מבוא לבינה מלאכותית - 236501

תרגיל בית 3

שם: עמית רוטנר

תעודת זהות: 312176050

שם: שקד דורון

תעודת זהות: 205819162

תאריך הגשה: 22.1.2020

חלק א':

1. מבנה עץ החלטה:

(א) בכל שלב עץ ההחלטה מפצל את הדוגמאות לפי תכונה מסויימת.
על מנת שעץ ההחלטה יוכל להפריד בין הדוגמאות החיוביות והשליליות תוך פיצול יחיד, אנו נרצה כי הישר $y = mx + n$ יהיה ישר מקביל לציר ה-x, כלומר אם $m = 0$ ואז הישר הוא $y = n$.
במקרה כזה נוכל להפריד בין הדוגמאות באמצעות צומת אשר בודק אם $y > n$ ואז מסווג את הדוגמה כשלילית ואחרת מסווג את הדוגמה כחיובית.

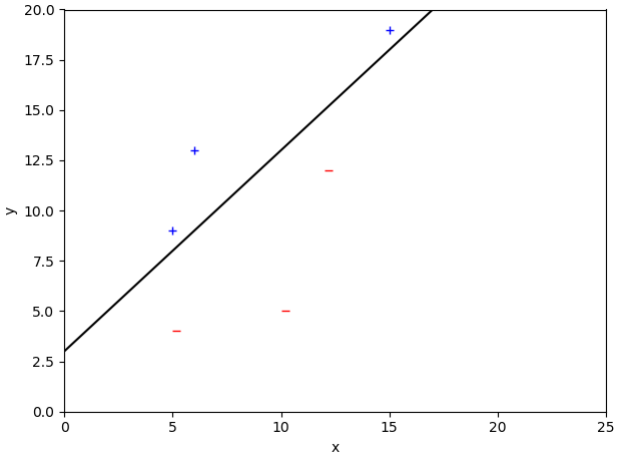
(ב) בסעיף 1 ראינו כי רק אם $m = 0$ ואז $y = n$ הוא הישר המפריד בין הדוגמאות, ניתן להפריד בין הדוגמאות תוך פיצול יחיד.
אולם, במקרה הכללי, סביר להניח כי לא קיים ישר המקביל לציר ה-x אשר מפריד בין הדוגמאות.
על כן, לא יהיה ניתן להפריד בין הדוגמאות החיוביות והשלילות תוך פיצול יחיד.

(ג) בעץ החלטה, פיצול צומת נעשה על ידי בחירת תכונה ופיצול לפיה. כפי שראינו בסעיפים הקודמים, לא תמיד ניתן להפריד בין הדוגמאות תוך פיצול יחיד. זה מתרחש כאשר הקו הישר המפריד בין הדוגמאות אינו קו המקביל לציר ה-x.
בנוסף אנו יודעים כי לכל דוגמה עם ערכי תכונות $x = a, y = b$, אם הדוגמה חיובית אז $ma + n > b$ ואחרת הדוגמה שלילית.
על כן, אם נשנה את כלל הפיצול כך שנוכל לפצל לפי שתי התכונות a, b במקביל, נוכל לבדוק האם $ma + n > b$ ובכך להפריד תוך פיצול יחיד את הדוגמאות השליליות והחיוביות.

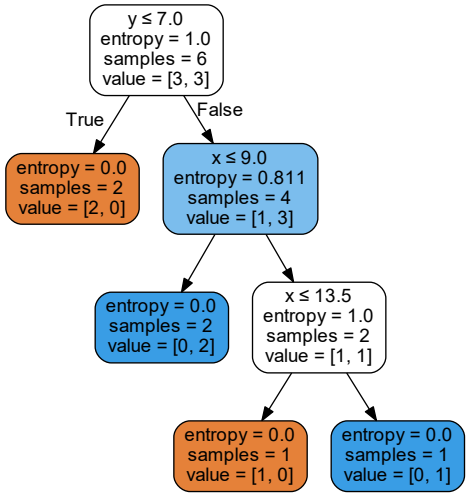
(ד) הטענה נכונה, נראה דוגמה. יהי הדאטה הבא:

x	y	$Outcome$
5	9	1
6	13	1
15	19	1
5	4	0
10	5	0
12	12	0

לפי הגרף, נשים לב כי הישר $y = x + 3$ מקיים את תנאי הבעיה:

