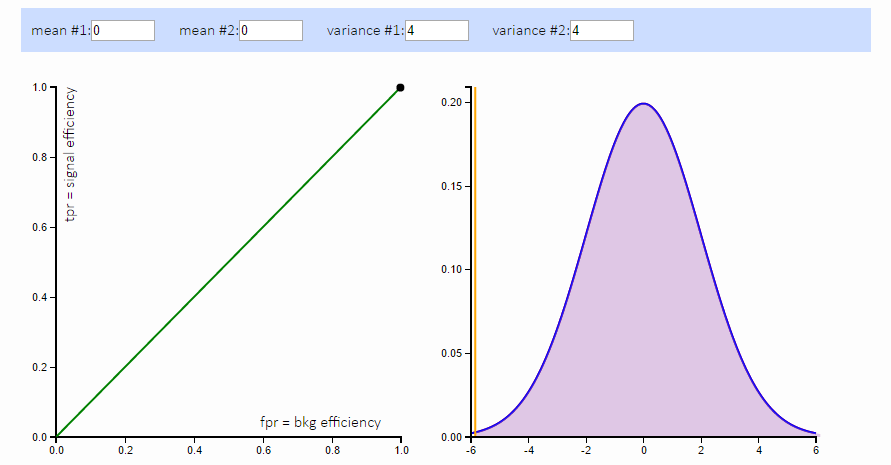
12.

|  |  |  |  |
| --- | --- | --- | --- |
| Negative | Positive |  | |
|  |  | Positive | סיווג האמת |
|  |  | Negative |

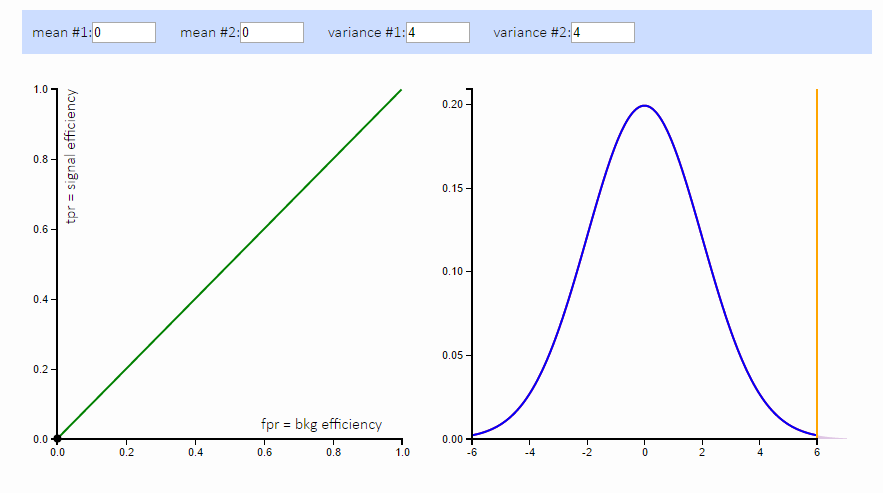
13.



כאשר הנתונים חופפים בצורה מושלמת, אזי ישנו יחס ישר בין התיוגים הנכונים לבין התיוגים הלא נכונים (של p) כתלות בכלל ההחלטה. כלומר, אם נחליט לתייג כל הנתונים ככחולים:

