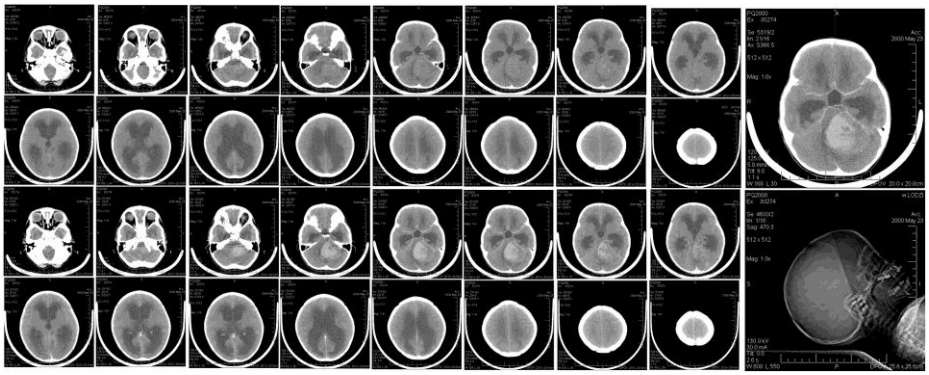
ביטוי גנים – סרטן מוח

סרטן המוח (מתוך ויקיפדיה)

גידול מוחי הוא גידול לא נורמלי ובלתי-מפוקח של תאים הנמצאים במוח: נירונים, תאי גליה, תאי אפיתל ועוד. הגידולים במוח מופיעים בעיקר בחלק האחורי של הראש בילדים ובשני השלישים הקדמיים של המוח במבוגרים. תאי המוח ניזוקים בגלל לחץ מתאי הגידול הגדלים, השפעה עקיפה על ידי תהליכים דלקתיים המתפתחים בתוך וסביב הגידול הסרטני, בצקות ו/או עלייה בלחץ בתוך הגולגולת (כתוצאה מהבצקת או מחסימת הזרימה של הנוזל המוחי).

סרטן המוח מסוג Medulloblastoma (MD)

גידול מסוג מדולובלסטומה הינו גידול מוחי ממאיר הנפוץ ביותר בקרב ילדים, הממוקם בחלק האחורי התחתון של המוח. גידול זה מתפשט במהירות לאיזורים אחרים בחלל המוח.
התסמינים הקליניים כוללים הקאות, כאבי ראש, אטקסיה, והידרדרות של הראיה.



סריקת CT של חולה בת 6 המראה גידול סרטני מסוג MD.

תיאור הנתונים

כל דגימה מבטאת 7129 סוגים של גנים, אינדיקציה אם החולה נפטר וכן תוחלת חיים של החולה (סה"כ 7131 פיצ'רים).
מהנתונים עולה כי נבדקו 60 חולים בסרטן מסוג MD. לכל 60 הנבדקים גידול סרטני מסוג Medulloblastoma.

מועד לקיחת דגימת ביטוי הגנים הוא לפני קבלת הטיפול, כאשר המטופלים קיבלו טיפולים כימותרפיים דומים ובוצע מעקב רפואי אחריהם.

התפלגות הנתונים באופן כמותי

מתוך הנבדקים 39 נשארו בחיים והשאר (21) נפטרו.

גיל האבחון נע בין 7 חודשים ל-38 שנים וחודשיים, כאשר החציון הינו 6 שנים וחודש.

מתוך החולים 21 בנות ו39 בנים. ל-14 יש סרטן MD מסוג Desmoplastic וליתר (46) סרטן MD מסוג Classic.
זמן ההישרדות של המטופלים שלא שרדו נע בין 5 ל-102 חודשים, כאשר החציון הינו 19 חודשים.
יתר הנתונים קשורים לביטוי הגנים של כל מטופל לפני מתן הטיפול. נסיר מן הנתונים את 59 הגנים המשמשים לבקרה טכנית, ונשתמש בהם רק לצורך איתור תצפיות חריגות.

תצפיות חריגות

במידה וימצאו תצפיות חריגות, נתעלם מהן, מכיוון שמספר התצפיות שלנו הוא קטן יחסית.
במידת הצורך נשתמש ב-59 הגנים המשמשים לבקרה טכנית לטובת בדיקת תצפיות חריגות.

שאלת מחקר ראשונה

חיזוי זמן ההשרדות הנובע מביטוי קבוצת גנים רלוונטיים. כלומר, בהינתן חולה, נרצה לנבא על סמך בדיקה של הגנים שבמאגר כמה זמן נותר לו לחיות.

רקע לבעיה

בשאלה זו אנו מנסים לחזות בכלים מתמטיים זמן השרדות של מטופל בהינתן ביטוי הגנים שלו.

דילמות ובעיות העומדות בפנינו:

- מספר הפיצ'רים גדול מאוד ביחס למספר התצפיות ($p \gg n$) ולכן יתכן שנבצע feature selection או הורדת מימד על מנת להגיע לתוצאות טובות.
- יתכן שישנן השפעות חיצוניות אשר אינן ידועות לנו העלולות להשפיע על תוצאות החיזוי.
- מתוך 60 החולים 39 נשאר בחיים. על כן, אנו יודעים כמה זמן הם שרדו עד כה אך לא יודעים מהי תוחלת החיים שלהם לאחר האבחון.

