מבוא לבינה מלאכותית

236501

תרגיל בית 3

22/1/2020

מגישים:

רוני אנגלנדר 312168354

נדב אורזך 311549455

שאלה 1

א. על מנת שנוכל להפריד בין הדוגמאות החיוביות והשליליות בפיצול יחיד נדרוש כי לא תהיה תלות בין שתי התכונות a ו- b. זאת על מנת שבפיצול יחיד נוכל להגדיר תכונה לפיה נוכל לחלק את הדוגמאות לחיובי או שלילי. לכן התנאי שנדרוש הוא שיתקיים $m = 0$ או $m = \pm\infty$ וכך נקבל שהישר מאונך לצירים ולמעשה נמיין רק לפי תכונה אחת.

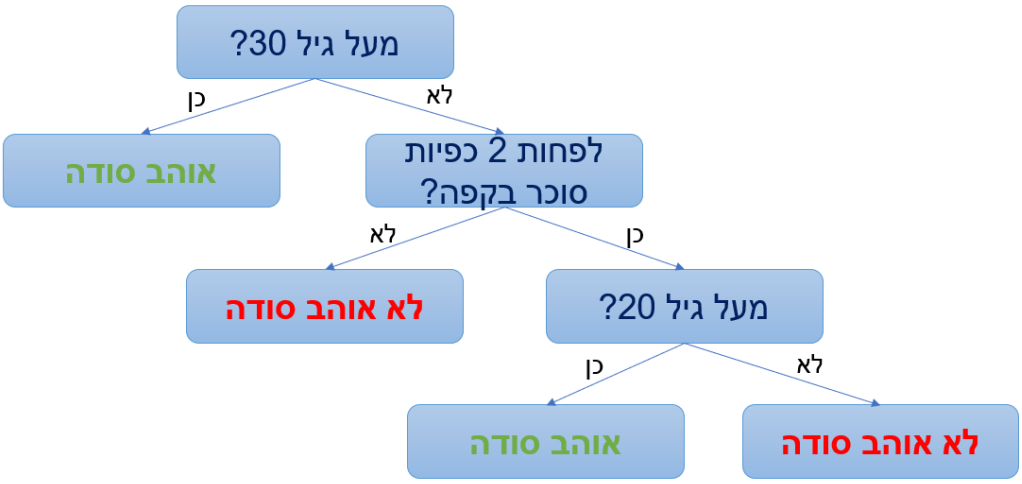
ב. נדרוש שמרחב ההיפותזות של עץ ההחלטה יהיה דו מימדי על מנת שנדרש ליותר מפיצול יחיד. כלומר נרצה שתי תכונות לפיהן נמיין את הדוגמאות בעזרת עצי החלטה, כאשר כל פיצול בעץ ההחלטה מתייחס לערך מסוים ביחס לציר אחר. ראינו בסעיף הקודם שכאשר נדרשנו לפיצול יחיד, הדרישה על המפריד הייתה ממימד יחיד, וכעת עבור 2 צירים שונים במישור נדרש ללפחות פיצול נוסף.

ג. בשונה מכלל הפיצול הסטנדרטי המתייחס לתכונה בודדת, על מנת שנוכל להפריד בין הדוגמאות בעזרת פיצול יחיד, נדרוש כי כל פיצול יתייחס למספר תכונות. במקרה שלנו, עבור הדאטה מהסוג המתואר נדרוש כי הפיצול יתייחס לשתי התכונות (x,y), ובכך נוכל לבנות עץ החלטה בעזרת פיצול יחיד. כלומר התנאי שייבדק יהיה עבור דוגמא (a,b) יהיה האם מתקיים b<ma+n או b>ma+n.