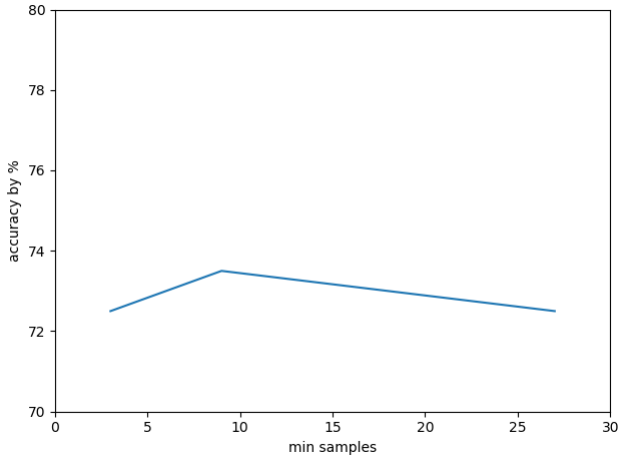


תרשים עץ ההחלטה הלא גזום $ID3$ אשר נבנה באמצעות הדאטה:



עבור המסלול הימני מהשורש לאחד העלים, מתבצעים שני פיצולים לפי אותה תכונה (x).

2. גרף המתאר את דיוק (באחוזים) עצי ההחלטה הגזומים על קבוצת המבחן כתלות בגודל $x \in \{3, 9, 27\}$:



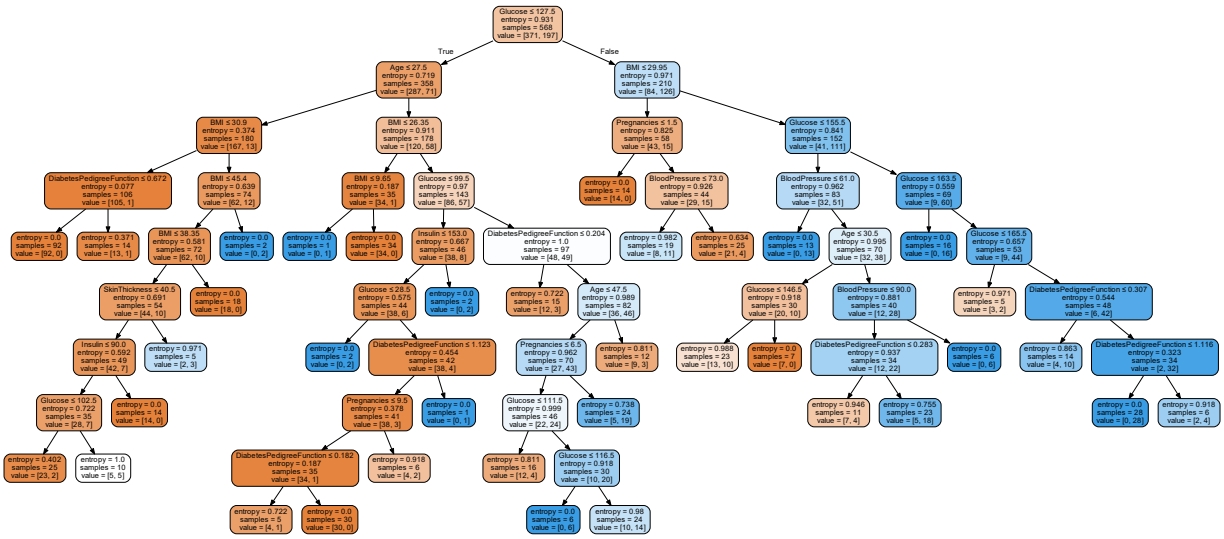
באופן כללי, גיזום עץ נעשה בכדי להקטין את העץ ולהחליש את אפקט התאמת היתר.