שאלת מחקר שניה

בעקבות הבעיה האחרונה שהצגנו (חוסר ידיעה של זמן ההשרדות האמיתי עבור מטופלים ששרדו) ובעקבות בחינה מחודשת של המאמר אותו קיבלנו, הגענו למסקנה שעדיף לשנות את שאלת המחקר לבעית הסיווג הבאה:

סיווג מטופל בעל דוגמת גנים, הנלקחת לפני מתן טיפול, כבעל פוטנציאל גבוה לחיות ("שורד") או כבעל פוטנציאל גבוה למות ("לא שורד").

רקע לבעיה

בשאלה זו אנו מנסים לחזות בכלים מתמטיים את סיווג המטופל בהינתן ביטוי הגנים שלו לאחת משתי המחלקות אותן נסמן כ"שורד" או "לא שורד".
דילמות ובעיות העומדות בפנינו:

- מספר הפיצ'רים גדול מאוד ביחס למספר התצפיות ($p \gg n$) ולכן יתכן שנבצע feature selection או הורדת מימד על מנת להגיע לתוצאות טובות.
- יתכן שישנן השפעות חיצוניות אשר אינן ידועות לנו העלולות להשפיע על תוצאות החיזוי.
- יתכן שלא יהיה צורך להשתמש בנתון "זמן השרדות" בעקבות הבעייתיות שהצגנו בשאלת המחקר הראשונה.
- לא נבדיל בין סוגי סרטן MD השונים (Desmoplastic ו-Classic).

כלים ותוכנות

נשתמש ב- R ובפייתון.

חלוקת עבודה מוצעת

נחקור את שאלת המחקר באמצעות ארבעה אלגוריתמים:

אלגוריתם 1 (LDA) – מאיה ואודליה.

אלגוריתם 2 (SVM) –עודד.

אלגוריתם 3 (רגרסיה לוגיסטית)– מאיה ואודליה.

אלגוריתם 4 (KNN) – אינה.

ניתוח התוצאות ומסקנות באחריות כל חברי הקבוצה.

שיטות

מבוא

אנו נשתמש ב-5 שיטות לניבוי.

1. LDA – נבצע סיווג באמצעות LDA, המבצע הורדת מימד, ונבחן את התוצאות באמצעות CV.
2. SVM – נבצע סיווג באמצעות אלגוריתם הסיווג SVM תוך בחינת פרמטרים שונים.
3. רגרסיה לוגיסטית – נבצע סיווג באמצעות רגרסיה לוגיסטית. נבחר ערך τ מתאים בהתאם לתוצאות שנקבל.
4. KNN – נבצע סיווג באמצעות K השכנים הקרובים ביותר. נבדוק ערכי K שונים ואת השפעתם על התוצאות.
5. נבצע אנסמבל של השיטות לפי 3 השיטות הטובות ביותר.

תיאור המשתנים

יהיו $X_1, X_2, .., X_{60}$ משתנים מקריים בממד 7070 (לאחר הסרת גני הבקרה) כאשר כל ממד הינו משתנה מקרי שנדגם מהתפלגות לא ידועה ומבטא רמה של גן, ומימד נוסף מבטא את זמן ההישרדות של כל מטופל.

יהי Y – מ"מ בינארי המבטא את תוצאת הסיווג.

$$\mathrm{Y} = \begin{cases} 1, & X\ patiant\ is\ classified\ as\ non-survivor \\ 0, & \mathrm{o/w} \end{cases}$$

המטרה:

מציאת פונקציה f המקבלת וקטור ביטוי גנים + זמן הישרדות של מטופל X ומסווגת: $f: X \to Y$

האילוצים:

לא ניתן להניח אי-תלות בין המשתנים המבטאים את רמות הגנים: X_{ij} , $1 \le j \le 7070$. נצטרך להתחשב זאת במבחנים סטטיסטיים, אם נבחר לבצע כאלה.

תיאור האלגוריתם הראשון

תיאור האלגוריתם- LDA

שיטה זו מוצאת קומבינציה לינארית של ה-features הגורמת להפרדה לינארית של 2 מחלקות ולהורדת מימד.

הנחת המודל – התפלגות רב נורמלית.

כלל ההחלטה: סיווג לקבוצה 1 אם מתקיים $(\hat{\mu}_1 - \hat{\mu}_2)^T S_{pooled}^{-1} \left(x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}\right) \ge \log\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right)$

א. נשתמש ב-10-fold Cross Validation על מנת לבחון את תוצאות הסיווג.
ב. נשתמש ב-5-fold Cross Validation על מנת לבחון את תוצאות הסיווג.
ג. נשתמש ב-Leave One Out Cross Validation על מנת לבחון את תוצאות הסיווג.