ד. נראה דוגמא לעץ החלטה DT לא גזום שקיים בו מסלול בעץ מהשורש לאחד העלים ובמסלול שני צמתים מפצלים לפי אותה תכונה.
נגדיר את קבוצת האימון הבאה:

	גיל	מספר כפיות סוכר בקפה	אוהב סודה
1	26	2	(+)
2	40	1	(+)
3	29	1	(-)
4	35	3	(+)
5	37	0	(+)
6	12	2	(-)

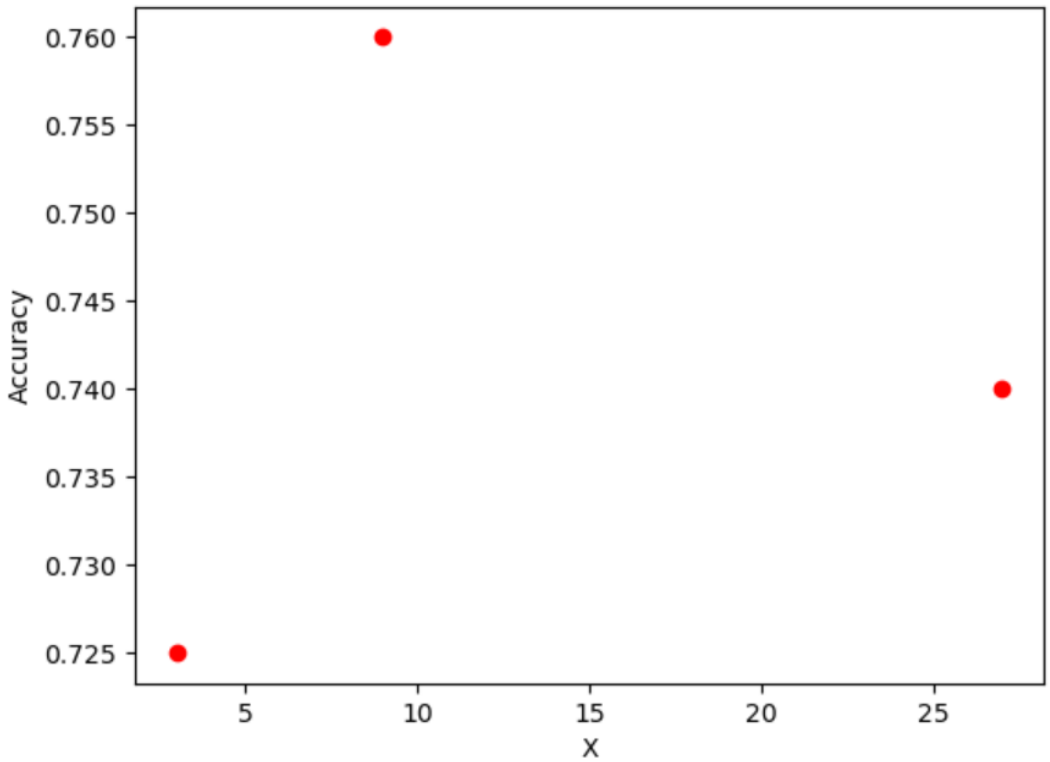
המסווג A, הבוחר תכונה הבאה לפיצול לפי התכונה שתביא לתוספת האינפורמציה הגדולה ביותר מיוצג ע"י העץ החלטה הנ"ל:



ניתן לראות כי אנו מפצלים פעמיים לפי אותה תכונה – גיל.