העץ הגזום יגדיל את שגיאת האימון בתקווה להקטנת שגיאת המבחן. נשים לב כי $DT_{x=9}$ הוא העץ בעל אחוזי הדיוק הגבוהים ביותר.

עבור $DT_{x=27}$, אמנם הקטנו הכי הרבה את העץ והגדלנו את שגיאת האימון, אך לא הצלחנו להקטין את שגיאת המבחן.

עבור $DT_{x=3}$, אמנם שגיאת האימון הקטנה ביותר, אך לא הצלחנו להקטין את שגיאת המבחן וזאת משום שכבר הנראה קיים רעש המפריע. לסיכום, כדי להגיע לדיוק המירבי תוך החלשת אפקט התאמת היתר, נבחר בעץ $DT_{x=9}$.

3. תרשים המתאר את מבנה עץ ההחלטה $DT_{x=27}$:



חלק ב':

5. עבור קבוצת אימון S לא מאוזנת:

(א) עבור עץ $ID3$ A לא גזום. נפריד למקרים:

- אם x שלילית אזי משום ש-$p \approx 1$, הרוב המוחלט של הדוגמאות בדאטה, הן שליליות. לכן, העץ A למד לסווג טוב מאוד את הדוגמאות השליליות. מכאן שההסתברות ש-A יסווג את x כשלילית גבוהה מאוד, כלומר שווה בערך ל-p.
- אם x חיובית אזי מהסיבות לעיל, העץ A למד לסווג טוב מאוד דוגמאות שליליות. לכן, הסיכוי שיסווג דוגמת מבחן חיובית כשלילית הוא $q = 1 - p$.

(ב) עבור עץ $ID3$ B גזום. נפריד למקרים:

בהנחה שבעת גיזום, סיווג עלה הוא על פי רוב הדוגמאות שבו, משום שיש פער גדול מאוד בין כמות דוגמאות האימון השליליות לבין כמות הדוגמאות החיוביות, נשער כי רוב העלים יסווגו כשליליים.
בכל פעם שיתקבל צומת עם פחות מ-y דוגמאות, נעצור את פיתוח העץ ונהפוך את הצומת לעלה.
משום שאנו בוחרים את סיווג העלה על פי הרוב, ככל ש-y גדול יותר, כך הסיכוי שרוב הדוגמאות יהיו חיוביות הולך וקטן.
כלומר, הסיכוי שהעלה יסווג כשלילי יהיה גדול מ-p. על כן, לא משנה מה הערך של x, סיכוי גדול מאוד כי יסווג כשלילי.
כלומר גדול ממש מ-p.

6. הטענה אינה נכונה. יהי הדאטה הלא מאוזן S הבא:

$Wings$	$Lay-Eggs$	$Weight$	$Outcome$
0	0	12	0
0	0	13	0
1	0	14	0
0	0	15	0
0	1	4	0
0	1	17	0
0	0	12	0
0	0	12	0
1	1	4	1
1	1	3	1

S מתאר חיות על פי התכונות הבאות:

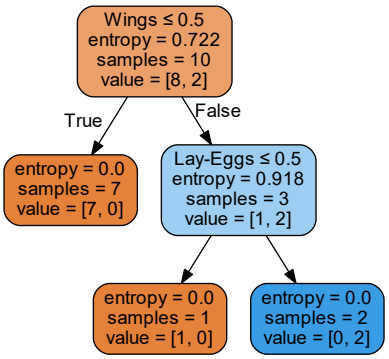
(א) בעל כנף

(ב) מטיל ביצים

(ג) משקל החיה

אנו נרצה, על פי הדאטה, לסווג האם החיה היא ציפור.