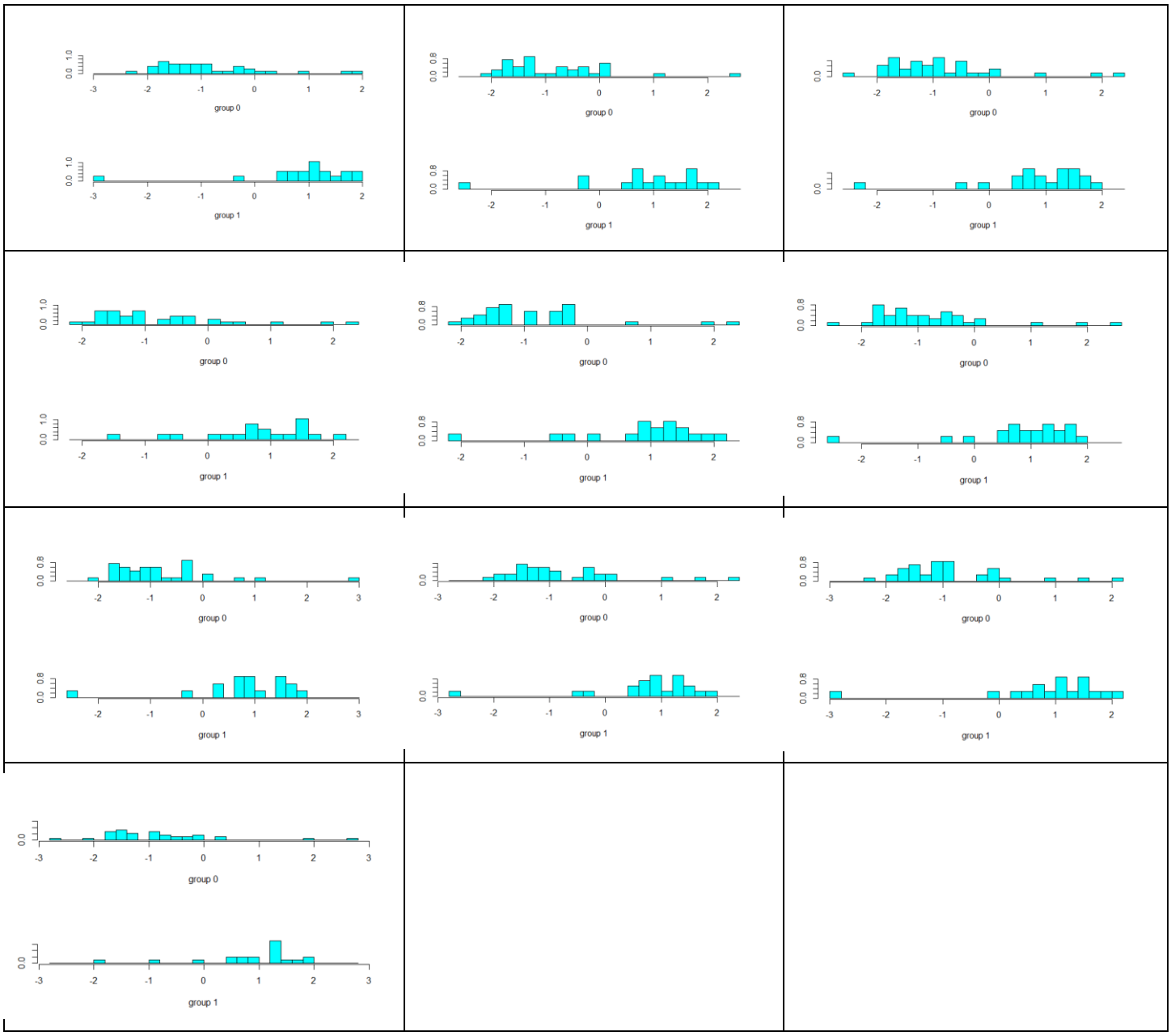
הצגת התוצאות

נציג לכל fold את ההיסטוגרמות שהתקבלו מהרצת ה-LDA.

כמו כן נציג מטריצות בלבול ממוצעות לכל ה-fold-ים, עבור קבוצת המבחן וקבוצת האימון.

:10-fold Cross Validation

- היסטוגרמות ה-LDA



ניתן לראות שיש הפרדה לינארית טובה ע"י העברת קו מפריד באזור ה-0.

נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול ממוצעת עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	32	3
Predicted as died	3	16

לכן: $Sensitivity = \frac{16}{19} = 0.84$, $Specificity = \frac{32}{35} = 0.91$

- מטריצת בלבול ממוצעת עבור נתוני המבחן:

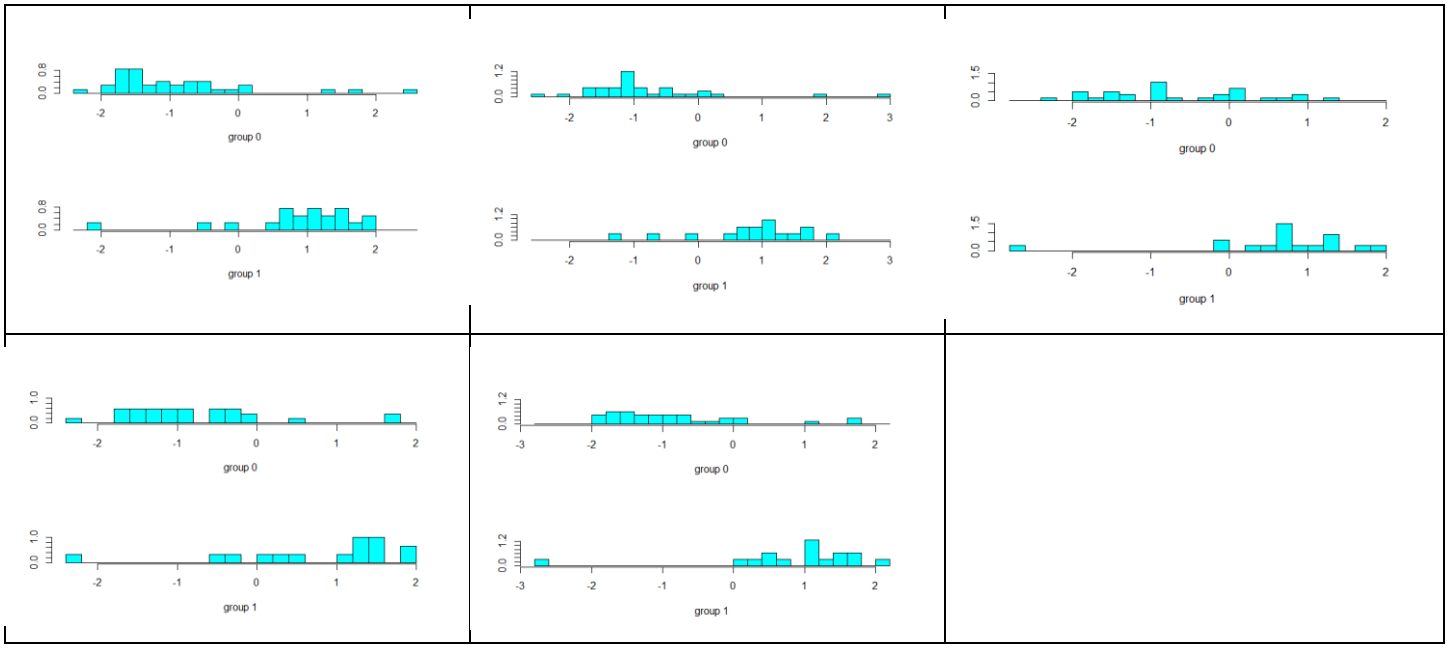
	True class survived	True class died
Predicted as survived	3	2
Predicted as died	1	0

לכן: $Sensitivity = \frac{0}{2} = 0 \ , Specificity \ = \frac{3}{4} = 0.75$

ניתן לראות שהתוצאות עבור קבוצת המבחן אינן טובות. יתכן שמקור הבעיה הוא בעובדה שקבוצת המבחן קטנה מאוד, לכן נסתכל ב-5-fold CV להשוואה.

:5-fold Cross Validation

- היסטוגרמות ה-LDA:



ניתן לראות שיש הפרדה לינארית טובה ע"י העברת קו מפריד באזור ה-0.

נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול ממוצעת עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	28	4
Predicted as died	3	13

לכן: $Sensitivity = \frac{13}{17} = 0.76 \ , Specificity \ = \frac{28}{31} = 0.9$

- מטריצת בלבול ממוצעת עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	7	4
Predicted as died	1	0

לכן: $Sensitivity = \frac{0}{4} = 0 \ , Specificity \ = \frac{7}{8} = 0.875$

הצלחנו לשפר את תוצאות המבחן עבור הדוגמאות של מטופלים ששרדו. אך עדיין קיימת בעיה עבור המטופלים שלא שרדו.

- ייתכן שמקור הבעיה הוא העובדה שמספר הדוגמאות של מטופלים שלא שרדו הוא קטן ממספר הדוגמאות של מטופלים ששרדו (1/3 לא שרדו ו-2/3 שרדו). נבחן זאת ע"י השוואה לשיטות הבאות בהן נשתמש לסיווג.
- ייתכן שמקור הבעיה הוא מדגם האימון הקטן מאוד. לכן ננסה לבחון את התוצאות באמצעות LOOCV.

Leave One Out Cross Validation:

נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול ממוצעת עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	35	3
Predicted as died	3	17

לכן: $Sensitivity = \frac{17}{20} = 0.85 \ , Specificity \ = \frac{35}{38} = 0.921$

- מטריצת בלבול ממוצעת עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	1	0
Predicted as died	0	0

לכן: $Sensitivity = 0 \ (or\ undefined) \ , Specificity \ = 1$

הצלחנו לשפר את תוצאות המבחן עבור הדוגמאות של מטופלים ששרדו. אך עדיין קיימת בעיה עבור המטופלים שלא שרדו.

כמו כן, ביצענו הרצה של 5-fold CV ו-10-fold CV ללא הפיצ'ר "זמן השרדות", מתוך חשד שהוא עלול להשפיע לרעה על תוצאת הסיווג. קיבלנו תוצאות דומות לתוצאות המוצגות להלן, ולכן לא צירפנו אותן לחלק זה של המסמך.

תיאור האלגוריתם- SVM

זהו אלגוריתם למידה חישובית מונחית. דוגמאות האימון מיוצגות כווקטורים במרחב לינארי. בשלב האימון נבנה מסווג אשר מטרתו היא להגיע להפרדה מקסימלית, כלומר יוצר מרווח גדול ככל האפשר בינו לבין הדוגמאות הקרובות לו ביותר בשתי הקטגוריות.