שאלה 3

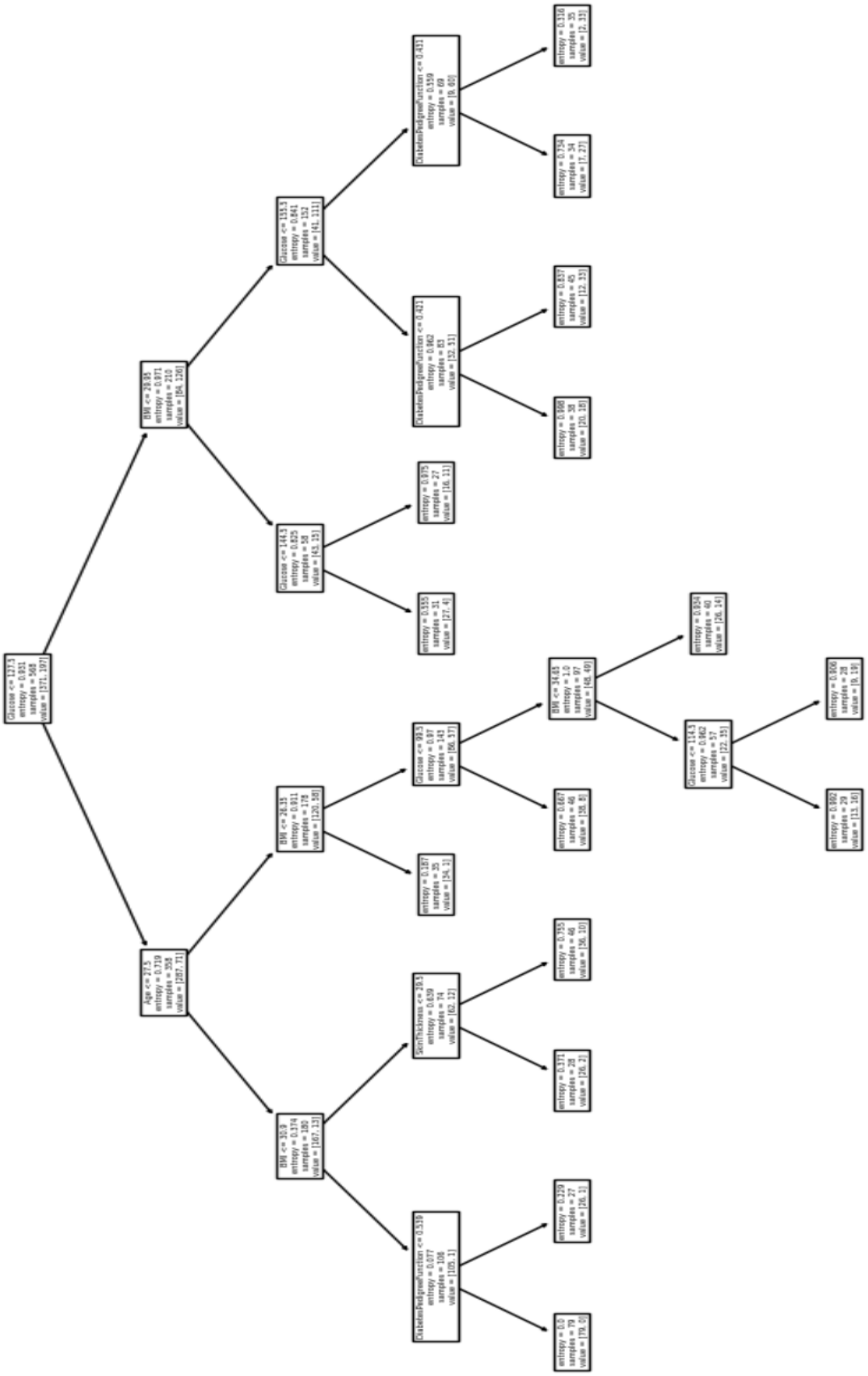
נציג גרף המראה את דיוק עצי ההחלטה כתלות בגודל הגיזום:



ניתן לראות כי הערך האופטימלי עבור X הוא 9. כמו כן נשים לב כי גיזום לפי X=27 נותן לנו דיוק טוב יותר מהמצב של X=3. ננסה להסביר את התוצאות כי מכיוון שכאשר אנחנו משתמשים בגיזום בו צומת הופך לעלה עם מינימום של 3 דוגמאות אין לנו אינדיקציה טובה לגבי סיווג העלה, כיוון שכמות הדוגמאות שהוא מייצג נמוכה מאוד. כלומר, במקרה הסביר בו יש רעש במרחב דוגמאות האימון, משקלו של רעש זה בבחירת סיווג הדוגמאות יהיה גבוה יחסית. לעומת זאת, כאשר אנחנו גוזמים לפי X=27 נקבל עץ החלטה פחות עמוק, כלומר נבדקות פחות תכונות ומתקבל עץ שאינו מראה על מגמה כלשהי ולכן לא מסווג בהתאם.

