חלק א

שאלה 1

<u>סעיף א'</u>

כדי שזה יקרה הישר צריך להיות מקביל לאחד הצירים. במקרה הזה נוכל לפצל את כל הדוגמאות על ידי בדיקה יחידה, כלומר פיצול אחד של העץ.

במקרה של קו מקביל לציר X, נדרוש כי $m = 0$ ונקבל את הישר $y = n$. בכל דוגמה שערך ה-y שלה מעל הקו (כלומר גדול מ-n) היא חיובית, והשאר שליליות. זו בדיקה יחידה שניתן לבצע בפיצול אחד.

לא ניתן לממש ישר המקביל לציר Y עם הפרמטרים הנ"ל (ישר זה הוא מהצורה $x = const$).

<u>סעיף ב'</u>

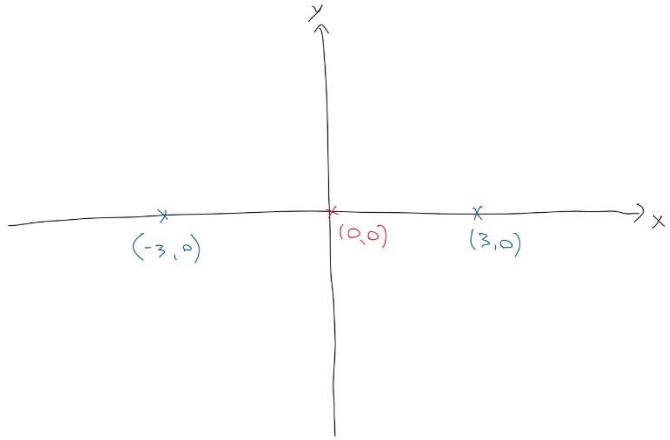
כל פיצול של עץ ההחלטה הוא ישר שמקביל לאחד הצירים, מרחב ההיפותזות הוא אוסף כל החלוקות של המרחב לאיחוד של מלבנים מקבילים לצירים. במקרה הכללי דרושה חלוקה של יותר מפיצול בודד כדי להפריד בין כל הדוגמאות.

<u>סעיף ג'</u>

נחליף את כלל הפיצול. במקום להיות חלוקה של דוגמאות לפי ערך תכונה יחידה, תהיה חלוקה של הדוגמאות לפי משוואת קו ישר שנובעת מהכנסת ערכי שתי התכונות. למשל עבור ישר $y = mx + n$ נפריד את הדוגמאות x_0, y_0 לפי הערך של הישר עליהן, כלומר האם $y_0 > mx_0 + n$ כדי לקבוע האם הוא מעל הישר (או מתחתיו אחרת). הערכים של m ו-n מוגדרים כחלק מכלל הפיצול.

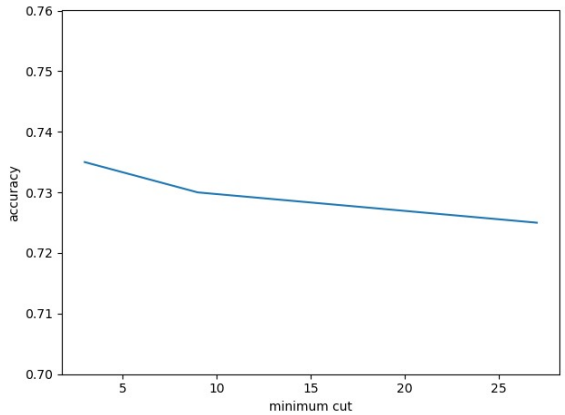
<u>סעיף ד'</u>

יכול להיות. נסתכל על המדגם הבא (אדום=שלילי, כחול=חיובי):