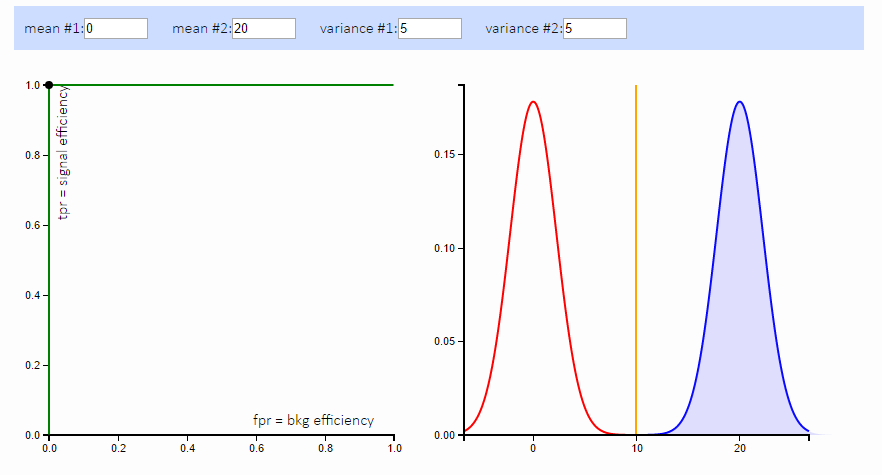


אזי כמות הכחולים שתויגו ככחולים היא 1.0 (100%) (tp – פירושו כחול שהמסווג תייג ככחול, fp – פירושו אדום שהמסווג תייג ככחול), וגם הערך של fp הוא 1.0 (100%) מכיוון שכלל ההחלטה הוא שהמסווג מסווג את כולם ככחולים – ולכן הוא צדק ב-100% לגבי הכחולים (כי באמת הוא הצליח לתייג את כל הכחולים ככחולים) וב-100% לגבי האדומים שהוא תייג ככחולים (כי 100% מהאדומים תויגו ככחולים).  
וכנ"ל לגבי כל נקודה שנבחר תמיד זה יהיה אותו יחס – אחוז הכחולים שיתויגו ככחולים יהיה זהה לאחוז האדומים שיתויגו ככחולים:

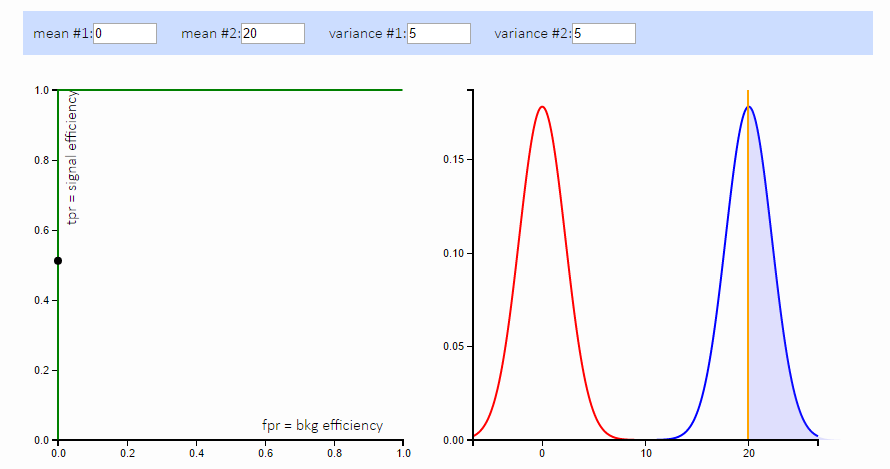


0% מהכחולים תויגו ככחולים ו-0% מהאדומים תויגו ככחולים (כי כל הנתונים תויגו ככחולים). ובגלל שהנתונים חופפים לגמרי, אזי זה תמיד יהיה אותו יחס.

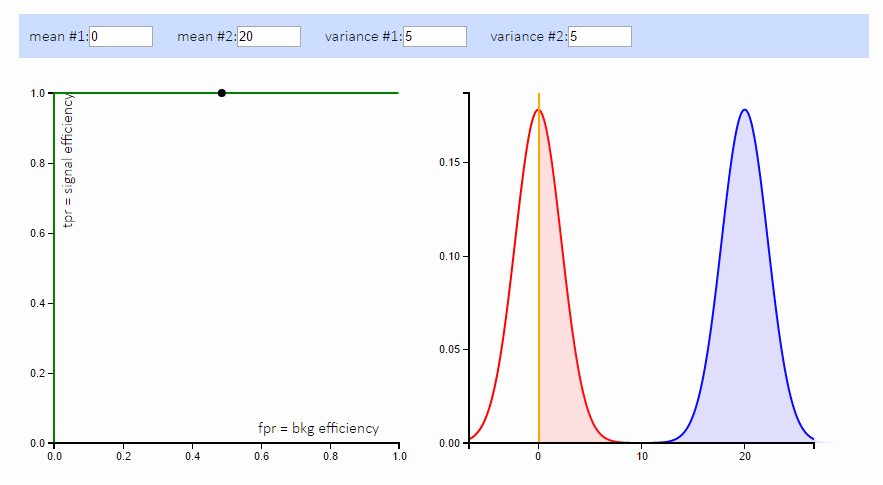
14. במקרה הזה נוכל לקבל שליטה מלאה על אלה שיתויגו נכון ועל אלה שיתויגו לא נכון, כל אחד בנפרד, כלומר, נוכל שערך אחד יהיה 1.0 ובשני לשלוט או ההפך:

למשל, באיור שלמטה ניתן לראות מה קורה כאשר החלטנו לשים את כלל ההחלטה באמצע (היכן שהוא נמצא) – כל הכחולים תויגו ככחולים (ולכן ערך ה-TP הוא 1.0) וכל האדומים תויגו כאדומים (ולכן ה-FP הוא 0.0 – כי אין אף אדום שתויג ככחול).

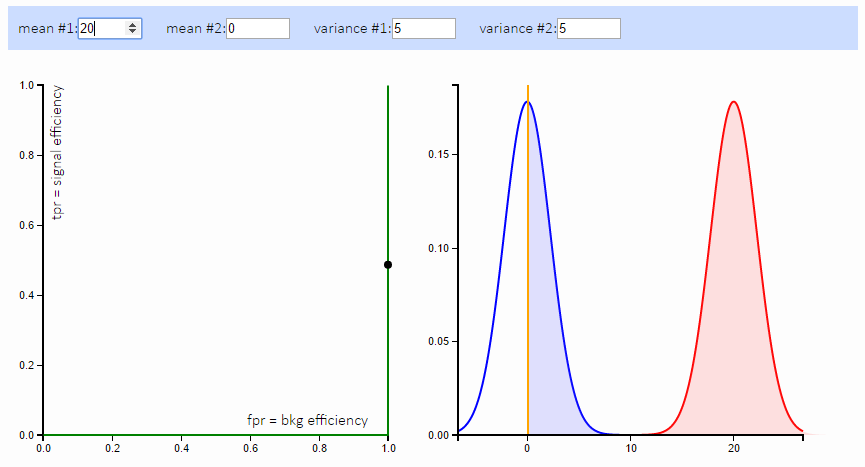
במקרה הבא למשל:



ניתן לראות שערך ה-FP הוא 0.0 – הסיבה לכך היא שמספר האדומים שתויגו כחולים הוא 0, לעומת זאת, בערך רק חצי מהכחולים תויגו ככחולים ולכך ערך ה-TP הוא חצי – כי רק חצי תויגו נכון.

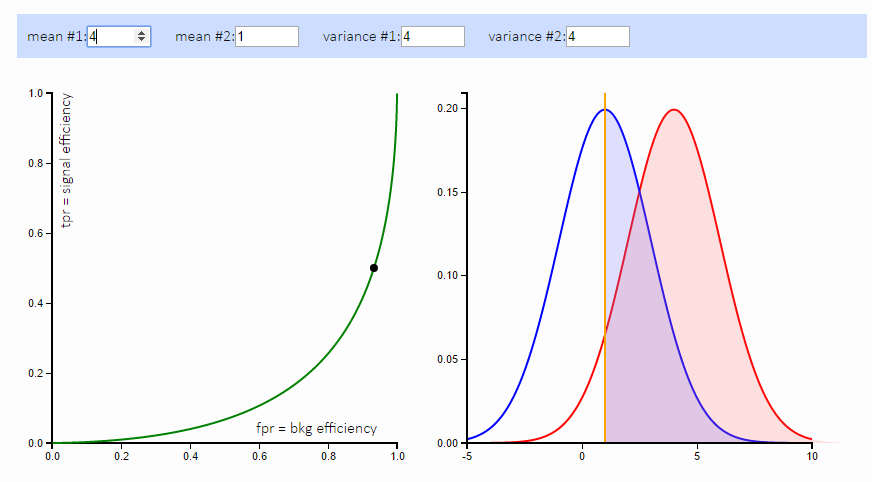
במקרה שלמטה, כל הכחולים תויגו נכון (ולכן ה-TP הוא 1) ולעומת זאת רק בערך חצי מהאדומים תויגו ככחולים, ולכן ערך ה-FP הוא בערך חצי (חצי אדומים תויגו ככחולים).  