שאלה 4

נציג את מבנה עץ ההחלטה DT(x=27)



שאלה 5

א. עבור עץ A לא גזום נעריך כי ההסתברות ש-A יסווג את x כשלילית גדולה מ-P. מכיוון שהעץ לא גזום והדאטה אינו מאוזן, ניתן להסיק כי יתרחש overfitting על הדאטה ולכן יהיו טעויות רבות על דוגמאות המבחן. מכיוון שבדוגמאות האימון מספר הדוגמאות החיוביות קטן משמעותית מהשליליות, סביר להניח כי המסווג A לא יסווג דוגמאות חיוביות באופן טוב לעומת דוגמאות שליליות עבורן תהיה למסווג התאמה יותר טובה. לכן ניתן להסיק כי מספר ה-FN יהיה גדול יותר ממספר ה-FP, ולכן נבין כי ההסתברות לסווג את x כשלילי גדולה יותר מ-P (מספר הדוגמאות השליליות בפועל).

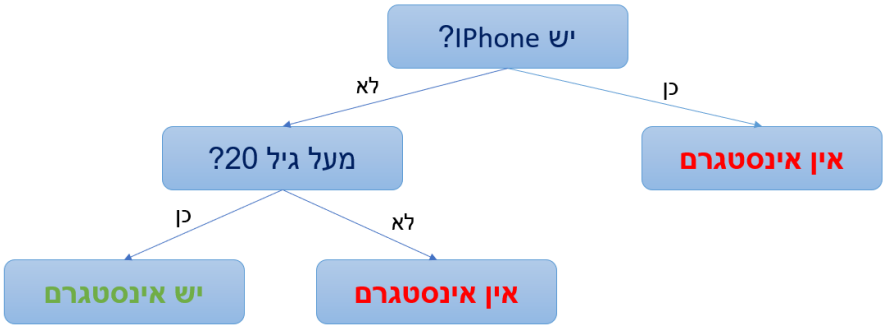
ב. עבור עץ B גזום נעריך כי ההסתברות ש-B יסווג את x כשלילית גם כן גדולה מ-P, מכיוון שהעץ גזום והדאטה אינו מאוזן. בגלל שהעלים מסווגים לפי רוב הדוגמאות שבעלה וקיימות יותר דוגמאות שליליות מחיוביות, סביר להניח כי קיימות דוגמאות חיוביות אשר ימצאו בעלים בהם רוב הדוגמאות שליליות ולכן העלה יסווג כשלילי. לפי הנחה זו, דוגמאות מבחן חיוביות עלולות להיות מסווגות כשליליות כיוון שיגיעו לעלה המסווג כשלילי, משמע מספר ה-FN יהיה גדול, ולכן ההסתברות לסווג את x כשלילי גדולה מ-p.

שאלה 6

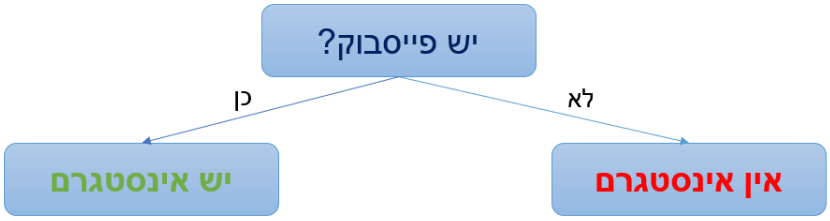
נפריך את הטענה ע"י דוגמא נגדית. נגדיר את הבעיה הבאה: אנו רוצים לסווג האם לבן אדם יש אינסטגרם או לא, ע"י 3 תכונות: האם יש לו פייסבוק, האם הוא בעל אייפון וגילו. תהא S קבוצת דאטה לא מאוזנת, כאשר 2 מהדוגמאות בה שליליות (כלומר אנשים ללא אינסטגרם) ודוגמא אחת חיובית. להלן קבוצת הדאטה הלא מאוזנת S:

	א	**ב**	**ג**
יש IPhone?	+	-	+
מעל גיל 20?	-	+	+
יש פייסבוק?	+	-	-
יש אינסטגרם	**-**	**-**	**+**

נבנה מסווג A לפי קבוצת האימון S, ע"י בחירת תכונה כל פעם המובילה לאנטרופיה הגבוהה ביותר:



נגדיר קבוצת אימון S', קבוצה מאוזנת עם דוגמא אחת חיובית ודוגמא אחת שלילית, ע"י זריקת דוגמא ב'. נבנה מסווג A' לפי קבוצת האימון S', הבוחר תכונה הבאה לפיצול לפי הגדלת האנטרופיה:



תהא x דוגמת מבחן שלילית, כלומר בן אדם שאין לו אינסטגרם, הוא מעל גיל 20 ואין לו פייסבוק ואייפון. לפי מסווג 'A - x יסווג כשלילי, כיוון שאין לו פייסבוק. אך לפי מסווג A - x יסווג כחיובי, כיוון שאינו בעל אייפון והוא מעל גיל 20.

שאלה 7

א. נצפה כי השגיאה בעץ הגזום B יניב שגיאה גדולה יותר ביחס לעץ שאינו גזום A. מכיוון שבעץ הגזום אנחנו מסווגים עלים לפי רוב הדוגמאות באותה צומת, סביר להניח כי מספר הדוגמאות החיוביות בקבוצת האימון שיוכנסו לעלה שלילי יהיה גדול, דבר זה יגרום לעלייה בסיווגים של ה-FN בקבוצת המבחן. לעומת זאת, בעץ שאינו גזום העץ עקבי כמובן, וייתכן כי דוגמאות מבחן יסווגו כ-FN עקב overfitting (היכול להיווצר מחוסר איזון בקבוצת הדאטה) אך אנחנו צופים כי קבוצה זו תהיה קטנה ביחס לקבוצה בעץ הגזום.
כיוון שמשוואת השגיאה נותנת משקל גדול משמעותית לשגיאת FN נסיק כי בהתאם להסבר לעיל השגיאה תהיה גדולה יותר בעץ הגזום.

ב. ערכי השגיאה על העצים DT1 ו- DT27 שיצאו לנו : Error1: 146, Error27: 155
ניתן לראות כי אכן כפי ששיערנו בסעיף הקודם, השגיאה בעץ הלא גזום קטנה ביחס לשגיאה בעץ הגזום.

שאלה 8

א. איזון הדאטה הנתון בתרגיל צפוי להקטין את שגיאת המבחן Error כיוון שיקטן כמות הFN על דוגמאות המבחן. בעץ גזום, אנו צופים כי כמות הFN תקטן כיוון שכמות דוגמאות האימון החיוביות והשליליות כעת שווה, ניתן להניח כי מספר העלים שמסווגים כחיובי או שלילי שווה יחסית בהמשך לאמור לעיל. לכן נניח כי הסיכוי לסווג דוגמא כ-FN עבור המקרה של דאטה מאוזן אינו גדול כמו שתיארנו בסעיף הקודם עבור המקרה עם דאטה לא מאוזן. מאותו הסבר של הסעיף הקודם, ערך הFN משפיע מאוד על פרמטר הError, ולכן עבור דאטה מאוזן בעץ הגזום הError יקטן.
עבור עץ שאינו גזום, שגיאת הFN גם כן תקטן כיוון שאנו צופים שה overfitting יקטן מהמקרה עבור דאטה לא מאוזן. לכן מאותו הסבר הError יקטן.