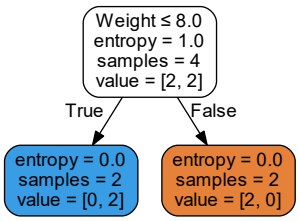
תרשים עץ ההחלטה $ID3\ A$ אשר נבנה באמצעות הדאטה הלא מאוזן:



לאחר זריקה אקראית של דוגמאות שלילות, על מנת להשוות בין גדלי המחלקות, התקבל הדאטה הבא:

$Wings$	$Lay-Eggs$	$Weight$	$Outcome$
0	0	12	0
0	0	13	0
1	1	4	1
1	1	3	1

תרשים עץ ההחלטה $ID3\ A'$ אשר נבנה באמצעות הדאטה המאוזן:



תהי דוגמת המבחן הבאה x המתארת תרנגול. תכונות x:

(א) בעל כנף - 1

(ב) מטיל ביצים - 1

(ג) משקל החיה - 4

העץ A' יסווג את x כשלילית, על פי משקלה. ואילו, העץ A יסווג את x כחיובית, משום שבעלת כנף ומטילה ביצים וזאת למרות שתרנגול אינו ציפור.

7. השגיאה הנתונה על ידי $Error_w = 4 \times FN + FP$:

(א) כאשר אנו גוזמים את העץ, אנו מסווגים כל עלה על פי רוב הדוגמאות שבו. משום שהדאטה אינו מאוזן, ישנה הסתברות גבוהה יותר כי עלה יסווג כשלילי. כלומר הדוגמאות החיוביות שנמצאות בצומת זה, יסווגו כשליליות.
כלומר, בעת גיזום עץ אנו מקטינים את TP ומגדילים את FN. מהגדרת $Error_w$, משקל FN בשגיאה הוא פי 4 מערכו המקורי ולכן משקל FN בשגיאה הוא משמעותי מאוד. מצד שני, בעץ לא גזום, אנו מגדילים את ההסתברות שדוגמה חיובית תסווג כחיוביות
וזאת משום שהדאטה בלתי תלוי (וכן סיווג עלה לא נקבע על פי רוב הדוגמאות בו).
לכן אנו מגדילים את TP ומקטינים את FN. כלומר, נקטין את $Error_w$.
לכן, נעריך כי ערך השגיאה $Error_w$ עבור עץ גזום יהיה גדול יותר מערך השגיאה עבור עץ לא גזום.

(ב) עבור עץ ההחלטה הלא גזום DT_1, ערך השגיאה הינו 122.
עבור עץ ההחלטה הגזום $DT_{x=27}$, ערך השגיאה הינו 145.
תוצאה זו אכן מתיישבת עם תשובתנו לסעיף הקודם, משום שגיזום הגדיל את ערך השגיאה.

8. איזון הדאטה:

(א) דאטה שאינו מאוזן, עשוי לגרום ליכולת ההכללה של עץ ההחלטה להיפגע. לכן, באופן אינטואיטיבי, אם נאזן את הדאטה נשפר את יכולת ההכללה של עץ ההחלטה ובכך להקטין שגיאת המבחן $Error_w$.
למשל, אם רוב הדאטה הוא דוגמאות של אנשים בריאים, אזי עץ ההחלטה שיבנה לפי דאטה זה, ככל הנראה יתקשה לסווג אדם חולה וכתוצאה מכך יגדיל את שגיאת המבחן.

(ב) ערך השגיאה $Error_w$ עבור הדאטה המאוזן הינו 138. ערך זה אינו מתיישב עם תשובתנו בסעיף הקודם.