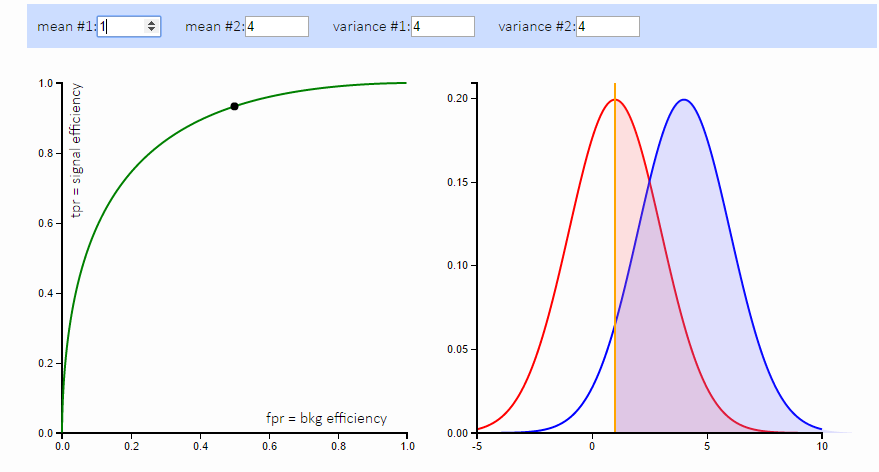
באותו מידה היינו יכולים לשנות את מיקום האדומים והכחולים וכך לקבל תוצאה הפוכה.  
בכל מקרה, כמו שנאמר בהתחלה – אם הנתונים מופרדים זה מזה – אחד היחסים תמיד אחד ובשני נוכל לשלוט ולקבל את היחס שאנחנו רוצים.  
לכן הגרף כאן הוא בצורה של "ר" (הפוכה) ה-TP תמיד יהיה 1 ואילו ה-FP משתנה בהתאם לקו הצהוב (כלל ההחלטה), ואם היינו הופכים את הצבעים אזי, זה היה נראה כך:



ערך ה-FP תמיד היה 1 והיינו יכול לשלוט ב-TP.

15. המשמעות היא שכמות התיוגים הלא נכונים גבוה מכמות התיוגים הנכונים, למשל:



ניתן לראות באיור שלמעלה שכמות האדומים שתויגו ככחולים גבוהה מכמות הכחולים שתויגו ככחולים.  
נוכל לשנות זאת ע"י כך שפשוט נוריד נחליף בין שני הצבעים (שני הסטים):

16. ה-recall של הספרה 8 הוא המשמעות היא שלא היו שגיאות עם הספרה הזאת והיא תמיד תויגה כ-8. הערך של העמודה הנ"ל היא הערך של התוצאה של הנוסחה[[1]](#footnote-1): והמשמעות של 1.0 היא ש-, כלומר סך השגיאות (כלומר, ספרות ה-8 שתויגו כספרות אחרות) הוא אפס. רואים את זה ב-confusion matrix ע"י כך שבשורה של הספרה 8 רק בעמודה של הספרה רואים מספר שהוא גדול מ-0 וכל שאר הערכים באותה שורה הם אפסים – כלומר, כל הספרות שהם 8 תויגו כ-8.