1. אנו נריץ את האלגוריתם תוך כדי שימוש בפונקציות kernel שונות ונבצע בחירת פרמטרים באמצעות grid-search (ובאמצעות 5-Cross Validation). על קבוצת האימון נבצע את בחירת הפרמטרים ואת הלמידה על ידי האלגוריתם בעל הפרמטרים שנבחרו. גודלה יהיה 90% מהתצפיות. 10% הנותרים יהיו קבוצת המבחן.
   א. Linear Kernel
   ב. RBF Kernel

2. נשים לב כי מספר הפיצ'רים שלנו גדול מאוד ביחס למספר הדגימות (7070 לעומת 60). לכן, מתבקש לנסות ולבצע feature selection. ננסה לצמצם את הפיצ'רים ל-100 הפיצ'רים אשר יעזרו לנבא באופן הטוב את השדה Died. לשם כך, נשתמש במבחן הסטטיסטי chi^2 ונבחר את ערכי 100 הפיצ'רים בעלי ערכי p הטובים ביותר. לאחר בחירת הפיצ'רים, נחזור על 1.

Linear Kernel:

נבחר את ערך הפרמטר C של המודל מבין הערכים הבאים $C\epsilon\{0.1, 1, 10\}$. נתאר את מהלך ריצת האלגוריתם (בעזרת דו"ח שכתבנו):

Tuning hyper-parameters:

Best parameters set found on development set:

{'kernel': 'linear', 'C': 0.1}

Grid scores on development set:

0.648 (+/-0.126) for {'kernel': 'linear', 'C': 0.1}

0.648 (+/-0.126) for {'kernel': 'linear', 'C': 1}

0.648 (+/-0.126) for {'kernel': 'linear', 'C': 10}

Detailed classification report:

The model is trained on the full development set.

The scores are computed on the full evaluation set.

	precision	recall	f1-score	support
0	1.00	0.75	0.86	4
1	0.67	1.00	0.80	2
avg / total	0.89	0.83	0.84	6

כעת, נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	35	0
Predicted as died	0	19

לכן: $Sensitivity = 1$, $Specificity = 1$

- מטריצת בלבול עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	3	0
Predicted as died	1	2

לכן: $Sensitivity = 1,\ Specificity\ = 0.75$

ממטריצת הבלבול של האימון, ניתן לראות כי אין טעויות בסיווג. כלומר, מצאנו מפריד לינארי (Hard Margin) לקבוצת האימון.
נציין, כי בחלק מההרצות הסיווג היה מושלם וקיבלנו $Sensitivity = Specificity\ = 1.0$

נבחן גם את התוצאות עבור קבוצת אימון המהווה 80% מהמדגם ונריץ באותו האופן (כולל כיוונון פרמטרים מחודש):

- מטריצת בלבול עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	31	0
Predicted as died	0	17

לכן: $Sensitivity = 1\ ,Specificity\ = 1$

- מטריצת בלבול עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	6	2
Predicted as died	2	2

לכן: $Sensitivity = 0.5\ ,Specificity\ = 0.75$

ניתן לראות כי קיבלנו תוצאות פחות טובות. ניתן להניח שמכיוון שמספר הדגימות נמוך גם כך, הקטנת קבוצת הדגימות שהמסווג לומד תפגע באופן מובהק יותר בביצועיו ותקשה עליו למצוא פיצ'רים שאכן מייצגים טוב את הנתונים ולא רק את קבוצת המדגם (סכנת התאמת היתר חמורה יותר ככל שקבוצת האימון קטנה).

הערה: מכיוון שאנו שואפים למודלים פשוטים יותר כדי להישאר נאמנים לעקרון תערו של אוקאם (בין השאר כדי למנוע מצב של התאמת יתר), לא נראה תוצאות של polynomial kernels.

RBF Kernel:

נבחר מבין הפרמטרים הבאים של המודל:

$C \in$ {1.00000000e-02, 1.00000000e-01, 1.00000000e+00,
1.00000000e+01, 1.00000000e+02, 1.00000000e+03,
1.00000000e+04, 1.00000000e+05, 1.00000000e+06,
1.00000000e+07, 1.00000000e+08, 1.00000000e+09,
1.00000000e+10}

gamma $\in$ {1.00000000e-09, 1.00000000e-08, 1.00000000e-07,
1.00000000e-06, 1.00000000e-05, 1.00000000e-04,
1.00000000e-03, 1.00000000e-02, 1.00000000e-01,
1.00000000e+00, 1.00000000e+01, 1.00000000e+02,
1.00000000e+03}

נתאר את מהלך ריצת האלגוריתם עבור קבוצת אימון של 90% (בעזרת דו"ח שכתבנו):

Tuning hyper-parameters:

Best parameters set found on development set:

{'kernel': 'rbf', 'C': 10000.0, 'gamma': 1.0000000000000001e-05}

Grid scores on development set:

0.667 (+/-0.057) for {'kernel': 'rbf', 'C': 0.01, 'gamma': 1.0000000000000001e-09}

0.667 (+/-0.057) for {'kernel': 'rbf', 'C': 0.01, 'gamma': 1e-08}

-
-
- (עקב ריבוי הנתונים לא הכנסנו את הדו"ח המלא)

0.667 (+/-0.057) for {'kernel': 'rbf', 'C': 10000000000.0, 'gamma': 100.0}

0.667 (+/-0.057) for {'kernel': 'rbf', 'C': 10000000000.0, 'gamma': 1000.0}

Detailed classification report:

The model is trained on the full development set.