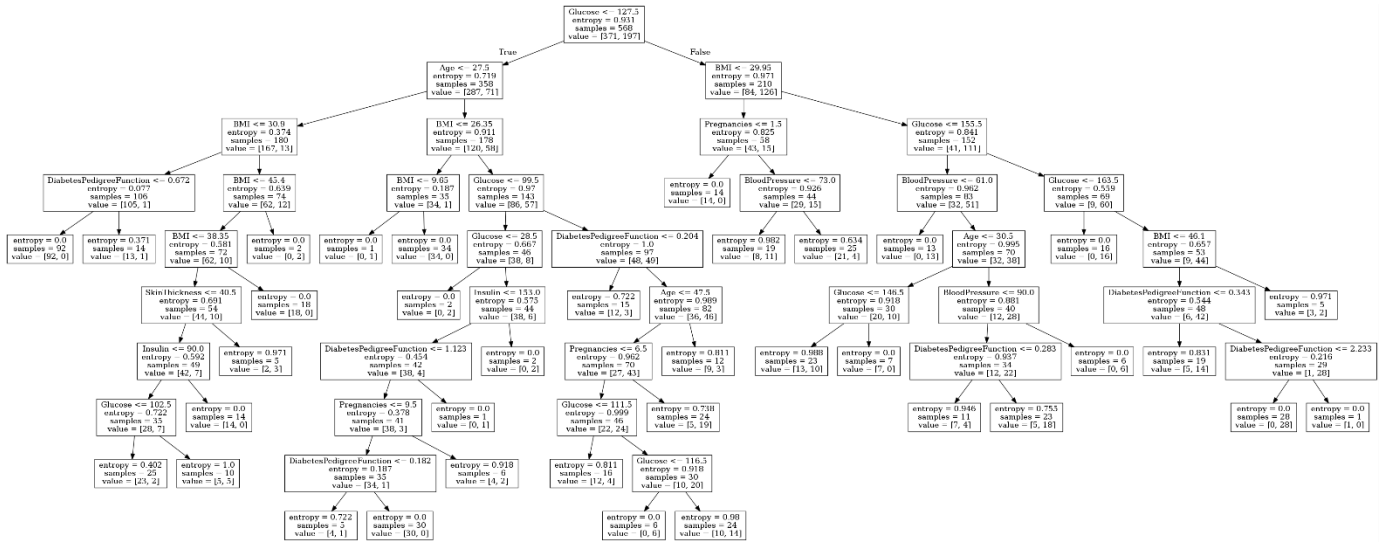
בתור התחלה נבחין כי הפרדה לפי *y* לא תשנה כלום, לכן הבחירה הראשונה בהכרח תהיה לפי ערך *x*. לכל ערך שנבחר להפריד על פיו נקבל כי בצד אחד יהיו דוגמא חיובית ושלילית, ובאחר חיובית (עלה). כעת נותר להפריד את הדוגמאות החיובית והשלילית. בדומה לקודם, הפרדה לפי ערך *y* לא תיתן כלום, ונצטרך להפרידם על פי ערך *x*. סך הכל, בעץ שנוצר הפרדנו פעמיים על פי אותה התכונה.

שאלה 3



נבחין כי בחיתוך גס יותר יש ירידה קלה (מאוד) בדיוק, כנראה מתוצאה של ויתור על פיצולים משמעותיים. עם זאת, כשמעלים את סף החיתוך מעבר לכך אין שינוי משמעותי. כנראה שפיצול ל-3 תופס טוב את החלוקה של הדוגמאות (מתמודד עם ה-*Overfitting* שיכל להיווצר לפני הגיזום).

שאלה 4



חלק ב'

שאלה 5

<u>סעיף א'</u>

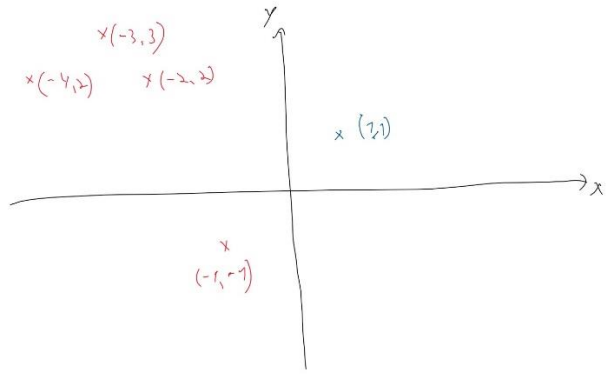
אנחנו מעריכים שההסתברות תהיה בערך זהה ל-p. זאת מכיוון שזה בערך היחס בין הדוגמאות, ולכן דוגמא אקראית כנראה תתאים ליחס זה.

<u>סעיף ב'</u>

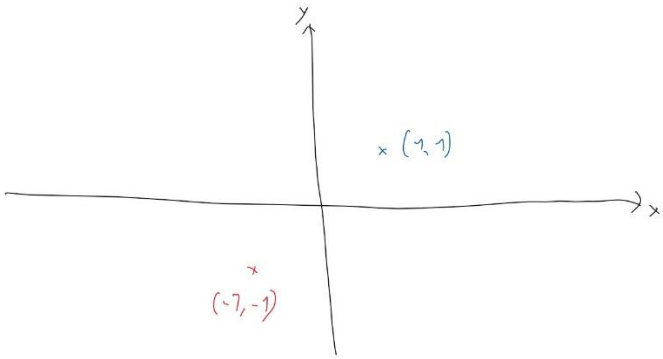
אנחנו מעריכים שההסתברות תהיה גדולה מ-p. מכיוון שכמות הדוגמאות השליליות גדולה ממש מהדוגמאות החיוביות, לכל עלה שנגזם יש הסתברות גבוהה יותר לצאת שלילי (יוכרע על ידי בחירת הרוב), ובכך לסמן דוגמאות חיוביות כשליליות.