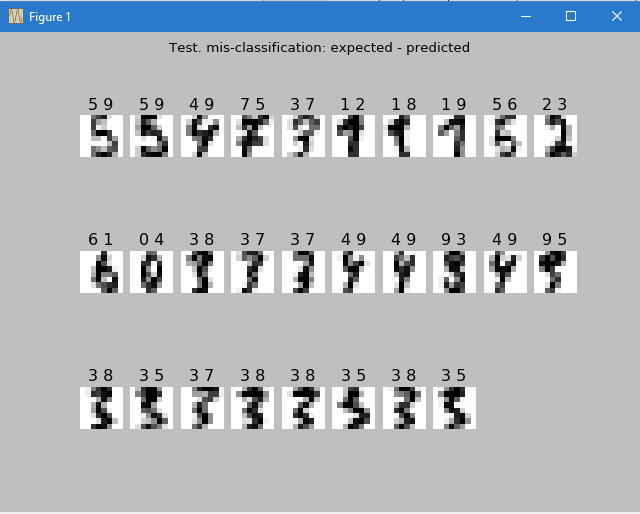
17. ה-precision של הספרה 0 הוא – המשמעות של זה היא שהדיוק של הספרה 0 היה מרבי, כלומר, מספר הספרות שהיו לא 0 ותויגו כ-0 הוא אפס, כלומר, שום ספרה אחרת שהיא לא 0 לא תויגה כ-0. הנוסחה[[2]](#footnote-2) היא:

והערך הוא 1 כאשר , כלומר המסווג לא סיווג שום ספרה שהיא אינה 0 כ-0.

18. רק שלוש שהן 5 סווגו בטעות כספרה אחרת –ספרה אחת תויגה כ-6 ושתי ספרות כ-9. אנחנו יכולים לראות את זה בטבלה השנייה ש-88 ספרות של 5 סווגו כ-5, רק ספרה אחת סווגה כ-6 (ולכן היא נמצאת עמודה אחת ימינה ממנה, כלומר, הספרה היא באמת 5, אבל המסווג סיווג אותה כ-6) ועוד שתי ספרות סווגו כ-9 (רואים את זה בעמודה הכי ימנית).

19. הספרה שסווגה הכי הרבה פעמים בטעות זאת הספרה 3 – ישנן 91 ספרות של 3 במאגר הנתונים, אבל המסווג סיווג 79 ספרות של 3 כ-3, שלוש ספרות של 3 כ-5, ארבע ספרות של 3 כ-7, ו-חמש ספרות של 3 כ-8. ניתן לראות זאת בשורה של 3.

20. הנה התוצאה והיא נמצאת גם בקובץ התרגיל.