The scores are computed on the full evaluation set.

	precision	recall	f1-score	support
0	0.60	1.00	0.75	3
1	1.00	0.33	0.50	3
avg / total	0.80	0.67	0.62	6

כעת, נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	36	0
Predicted as died	0	18

לכן: $Sensitivity = 1$, $Specificity = 1$

- מטריצת בלבול עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	3	2
Predicted as died	0	1

לכן: $Sensitivity = 0.3333$, $Specificity = 1$

ניתן לראות כי קיבלנו התאמת יתר. בעוד המסווג מסווג נכונה את כל מרחב המדגם, הוא טועה עבור קבוצת המבחן ומסווג אנשים מתים לחיים (ייתכן שמשום שיחס החיים למתים עומד על 2:1). התאמת היתר נובעת ממיעוט הדוגמאות לעומת ריבוי הפיצ'רים ומגמישות הקרנל – קל לו לסווג את נתוני האימון ולמצוא בהם חוקיות שתקפה לקבוצה זו אך אינה תופסת את מהות הנתונים ולכן אינה תקפה בעבור מדגם המבחן.

Feature Selection:

כפי שכבר הזכרנו, בדאטא, כמות הפיצ'רים שאנו מקבלים גדולה מאוד ביחס למספר הדגימות. לכן, סביר להניח שיש רעש רב והסכנה ל-overfitting גדולה משום שייתכן שההפרדה תעשה ע"פ גנים אשר גורמים להפרדה טובה במדגם אך אינם מבטאים את המציאות נכונה. כדי להתגבר על כך, נרצה לבצע feature selection. נרצה לשמור את הפיצ'רים שיבטאו באופן הטוב ביותר את המציאות. נציע את הדרך הבאה: בעבור על פיצ'ר, נבדוק כמה מובהק הקשר שלו לפרמטר Died אותו אנו מנסים להעריך. זאת נעשה על פי המבחן הסטטיסטי chi^2. נאסוף את ערכי ה-p, ונבחר את k הפיצ'רים המתאימים לערכי ה-p הטובים ביותר.

כעת, נראה תוצאות בעבור k=100. נתחיל מהמסווג שהתקבל תוך כדי שימוש ב RBF kernel (מלבד בחירת הפיצ'רים, לא נעשו שינויים בהרצת האלגוריתם).

- מטריצת בלבול ממוצעת עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	35	3
Predicted as died	1	15

לכן: $Sensitivity = 0.8333 \ , Specificity \ = 0.9722$

- מטריצת בלבול ממוצעת עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	3	1
Predicted as died	0	2

לכן: $Sensitivity = 0.6667 \ , Specificity \ = 1$

ניתן לראות כי בעיית התאמת היתר קטנה והתוצאות השתפרו. כדי להראות שהתוצאות אינן מקריות, נבחן כעת 100 אקראיים:

	True class survived	True class died
Predicted as survived	3	3
Predicted as died	0	0

לכן: $Sensitivity = 0 \ , Specificity \ = 1$

כלומר, קיבלנו סיווג על פי הרוב. מכאן, נראה כי השיטה שלנו צלחה.

ננסה גם עבור המסווג שקיבלנו עבור הקרנל הלינארי, כאשר מדגם האימון היה 80% והמבחן 20%:

	True class survived	True class died
Predicted as survived	6	1
Predicted as died	2	3

לכן: $Sensitivity = 0.75 \ , Specificity \ = 0.75$

ניתן לראות כי גם במקרה זה השגנו שיפור.

הערה: בכל הריצות קבוצת המבחן והאימון חולקו באופן אקראי. בנוסף הפיצ'רים נורמלו לפי דרישות האלגוריתם.

תיאור האלגוריתם- רגרסיה לוגיסטית

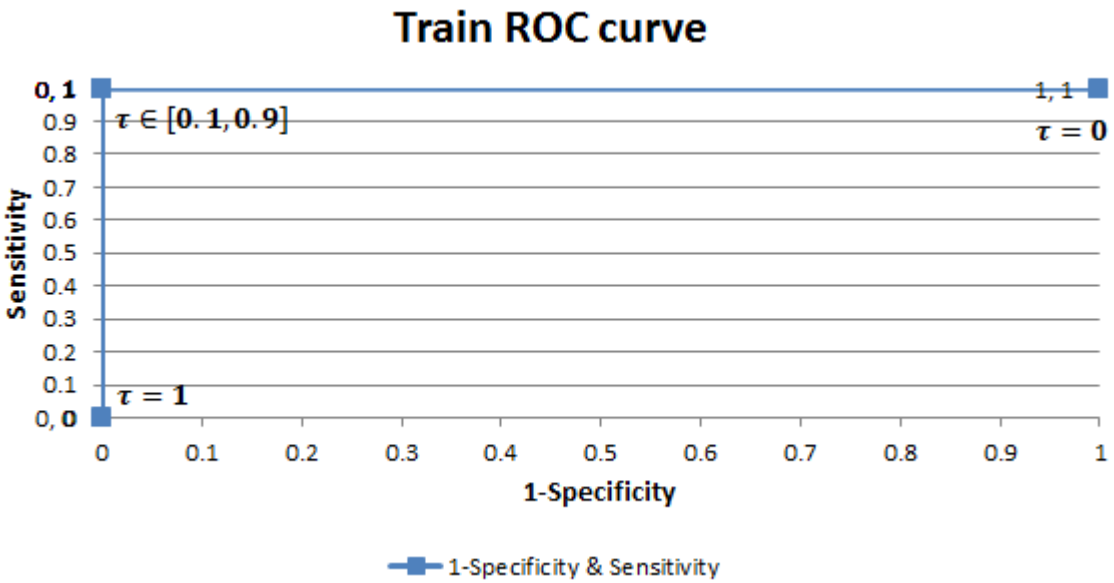
שיטה זו מתאימה מודל רגרסיה ומחזירה עבור קלט x פלט שהוא משתנה קטגוריאלי (במקרה שלנו, "שורד"/"לא שורד").