9. שינוי הסיווג הנקבע על ידי העץ בדיעבד:

(א) נצייר את מטריצת הבלבול:

Classified \ Actual	$Sick$	$Healthy$
$Sick$	TP	FP
$Healthy$	FN	TN

כאשר $p = 1$ ואנו בשורה התחתונה, כלומר הדוגמה מקבלת סיווג שלילי, תמיד נעביר את הערך שורה אחת למעלה.
מ-FN ל-TP נרצה לעשות זאת תמיד וזאת על מנת להקטין את השגיאה. אולם, אם נעשה זאת מ-TN ל-FP, נגדיל את השגיאה.
השגיאה לפני שינוי הסיווג הינה:

$$Error_w\ before = 4 \times FN + FP$$

השגיאה לפני שינוי הסיווג הינה:

$$Error_w\ after = TN + FP$$

על מנת שלא נגדיל את ערך השגיאה, נדרוש שהשגיאה לפני שינוי הסיווג תהיה גדולה מהשגיאה אחרי שינוי הסיווג, כלומר:

$$Error_w\ before > Error_w\ after$$

ומכאן נקבל כי:

$$4 \times FN > TN$$

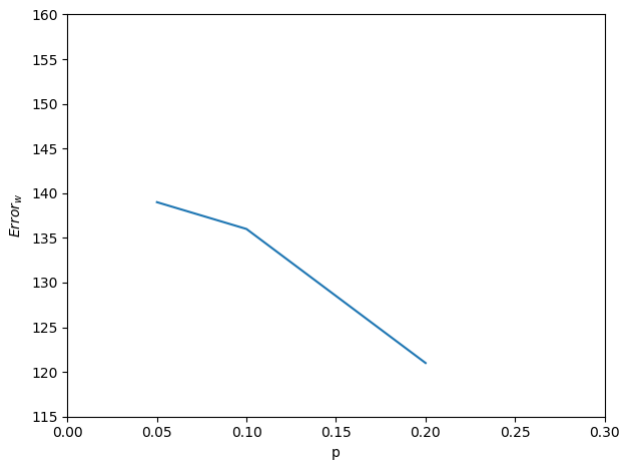
כלומר, אם $4 \times FN > TN$, כדאי יהיה לקבוע $p = 1$.

(ב) נצייר את מטריצת הבלבול עבור העץ DT_1 לפי השימוש בפרוטוקול המוצע:

Classified \ Actual	$Sick$	$Healthy$
$Sick$	$\|P\|$	0
$Healthy$	0	$\|F\|$

כאמור אנו מעבירים רק מהשורה התחתונה לשורה מעל.
בהסתברות P נעביר מהקבוצה TN בגודל $|F|$ לקבוצה FP. כלומר תוחלת גודל הקבוצה FP היא $p \cdot |F|$.
גודל הקבוצה FN הוא 0 ולכן לא נעביר אף איבר ממנה לקבוצה TP.
על כן, תוחלת ערך השגיאה $Error_w$ שווה ל-$4 \times 0 + p \cdot |F| = p \cdot |F|$.

(ג) גרף המציג את שגיאת המבחן $Error_w$ של העץ DT_1 כאשר משתמשים בפרוטוקול המוצע כתלות בגודל p:



נשים לב כי עבור $p = 0.05$ ערך השגיאה הוא 139, עבור $p = 0.1$ ערך השגיאה הוא 136 ועבור $p = 0.2$ ערך השגיאה הוא 121.
ואכן הצלחנו להקטין את ערך השגיאה עבור DT_1 באמצעות שימוש בפרוטוקול זה עבור $p = 0.2$ וזאת
משום שהשגיאה הקודמת הייתה 121. מהגרף נסיק ככל ש-p גדל כך $Error_w$ קטן. שינוי סיווג מבריא לחולה
מקטין את FN ומגדיל את TP אם האדם חולה באמת, או שמקטין את TN ומגדיל את FP, אם האדם בריא באמת.
מעבר מ-FN ל-TP מקטין את ערך השגיאה בפקטור 4 ומעבר מ-TN ל-FP מגדיל את ערך השגיאה בפקטור 1.
בהרצה זו, ניתן לראות על פי הגרף, כי ככל ש-p גדל כך העברנו יותר משקל מ-FN ל-TP מאשר העברנו מ-TN ל-FP
ובכך הצלחנו להקטין את $Error_w$.