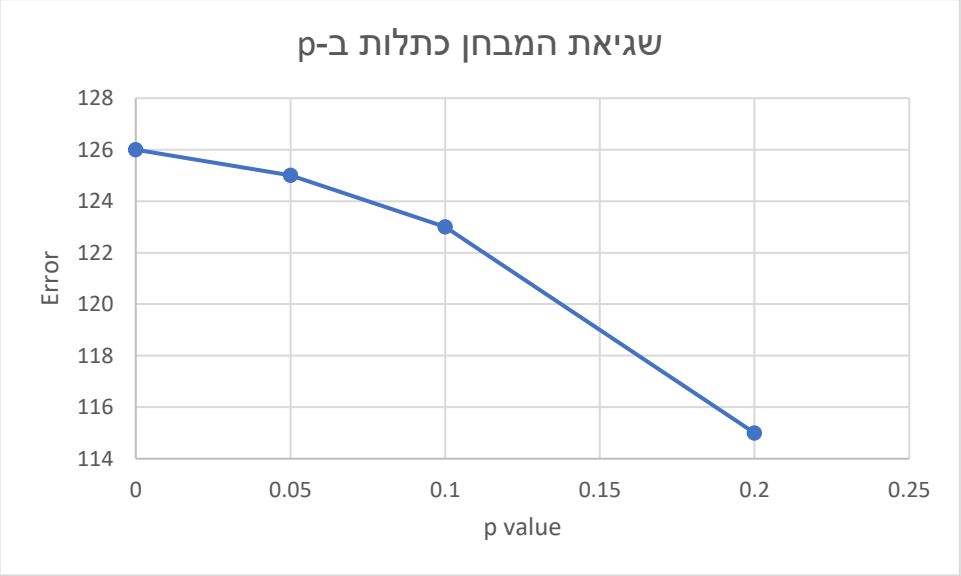
ב. תוצאת הסיווג של קבוצת המבחן שקיבלנו בפורמט f1 הינה:
[[46, 35]
[25, 94]]
כלומר מתקיים Error = 25*4 + 35 = 135
כלומר קיבלנו ערך שגיאה נמוך מערך השגיאה בעץ הלא גזום עם דאטה לא מאוזן מהסעיף הקודם (Error1: 146), מה שמתיישב עם תשובתנו בסעיף א.

שאלה 9

א. נשים לב כי השינוי המוצג בסעיף מתייחס רק לדוגמאות שמסווגות כשליליות לכן לא נדרוש תנאי כלשהו עבור דוגמאות שמסווגות שחיוביות, כלומר לא נדרוש תנאי על הערכים TP,FP. עבור הסתברות p=1 נקבל כי כל דוגמא שמקבל סיווג שלילי נשנה את סיווגה לחיובי. לכן נדרוש כי מספר הדוגמאות שמסווגות כשלילי ואכן שליליות יהיה קטן ממספר הדוגמאות המסווגות כשלילי ובעצם חיוביות, כלומר נדרוש TN < 4*FN וכך נקבל ששינינו יותר דוגמאות המסווגות לא נכון לסיווג נכון מאשר ההיפך ובכך הגדלנו את רמת הדיוק. בנוסף הקטנו בהכרח את ערך הError למרות שהגדלנו את ערך הFP (מה שסווג כ-TN לאחר הפרוטוקול הפך ל-FP), כיוון שאחרי ביצוע הפרוטוקול עם p=1, ערך הFN=0 והקטנתו הורידה יותר מאשר הגדלת הFP על הError.

ב. נשים לב כי שגיאת האימון נוצרת בעקבות הפיכת דוגמאות שמסווגות כשליליות לחיוביות בהסתברות P. כלומר מדובר בהתפלגות בינומית כאשר מתוך קבוצה בגודל F אנחנו בוחרים להפוך איבר לחיובי בהסתברות p לכן נקבל כי התוחלת עבור מקרה זה היא F*p.