כלל ההחלטה: סיווג לפי הכלל הבא- $y(x) = \begin{cases} 1, & p(x) > \tau \\ 0, & p(x) \leq \tau \end{cases}$, כאשר 1 מציין מטופל שלא שרד.

נבחן ערכי τ שונים ואת השפעתם על טיב הסיווג. נשתמש ב-80% אימון ו-20% מבחן.

הצגת התוצאות

עקומת ה-ROC עבור נתוני האימון



הערך האופטימלי של τ מתקבל עבור הערך הקרוב ביותר ל- $sensitivity = 1, 1 - specifity = 0$.

כלומר במקרה שלנו, הנקודה האופטימלית היא $\tau \in [0.1,0.9]$.

- מטריצת הבלבול המתאימה לנקודה זו:

	True class survived	True class died
Predicted as survived	32	0
Predicted as died	0	12

$(Sensitivity = 1 \ , Specificity \ = 1)$

ניתן לראות כי הערכים שקיבלנו עבור נתוני האימון הינם הערכים האופטימליים.

תוצאות המבחן עבור ערך ה-τ האופטימלי

הערך האופטימלי של τ מתקבל עבור $\tau \in [0.1,0.9]$

- מטריצת הבלבול עבור נתוני המבחן המתאימה לנקודה זו:

	True class survived	True class died
Predicted as survived	3	4
Predicted as died	2	3

$(Sensitivity = 0.4285714\ ,Specificity\ = 0.6)$

ניתן לראות כי ערך ה- $Specificity$ שקיבלנו עבור נתוני המבחן נמוך לעומת ערכו עבור נתוני האימון ובנוסף גם ערך ה- $Sensitivity$ המתקבל נמוך משמעותית, כלומר קיימת בעיה של $overfitting$, וכן שוב נתקלנו בבעיית סיווג המטופלים שאינם שורדים.

תיאור האלגוריתם הרביעי

תיאור האלגוריתם- KNN

לפי אלגוריתם KNN דוגמאות האימון הם וקטורי תכונות במרחב רב ממדי, כל אחת עם סיווג. שלב האימון של האלגוריתם מתבסס רק על אחסון הווקטור והסיווג של דוגמאות האימון.

בשלב הסיווג נבדקים K הדוגמאות השכנות הקרובות ביותר לפי מטריקת מרחק אוקלידי, והסיווג נקבע לפי סיווג רוב הדוגמאות מתוך ה-K הנבחר.

נבחן ערכי k שונים ואת השפעתם על טיב הסיווג.

נחלק את הנתונים ל -90% אימון ו-10% מבחן:

מכיוון שאנו משתמשים במסווג בינארי נבחר ב-k אי זוגי על מנת להימנע ממצב של תיקו.

נבצע את המבחן עבור k = 5, k = 7. בחרנו ערכי K הקרובים לשורש מספר התצפיות.

עבור k = 5 מתקבלות התוצאות הבאות:

נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	30	4
Predicted as died	8	11

לכן: $Sensitivity = 0.733,\ Specificity\ = 0.789$

- מטריצת בלבול עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	5	0
Predicted as died	1	1

לכן: $Sensitivity = 1\ ,Specificity\ = 0.833$

עבור k = 7 מתקבלות התוצאות הבאות:

נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	32	2
Predicted as died	10	9

לכן: $Sensitivity = 0.818$, $Specificity = 0.762$

- מטריצת בלבול ממוצעת עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	5	0
Predicted as died	0	2

לכן: $Sensitivity = 1$, $Specificity = 1$

נשתמש ב-80% אימון ו-20% מבחן:

עבור k = 5 מתקבלות התוצאות הבאות:

נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	27	4
Predicted as died	7	10

לכן: $Sensitivity = 0.714$, $Specificity = 0.794$

- מטריצת בלבול עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	6	2
Predicted as died	2	2

לכן: $Sensitivity = 0.5$, $Specificity = 0.75$

עבור k = 7 מתקבלות תוצאות הבאות:

נבחן את התוצאות באמצעות מטריצת בלבול:

- מטריצת בלבול עבור נתוני האימון:

	True class survived	True class died
Predicted as survived	29	2
Predicted as died	9	8

לכן: $Sensitivity = 0.8$, $Specificity = 0.763$

- מטריצת בלבול עבור נתוני המבחן:

	True class survived	True class died
Predicted as survived	6	2
Predicted as died	2	2

לכן: $Sensitivity = 0.5 \ , Specificity \ = 0.75$

מסקנה: שוב ניתן לראות שהמודל בעל מדגם האימון הגדול יותר (90%) הציג תוצאות טובות יותר מאשר מודל ה-80%. כמו כן, בעבור בחירה של k=7 התקבלו תוצאות זהות או טובות יותר במדגם המבחן, מאשר בחירה של k=5. זאת מכיוון ש-k=5 מניב מודל בעל שונות גבוהה, שכן הוא מותאם יותר לנתוני האימון (overfitting).