10. נצייר את מטריצת הבלבול:

		$Actual$	
		$Sick$	$Healthy$
$Classified$	$Sick$	TP	FP
	$Healthy$	FN	TN

בהינתן עלה v בעץ ההחלטה, עם $|T|$ דוגמאות חיוביות ו-$|F|$ שליליות, נשים לב כי אם נסווג את העלה כחיובי,
כל הדוגמאות השליליות (האנשים הבריאים) יחשבו בקבוצה FP משום שיסווגו כחולים וכל הדוגמאות החיוביות (האנשים החולים)
יחשבו בקבוצה TP. כלומר ערך השגיאה $Error_w = 4 \cdot FN + FP$, גדל ב-$|F|$.
אולם, אם נסווג את העלה כשלילי, כל הדוגמאות החיוביות (האנשים החולים) יחשבו בקבוצה FN משום שיסווגו כבריאים
וכל הדוגמאות השליליות (האנשים הבריאים) יחשבו בקבוצה TN. כלומר ערך השגיאה $Error_w = 4 \cdot FN + FP$, גדל ב-$4 \cdot |T|$.
אנו רוצים להקטין את ערך השגיאה ולכן נעדיף לסווג את העלה באופן שיקטין את ערך השגיאה. כלומר לבחור באופציה הקטנה
מבין $4 \cdot |T|$ ו-$|F|$. כלומר אם $4 \cdot |T| > |F|$ נעדיף לסווג את העלה כחיובי ואחרת כשלילי.
לכן $\alpha = 4$. באותו האופן, אנו נרצה לבחור את התכונה לפיצול,על פי התכונה שממזערת את FN וזאת משום
שערך $Error_w$ גדל בפקטור 4 על כל דוגמה ב-FN. מכאן, באופן אינטואיטיבי, נעדיף תכונות
אשר תוספת האינפורמציה המחושבת על ידי פונקציית האנטרופיה הממושקלת (כפול מינוס 1) היא הגבוהה ביותר,
ביחס לכך שנרצה לתת משקל גדול יותר להסתברות שעלה יסווג כחיובי ובכך להקטין את FN.
נרצה לתת משקל פי 4 לדוגמה חיובית מאשר לדוגמה שלילית ולכן נבחר את δ להיות 0.8.

12. המניפולציה אותה נבצע על הדאטה היא הוספת 3 דוגמאות חיוביות על כל דוגמה חיובית.
כלומר, בכך הגדלנו את מספר הדוגמאות החיוביות פי 4 ומשום שלא שינינו את מספר הדוגמאות השליליות, נתנו באופן מלאכותי, משקל פי 4 לדוגמה חיובית מאשר היה בעבר. כלומר כעת, פעולת עץ $ID3$ סטנדרטי גזום עם $x = 9$ על הדאטה החדש, תהיה זה לפעולת העץ DT_2.

חלק ג':

13. כפי שנלמד בכיתה, בעת סיווג דוגמת מבחן x, מסווג KNN מודד את המרחק מדוגמאות האימון ובוחר את k הכי קרובים ל-x. סיווג x, נקבע לפי הרוב של k השכנים הללו. לכן, עבור קבוצת האימון S שהוגדרה בשאלה 5, משום ש-$p \gg q = 1 - p$, ככל ש-k יעלה, כלומר "נתייעץ" עם יותר שכנים, כך גדל הסיכוי שהרוב יהיו שליליים.
לכן, ככל ש-k יהיה גדול יותר, ההסתברות ש-C יסווג את x כשלילית, תגדל ותעלה על p.