ג. נציג גרף המתאר את שגיאת המבחן כתלות ב-p:



ניתן להבחין כי שגיאת המבחן יורדת ככל שערך ה-p גדל. תוצאה זו ניתנת להסבר כיוון שלפי הנאמר בסעיפים הקודמים, כאשר הדאטה אינו מאוזן ערך הFN המתקבל גבוה, ולכן הפרוטוקול המוצע משפר את ערך השגיאה כי הוא מקטין את הFN ע"י היפוך סיווגים שיוצאים שליליים לחיוביים. כפי שהוסבר בסעיפים הקודמים, ערך הFN משפיע רבות על פרמטר הError, והקטנתו גורמת להקטנת ערך השגיאה. כמו כן נציין כי היפוך הסיווגים משנה גם דוגמאות מ-TN ל-FP אבל השפעתו של שינוי זה קטנה יותר על ערך ה-error ביחס ל-FN.

כצפוי, ככל שערך הp גדל אנו מקטינים יותר את הFN ולכן קטנה שגיאת האימון.

שאלה 10

מכיוון שמשקל הטעות לגבי סיווג שלילי לא נכון גדול פי 4 ממשקל הטעות עבור סיווג חיובי לא נכון, נרצה כי בחירת סוג הסיווג של העלה תעשה בהתאם לאותו יחס. כלומר רק כאשר מספר הדוגמאות השליליות יהיה גדול או שווה לפי 4 מהדוגמאות החיוביות נבחר את העלה להיות שלילי, אחרת העלה יהיה חיובי. כלומר נבחר את $\alpha = 4$. באותו אופן, כאשר נחשב את כלל הפיצול, נרצה לתת משקל גדול יותר פי 4 עבור דוגמא חיובית כך שייבחר הפיצול המתאים ביחס לפונקציית השגיאה הנ"ל. לכן נבחר את $\delta = 0.8$ כאשר יתקיים באותו אופן $(1-\delta) = 0.2$.

שאלה 12

לפי האמור לעיל, משקל הטעות לגבי סיווג שלילי לא נכון גדול פי 4 מטעות עבור סיווג חיובי לא נכון, ולכן בעץ ה*DT2* נתנו לכל דוגמא חיוביות משקל גדול פי 4. ניתן לבצע מניפולציה ולהכפיל את כמות הדוגמאות החיוביות פי 4, ובכך נקבל שפעולת עץ *ID3* סטנדרטי גזום עם *x=9* פועל על הדאטה החדש כמו עץ ה*DT2* עם הדאטה הישן. כלומר רק כאשר מספר הדוגמאות השליליות יהיה גדול פי 4 ממספר הדוגמאות החיוביות בדאטה המקורי, נקבע את העלה להיות שלילי ואם לא נקבע לחיובי, ובכך נביא למינימום את השגיאה.

שאלה 13

לפי הנתון *q* מהדוגמאות הן חיוביות ו-*p* מהדוגמאות שליליות כאשר *q<<p* ומתקיים כי $q + p = 1$ כאשר נגדיר את מספר דוגמאות האימון להיות *n*. אלג' *KNN* מסווג דוגמאות מבחן לפי סיווג רוב ה-*k* שכנים הקרובים ביותר לדוגמת המבחן הנבדקת. לכן, עבור $k \geq 2q + 1$ נקבל כי תמיד יהיה רוב של דוגמאות שליליות בשכנים של *x*, לכן המסווג *C* תמיד יקבע את *x* להיות שלילי, כלומר יקבע אותו לשלילי בהסתברות 1. עבור $k < 2q + 1$ נקבל כי ההסתברות שהמסווג *C* יקבע את *x* להיות שלילי היא התפלגות בינומית כלומר $P(x =' -') = \binom{k/2}{n} * q^{\frac{k}{2}} * p^{n-\binom{k}{2}}$.