שאלה 6

דוגמה נגדית, נניח כי קיבלנו את המדגם הנ"ל (אדום = שלילי, כחול = חיובי)



איזנו את הדוגמאות באופן אקראי וקיבלנו את המדגם הבא:



נסכל על דוגמת מבחן t שנופלת ב-$(1,-1)$. בהתחלה לפני הדילול, משתלם לנו (מבחינת *information gain*) לבנות את A כך שיפצל לפי ציר y, כי זה מפריד בבדיקה אחת את כל הדוגמאות. לכן, A יסווג את t כחיובית.

עם זאת, לאחר הדילול, יכול להיות שנבחר לבנות את A' כך שיפריד לפי ציר x (שכן זו דוגמה לבנייה שמביאה *information gain* מקסימלי). בבנייה זו, הדוגמה t דווקא תסווג כשלילית, בהפרכה לטענה.

שאלה 7

<u>סעיף א'</u>

העץ הלא-גזום יביא שגיאה קטנה יותר. זאת מכיוון שעץ גזום מעלה את הסיכוי לקבל *Negative*, ובפרט *False Negative*. מכיוון שהחישוב של $Error_w$ מושפע יותר מ-*False Negative* נקבל שגיאה גדולה יותר בעץ הגזום.

<u>סעיף ב'</u>

קיבלנו כי השגיאה בעץ הגזום היא 145, בעוד שבעץ הלא גזום היא 129. זה מתיישב עם ההשערה שלנו מהסעיף הקודם.

שאלה 8

<u>סעיף א'</u>

מכיוון שאנחנו מקטינים את כמות הדוגמאות השליליות, אנחנו מקטינים את כמות ה-*False Negatives* (וכנראה גם את כמות ה-*True Negative*, כלומר מעלים *False Positive*). מכיוון שהשגיאה מושפעת יותר מה-*False Negatives* אנחנו בפועל מקטינים את השגיאה.

<u>סעיף ב'</u>

כן, התשובה מתיישבת עם הסעיף הקודם (ערך השגיאה בדאטה המאוזן הוא 128 לעומת 129 בלא מאוזן). ההפרשים קטנים ועלולים להשתנות עם *Seed* שונה (נובע מכך שיש יחסית מאט דאטה וחוסר האיזון לא מאוד קיצוני).

שאלה 9

<u>סעיף א'</u>

על ידי המרה של p להיות 1, נהפוך את הערכים של *true negative* להיות *false positive*. כלומר, משתלם להגדיר את $p = 1$ אם מתקיים:

$$4 * FN + FP > TN + FP$$
$$4 * FN > TN$$

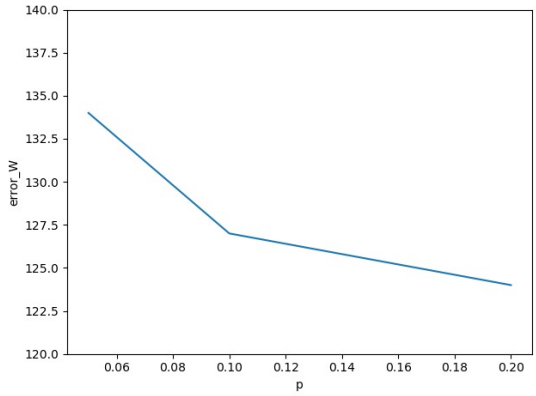
<u>סעיף ב'</u>

$$P(T) = \frac{|T|}{|T| + |F|}$$

$$P(F) = \frac{|F|}{|T| + |F|}$$

$$P(PredictTrue) = P(T) + p * P(F)$$
$$P(PredictFalse) = (1 - p) * P(F)$$
$$FN = P(T) * P(PredictFalse)$$
$$FP = P(F) * P(PredictTrue)$$
$$Error_w = 4 * FN + FP = 4 * \frac{|T|}{|T| + |F|} * (1 - p) * \frac{|F|}{|T| + |F|} + \frac{|F|}{|T| + |F|} * \left(\frac{|T|}{|T| + |F|} + p * \frac{|F|}{|T| + |F|}\right)$$

סעיף ג'



בעץ הלא גזום קיבלנו שהשגיאה הינה 129. עם מקדם נמוך לא נצפה הבדל משמעותי, וזה כנראה נובע מהאקראיות (שבמקרה הזה קטנה מכדי שתורגש). כאשר המקדם עולה יש ירידה בשגיאה, כבר ב-*0.1* יש הבדל לטובת השיטה (אם כי גם כאן ניתן לטעון שזה נובע מהאקראיות). כשהמקדם *p* על *0.2* נראה כי השגיאה ירדה ל-124, שזה הבדל קצת יותר משמעותי.