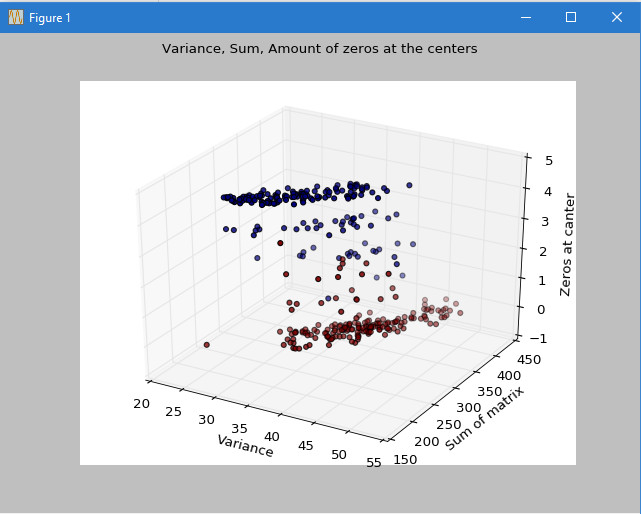


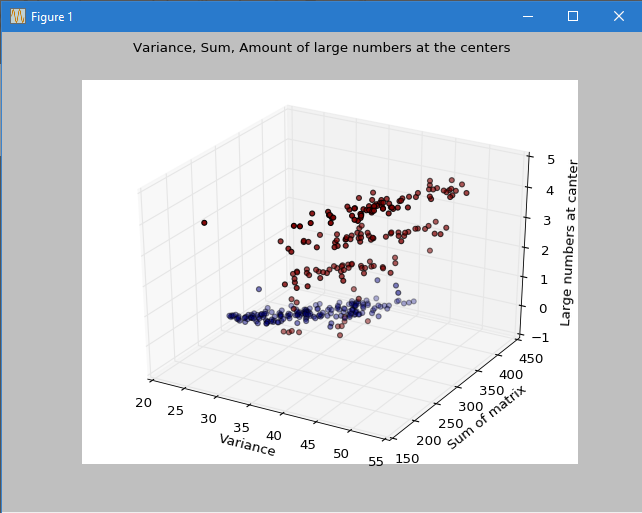
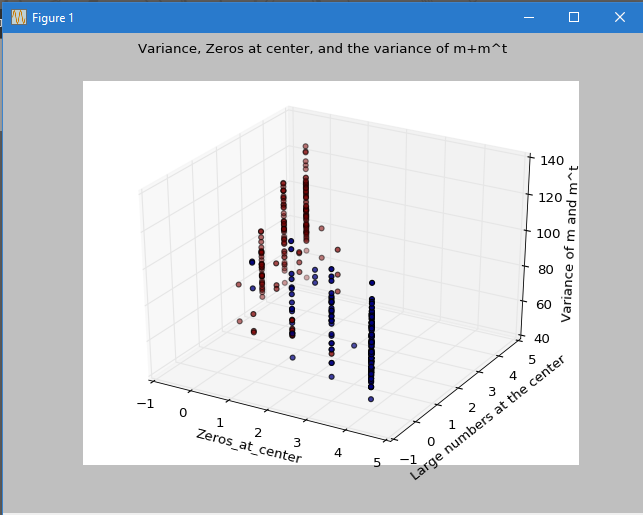
21.

כמה הערות לשאלה:

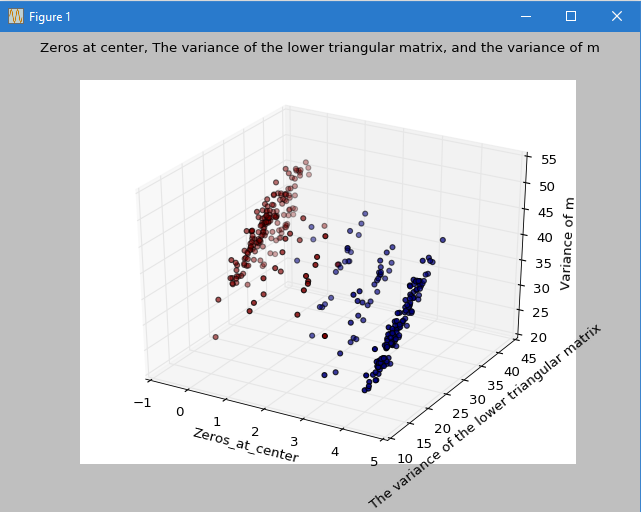
1. היות והיו המון שינויים לא כתבתי קוד יעיל, כלומר, ניתן לשפר אותו, אבל לא על זה שמתי את הדגש.
2. אמרת לי שניתן לכתוב כמה מאפיינים שרוצים, אבל צריך להשתמש מקסימום ב-10, לכן יש סה"כ 13 מאפיינים אבל בסעיפים ו'+ז' השתמשתי רק ב-10 (אבל לא באותם עשרה מאפיינים).
3. לגבי הגרפים – לא הצלחתי ליצור הפרדה טובה כמו שיש בגרף ששמת, אבל כן מצאתי כמה פרמטרים שנותנים הפרדה ברורה (אם כי אולי יש פה ושם נקודה שנמצאת "בצד השני").
4. בתוכנית עצמה שמתי כל מיני פלטים שיופיעו, לא הורדתי אותם כי למיטב הבנתי אין בעיה אם זה, וזה עוזר להבין יותר טוב כל מיני דברים בהקשר של זיהוי הספרות.
5. בחלק של זיהוי הספרות (כולן) הוא כותב איזו אזהרה על מספר גדול מדי. ניתן להתעלם ממנה כי למיטב הבנתי (קראתי את האזהרה) מדובר על כך שזה עלול להגיע למספרים גדולים. אבל בגלל שאצלנו זה משהו קובע – אפשר להתעלם.
6. לגבי שמות הפונקציות והמשתנים – ניסיתי לתת להם שמות כמה שיותר משמעותיים, אבל מצד שני שמות שלא יהיו ארוכים מדי, לכן לא תמיד הצלחתי לקלוע במדויק. בד"כ התיעוד של הפונקציה מסביר הכי טוב, אבל גם שמה לפעמים הסתבכתי עם ההסברים.
7. כתוב בדף התרגיל שצריך לשים את הפונקציות למעלה, לכן גם את הפונקציות של המאפיינים (אלה שיוצרות מאפיינים) שמתי למעלה.