שאלה 14

תוצאת הסיווג של *KNN1* המתקבלת על קבוצת המבחן בפורמט f1 הינה:

[[37 18]

[[34 111]

לכן השגיאה Error של מסווג זה הינה: Error = 34*4 + 18 = 154

שאלה 16

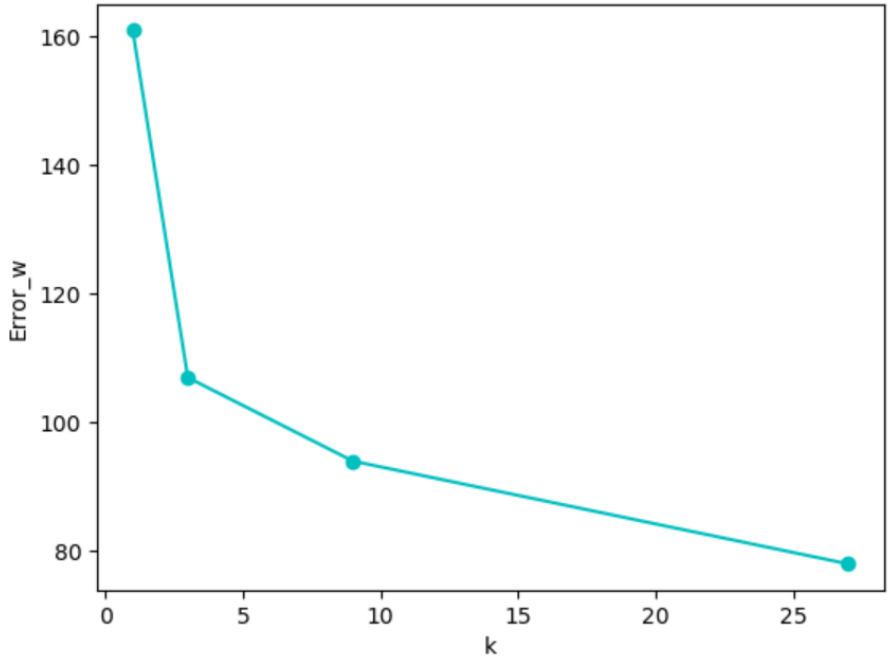
תוצאת הסיווג של *KNN2* המתקבלת על קבוצת המבחן בפורמט *f1* היא:

[[74 66]]

[[55 5]]

שאלה 17

נציג את הגרף המתקבל עבור מסווג המשתמש בכלל ההחלטה הנתון המציג את $Error_w$ כתלות ב-k:



נשים לב כי ככל ש-K גדל כלומר מסווגים לפי מספר גדול יותר של שכנים עבור כלל ההחלטה הנתון כך השגיאה יורדת. עבור $k=1$, אנחנו למעשה מסווגים לפי השכן הקרוב ביותר לכל דוגמת מבחן, אזי אין משמעות להוספת המשקל בכלל ההחלטה החדש. בנוסף כיוון שהדאטה אינו מאוזן ומספר הדוגמאות השליליות גדול משמעותית ממספר הדוגמאות החיוביות קיים סיכוי גדול יותר שנסווג דוגמת מבחן כשלילית ללא תלות בסיווג האמיתי שלה. לכן נסיק כי נקבל מספר FN גבוה וכפי שהסברנו בסעיפים הקודמים נקבל ערך $Error_w$ גבוה, כפי שניתן לראות בגרף המצורף.

אמנם כלל ההחלטה מעניק לשכן חיובי משקל גבוה פי 4 ממשקל המוענק לשכן שלילי אך כיוון שהדאטה אינו מאוזן וקיים מספר רב יותר של דוגמאות שליליות עדיין ההסתברות לסווג דוגמא כחיובית אינה גדולה. ככל שנסתכל על יותר שכנים אמנם נצטרך מספר גדול יותר של שכנים חיובים על מנת לסווג את הדוגמא כחיובית אבל מספר השכנים הכולל אותם אנחנו בודקים גדל באופן מהיר יותר. לפי ההסבר שמובא לעיל נקבל ככל שערך ה-K יותר קטן נקבל ערך FN יותר גדול ולכן ערך $Error_w$ יותר גדול, דבר התואם את הממצאים בגרף.