נסתכל על המגמה הכללית. השיטה להפיכת התוצאות היא אקראית, לכן היא הופכת באקראי שלילי לחיובי, ועלולה להרוס או לתקן דוגמא שסווגה כשלילית. בגלל ש-*fn* חמור פי 4 מ-*fp*, בממוצע השגיאה יורדת.

שאלה 10

הערכים הם $\alpha = 4, \delta = \frac{4}{5}$

בחרנו בערך של אלפא להיות 4 מכיוון שהשווי של כל *FN* הוא פי 4 מ-*FP* ולכן נרצה יחס זהה בקבלת החלטה האם לתייג את העלה כחיובי או שלילי. כלומר, עדיף לטעות שלוש פעמים בסיווג לא נכון כחיובי מאשר פעם בסיווג לא נכון כשלילי.

בחרנו ערך $\frac{4}{5}$ לדלתא מכיוון שאנחנו נותנים ערך גדול פי 4 לכל דוגמא חיובית ולכן היא נותנת לנו פי 4 יותר מידע מדוגמא שלילית. בהצבת המספר הזה נקבל יחס של $\frac{1}{5}$ לעומת $\frac{4}{5}$ שהוא בדיוק 1:4.

חלק ג'

שאלה 13

$$P(N) = \sum_{i=\left\lceil \frac{k}{2} \right\rceil}^{k} \binom{k}{i} p^i (1-p)^{k-i}$$

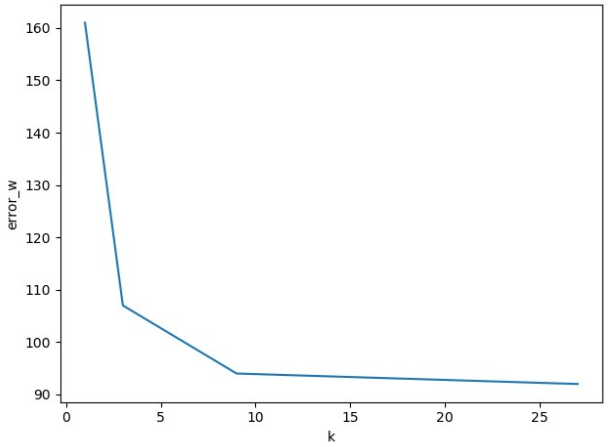
הסבר:

על מנת שדוגמא תסווג כשלילי נצטרך שלפחות מחצית מהשכנים הקרובים יהיו שליליים, מכיוון שהדוגמאות מתפלגות ביחס של $p, 1-p$ נקבל שההסתברות לקבל שכן קרוב שלילי יחיד היא p. לכן נסכום על כל האפשרויות שבהם לפחות מחצית מהשכנים הם שליליים (סכימה על מבחן בינומי).

שאלה 14

השגיאה של מסווג זה היא 154

שאלה 17



קל להבחין כי השגיאה יורדת ככל שערכו של *k* עולה. היחס בין החיוביים לשליליים הוא 197 ל-371, מכאן שבערכי *k* נמוכים יש סבירות גבוהה יותר לתפוס דוגמאות שליליות בלבד. במקרים אלה המשקל שהוגדר עבור החיוביים אינו פקטור בהחלטה, ולכן טעויות הסיווג של חיובי לשלילי (שיוצרות *FN*) יהיו נפוצות יותר, ולכן ערכו של $error_w$ יעלה.

כאשר ערכו של *k* עולה, אנו מסתכלים על מספר רב יותר של שכנים, ולכן ניתן להניח שעבור דוגמא כלשהי נקבל יחס שכנים שמתאים ליחס הדוגמאות. מכיוון שיחס החיוביים-שליליים הוא בערך אחד לשניים, בממוצע שני שליש מהדוגמאות יהיו שליליות, ושליש חיוביות. במקרה זה, מכיוון שלדוגמא חיובית יש משקל גדול פי 4 משלילית, אנחנו נעדיף בממוצע להכריע לטובת הדוגמאות החיוביות, ויהיו פחות *FN*, מה שמקטין את השגיאה.