שאלה 1

1. 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 שהישר יהיה מאונך לאחד הצירים, כלומר שיתקיים m=0 או n=0.
2. 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.
3. 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.

שאלה 7

1. 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.

1. 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

1. 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.

1. תוצאת הסיווג של קבוצת המבחן שקיבלנו בפורמט f1 הינה:

[[34 95]

[48 23]]

כלומר מתקיים Error = 23\*4 + 34 = 126

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

1. 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 0 על מנת שלא ייווצר מצב בו אנחנו משנים סיווג נכון לסיווג לא נכון. כלומר נדרוש כי TN=0. מעבר לכך לא נדרוש תנאי נוסף על הערך FN כיוון שאנחנו יודעים שזהו סיווג שלילי לא נכון ולאחר הטלת המטבע יהפוך לסיווג נכון בוודאות.
2. 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.
3. 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 דוגמאות שליליות גדול פי 4 ממשקל הטעות עבור דוגמאות חיוביות, נרצה כי בחירת סוג הסיווג של העלה תעשה בהתאם לאותו יחס. כלומר אם מספר הדוגמאות השליליות יהיה גדול פי 4 מהדוגמאות החיוביות נבחר את העלה להיות שלילי, אחרת העלה יהיה חיובי. כלומר נבחר את *. 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 כאשר יתקיים באותו אופן .*