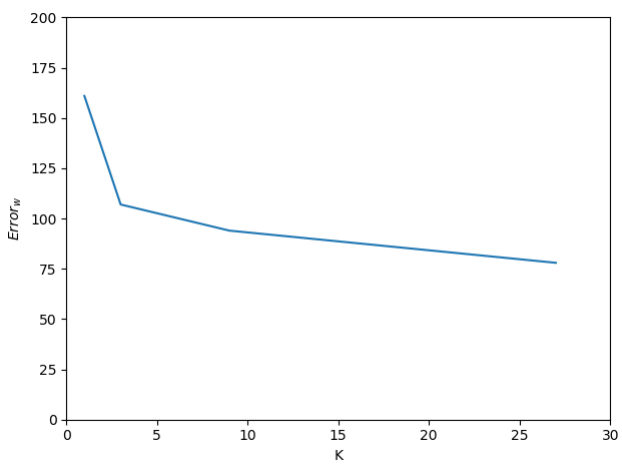
14. מטריצת הבלבול עבור מסווג זה:

$Classified$ \ $Actual$	$Sick$	$Healthy$
$Sick$	37	18
$Healthy$	34	111

ומכאן ששגיאת $Error_w$ של מסווג זה הינה:

$$\begin{aligned} Error_w &= 4 \cdot 34 + 18 \\ &= 154 \end{aligned}$$

17. גרף המתאר את $Error_w$ כתלות בגודל $k \in \{1, 3, 9, 27\}$:



על מנת לנתח את תוצאות אלו, נתבונן במטריצות הבלבול שהתקבלו:

k	$Confusion\_matrix$				$Error_w$
1			$Actual$		161
	$Classified$		$Sick$	$Healthy$	
		$Sick$	39	33	
		$Healthy$	32	96	
3			$Actual$		107
	$Classified$		$Sick$	$Healthy$	
		$Sick$	60	63	
		$Healthy$	11	66	
9			$Actual$		94
	$Classified$		$Sick$	$Healthy$	
		$Sick$	66	74	
		$Healthy$	5	55	
27			$Actual$		78
	$Classified$		$Sick$	$Healthy$	
		$Sick$	70	74	
		$Healthy$	1	55	

נשים לב כי ככל ש-k עלה, כך התייצענו עם יותר שכנים ולפיכך יחס השכנים החיוביים מול השכנים השליליים מתקרב ליחס שקיים בדאטה $train.csv$. בדאטה זה, אשר הוא לא מאוזן, ישנן בערך פי 2 דוגמאות שליליות מחיוביות.

לכן, כאשר אנו נותנים משקל של פי 4 לשכן חיובי מתוך k השכנים, דעת הרוב, תיטה לכיוון דעה חיובית.

ואכן, כפי שניתן לראות, ככל ש-k עולה, כך דעת השכנים הוסטה לכיוון חיובי. כלומר, דוגמאות מ-FN עברו ל-TP ודוגמאות מ-TN עברו ל-FP. כאשר אנו מעבירים דוגמאות מ-FN ל-TP, אנו מקטינים את ערך השגיאה ואילו כאשר אנו מעבירים דוגמאות מ-TN ל-FP, אנו מגדילים אותה.

מהגדרת $Error_w$, ישנו משקל פי 4 ל-FN ביחס ל-FP. כלומר, על מנת להקטין את $Error_w$, נשאף כי FN יירד כמה שיותר אך שהעברת הדוגמאות מ-TN ל-FP לא תעלה על השיפור שנובע מהעברת FN ל-TP.

מכאן נסיק כי אנו מוכנים "לשלם" ולהגדיל את השגיאה כתוצאה מהעברת דוגמאות מ-TN ל-FP ושהשיפור אשר ינבע מהחסרת $4 \times FN$ מהשגיאה, יהיה משמעותי יותר. ואכן, המגמה הכללית המוצגת בטבלה מתארת את הקטנת FN והגדרת TP וכן את הגדלת FP והקטנת TN באופן כזה שמקטין את $Error_w$.