תיאור האלגוריתם החמישי

תיאור האלגוריתם- ensemble learning

נשתמש באלגוריתמים בעלי הפרמטרים הטובים ביותר שהתקבלו מכל שיטה.

קיבלנו תוצאות אופטימליות עבור מדגם אימון גדול (90%) ולכן נשתמש בגודל מדגם זה. תוצאת הסיווג תקבע לפי עקרון הרוב.

טבלת תוצאות

Method/sample	LDA	SVM (Linear, no feature selection)	Logistic Regression	KNN (k=7)	True Class
5	0	1	1	1	1
8	0	1	1	1	1
24	0	0	0	0	0
28	0	0	0	0	0
42	0	0	0	0	0
54	0	0	0	1	0

מטריצת בלבול:

	True class survived	True class died
Predicted as survived	4	0
Predicted as died	0	2

לכן: $Sensitivity = 1 \ , Specificity \ = 1$

תוצאות

האלגוריתם שנתן את הביצועים האופטימליים (והמושלמים)היה ה-ensemble.

מבין האלגוריתמים הבודדים, האלגוריתמים הבאים נתנו את התוצאות הטובות ביותר:

- SVM עם הפרמטרים: קרנל לינארי ו-c=0.1
- KNN עם הפרמטר k=7
- Linear Regression

ניתוח ומסקנות

במהלך העבודה ניתחנו את הנתונים באמצעות 4 שיטות עיקריות, כל אחת עם יתרונותיה וחסרונותיה. לבסוף ניסינו לבצע שילוב בין 4 השיטות תוך מחשבה שכך נוכל למקסם את היתרונות ולכפר על החסרונות של כל שיטה.

במהלך העבודה נתקלנו במספר בעיות, להן נתייחס כעת:

- תצפיות חריגות
לאורך כל העבודה, בבחינת השיטות השונות, לא נתקלנו בבעיות סיווג שעלולות לנבוע מתצפיות חריגות. כלומר, ראינו ביצועי סיווג סבירים עד טובים בכל השיטות, ולכן לא חשדנו כי קיימות תצפיות חריגות. מפאת אורך המסמך, לא הוספנו בדיקה זו לפרויקט.
- שימוש בנתון זמן הישרדות
לאורך כל העבודה ראינו ביצועי סיווג סבירים עד טובים בכל השיטות. כאשר בדקנו עבור שיטת ה-LDA את השפעת פיצ'ר זה על ביצועי הסיווג לא ראינו הבדל, ולכן לא הכנסנו את התוצאה הזו למסמך והפסקנו לחקור את הנושא.
- נתונים ממימד גבוה (הורדת מימד)
תחילה, בכל השיטות הסרנו את 59 גני הבקרה (control) – הורדת מימד ראשונית.
התחלנו בשיטת ה-LDA שמבצעת הורדת מימד כחלק מהאלגוריתם. משם המשכנו לחקור את השפעת הפיצ'רים בשיטת ה-SVM, שם התמקדנו בבעיה זו באופן מורחב, וניתן למצוא הסברים רבים בחלק זה של המסמך. בסופו של דבר, מסקנתינו היא שע"י ביצוע הורדת מימד בעיית התאמת היתר קטנה והתוצאות השתפרו.
- בעיה בסיווג מטופלים כ"לא שורדים" (סיווג מסוג TP)
ראינו כי בעיקר בשיטת ה-LDA היה קושי רב בסיווג מטופלים כ"לא שורדים". כמו כן בשאר השיטות ערך מדד ה-Sensitivity היה נמוך או כמעט זהה לערך ה-Specificity .
ניתן לייחס בעיה זו לנתוני האימון: כפי שראינו, יש לנו מס' דוגמאות קטן מאוד (60). כמו כן, מספר הדוגמאות המסווגות כ"לא שורד" הוא נמוך עוד יותר- רק כ-$\frac{1}{3}$ מהדוגמאות. לכן ניתן להניח שהמודלים השונים מוטים לטובת ערך ה-prior הגבוה יותר, שהוא הסיווג כ"שורד", ולכן הטעות מסוג FN היא גדולה יותר.

דיון והצעות לעתיד

לפי התוצאות שקיבלנו, המלצתנו היא לשלב בין השיטות ולקבוע את הסיווג לפי עקרון הרוב, כאשר כדאי לבחור בשיטות SVM, KNN ו-Linear Regression עם הפרמטרים האופטימליים הנ"ל.

כדאי לנסות לשחזר את המחקר ובכך להשיג מספר דגימות נוסף, כלומר להגדיל את מאגר האימון (והמבחן). מהממצאים אליהם הגענו ומהמסקנות ניתן להניח שהוספת דגימות מטופלים נוספות (בפרט מהסוג שמסווג כ"לא שורד") ישפרו את יכולת הלמידה והסיווג של האלגוריתמים השונים וכן של האלגוריתם המשולב.

נספחים

- Code : https://github.com/OdedH/DMBI.git

מקורות

- *Prediction of central nervous system embryonal tumour outcome based on gene expression,* NATURE (VOL 415) 24 January 2002
- Brain Tumor, Medulloblastoma from Wikipedia
- Performing CV in R: https://gist.github.com/bhoung/11237681
- R library for ROC plots: http://rocr.bioinf.mpi-sb.mpg.de/ROCR.pdf
- Python library for machine learning: http://scikit-learn.org/stable/