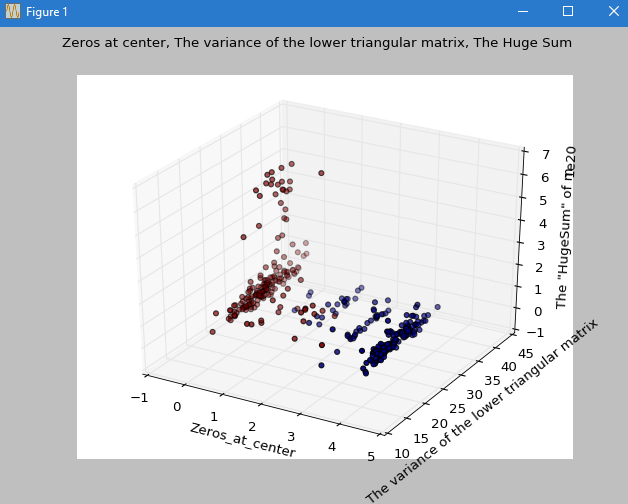
ה.

לא ממש הצלחתי ליצור גרפים עם הפרדה ברורה, אבל מכל מה שהצלחתי למצוא אלה הכי טובים:



באיור למעלה ניתן לראות שהעמודות הן לפי כחול/אדום ואולי פה ושם ישנה נקודה שלא נמצאת בצד "שלה", אבל רוב הנקודות מסודרות.





*הערה: בקובץ הקוד שמתי את הקריאה לפונקציה שמציגה את הגרפים בהערה ולכן היא אינה מופעלת כאשר התוכנית רצה.*

ז. בתיאור המאפיינים שמתי את שמות המשתנים (המערכים) שבהם השתמשתי, בנוסף לא הייתי צריך להשתמש בכל האמפיינים, ברגע שראיתי שקיבלתי תוצאה טובה (הטבלה שלמטה) הפסקתי לחפש צירופים טובים יותר (ניסיתי כמה אבל זה רק הוריד...):

Logistic regression using [sumx, varx, zeros\_at\_center, sum\_1, max\_1, large, var\_of\_m\_and\_t,tri\_down, tri\_up] features cross validation:

precision recall f1-score support

0 0.98 1.00 0.99 178

1 1.00 0.98 0.99 182

avg / total 0.99 0.99 0.99 360

Confusion matrix:

[[178 0]

[ 4 178]]

ח.

Logistic regression using [hugesum, varx, tri\_down , tri\_up , ratio\_up\_low , var\_of\_m\_and\_t , sum\_1 , max\_1 ,zeros\_at\_center ,large] features cross validation:

precision recall f1-score support

0 0.95 0.98 0.96 178

1 0.78 0.79 0.78 182

2 0.85 0.76 0.80 177

3 0.74 0.78 0.76 183

4 0.89 0.88 0.89 181

5 0.67 0.65 0.66 182

6 0.93 0.92 0.92 181

7 0.90 0.87 0.88 179

8 0.72 0.71 0.71 174

9 0.65 0.73 0.69 180

avg / total 0.81 0.81 0.81 1797

Confusion matrix:

[[174 2 0 0 2 0 0 0 0 0]

[ 1 144 11 0 1 1 3 2 8 11]

[ 1 14 134 12 0 9 3 1 1 2]

[ 0 3 2 143 1 5 1 3 7 18]

[ 5 0 0 0 159 2 6 6 3 0]

[ 2 2 2 4 6 118 0 2 22 24]

[ 0 0 4 2 3 3 166 0 3 0]

[ 0 2 0 0 4 3 0 156 0 14]

[ 0 17 0 8 2 22 0 0 123 2]

[ 1 1 4 23 0 12 0 4 4 131]]

1. מקור הנוסחה: <http://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html> [↑](#footnote-ref-1)
2. מקור הנוסחה: <http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html> [↑](#footnote-ref-2)