**מבוא לבינה מלאכותית**

**תרגיל בית 3**

מגישים:

אביב כספי 311136691

יקיר יהודה 205710528

**חלק א'**

1. א. עץ החלטה מפריד בין דוגמאות על ידי קווים ישירים המקבילים לצירים, לכן נדרוש m=0 ו ממשי כלשהו.

ב. עץ החלטה מפריד בכל איטרציה על ידי מציאת קו ישר המקביל לאחד הצירים, המפריד בין הדוגמאות, כאשר העקום המתקבל, בסיום בניית העץ, יהיה איחוד של כל הקווים שמצאנו. כלומר מרחב ההיפותזות מכיל עקומים המורכבים מישרים מקבילים לצירים.

כאשר כל פיצול מוסיף לנו קו נוסף , לכן כל מקרה בו הדוגמאות לא ניתנות להפרדה על ידי קו ישר מקביל לצירים, נצטרך לבצע יותר מפיצול אחד.

ג. נשנה את כלל הפיצול שיכיל יחס בין הפיצ'רים של כל אובייקט. לדוגמא אם ניתן להפריד בין הדוגמאות על ידי ישר מהצורה y=mx+n , כלל הפיצול שלנו יהיה y<mx+n , כאשר אם התשובה היא כן, נסווג בצורה אחת, ואם התשובה היא לא נסווג בצורה שנייה.

ד. כן, כאשר התכונות רציפות, ייתכן כי בין שתי דוגמאות חיוביות נמצאת דוגמא שלילית ונפריד על ידי שימוש באותה תכונה פעמיים . להוסיף דוגמא עם גרף

1. ש
2. 3
3. 4

**חלק ב'**

1. נעריך כי ההסתברות שהמסווג שלנו יסווג את X כשלילית בהסתברות קרובה ל-p.

A עץ ID3 לא גזום, כלומר הוא ממשיך לפצל בזמן האימון, עד אשר כל העלים הומוגניים.  
בגלל שנתון כי יש הרבה יותר דוגמאות שליליות מחיוביות, נוכל להסיק כי יהיו הרבה יותר מסלולים לעלים שליליים מאשר חיוביים. כאשר נוכל להסיק כי ההסתברות לסווג את X כשלילית הינה בערך p בגלל שהדוגמאות בלתי תלויות (כלומר, יש הסתברות זהה להגיע לכל עלה בעץ).

1. כאשר יש הרבה יותר דוגמאות "-" מאשר "+" , אם נגזום את העץ, ייתכן כי נגיע למצב

בו בכל העלים יש מספר גדול יותר של דוגמאות "-" מאשר "+", לכן עץ ההחלטה שלנו יסווג כל דוגמא כשלילית בהסתברות גבוהה מאד (גדולה מ-p). להוסיף ניסוח בדומה לסעיף קודם עם הסתברויות (הסיכוי שרוב העלים יהיו שליליים)

1. לחשוב על דוגמא נגדית
2. א. כמו שראינו בסעיף 5 , כאשר הדאטה שלנו לא מאוזן, ישנה הסתברות גבוהה יותר שדוגמא כלשהי תסווג בסיווג אשר מופיע יותר בדוגמאות האימון כאשר אנחנו גוזמים את העץ. כלומר בדוגמא שלנו, יש יותר דוגמאות שליליות מאשר חיוביות, לכן בעץ גזום ההסתברות שנסווג דוגמא כשלילית גבוהה יותר (ההסתברות שבעלים יהיו יותר דוגמאות שליליות מאשר חיוביות). כלומר בהתאם ישנו סיכוי גדול יותר כי נגדיל את מספר הFN (ונקטין את FP) ולכן השגיאה תגדל.
3. א. כמו שראינו בסעיפים הקודמים, ההסתברות לסווג דוגמא כשלילית כאשר הדאטה שלנו לא מאוזן הינה גדולה מאד ולכן נסווג הרבה דוגמאות כשליליות ומכך יעלה ערך השגיאה . כאשר נאזן את הדאטה שלנו כך שיהיו מספר דוגמאות שווה לכל סיווג, נעלה את ההסתברות (כאשר הדוגמאות בלתי תלויות אחת בשנייה) לסווג את הדוגמא כחיוביות, כלומר ערך השגיאה יקטן.
4. א. על מנת שיהיה כדאי להגדיר p=1 , נרצה כי השגיאה שלנו אחרי ההגדרה תהיה קטנה יותר.

נסמן TP,FP,TN,FN את הערכים של המודל שלנו לפני הגדרת p וב- TP',FP',TN',FN' את הערכים אחרי הגדרת p .

נרצה שיתקיים :

4FN + FP >4FN' + FP'

נשים לב כי כאשר p=1 , נסווג כל דוגמא כחיובית, כלומר מתקיים :

FN'=0, TN'=0, FP'=TN+FP

כאשר השגיאה החיובית שלנו כרגע היא כל הדוגמאות השליליות הקיימות.

כלומר :

במקרה זה יהיה לנו עדיף להגדיר p=1.

ב. נרצה למצוא את התוחלת של השגיאה שלנו.  
נשים לב כי שגיאת האימון של עץ DT1 מבלי להשתמש בp הינה 0, מפני שממשיכים לפצל את הצמתים עד אשר מגיעים לעלים הומוגניים, כלומר לכל דוגמא מקבוצת האימון נסווג אותה בצורה נכונה (עץ DT1 הינו עץ עקבי).

כלומר, לאחר שימוש בשיטה המוגדרת, יתכן כי נשנה רק דוגמאות שמסווגות כשליליות לחיוביות. כלומר, לא יתכן כי נסווג דוגמת אימון חיובית כשלילית ולכן תמיד יתקיים FN=0.

ניתן לראות כי FP מתפלג בצורה בינומית, כאשר מספר הניסויים שלנו הוא כגודל קבוצת הדוגמאות השליליות - |F| , ונגדיר הצלחה כשינוי של סיווג שלילי לחיובי, בהסתברות p.

*לפי תוחלת של התפלגות בינומית נקבל כי :*

ג. להוסיף גרף

10. הסתברות שדוגמא x תסווג לפני השינוי (בהנחה כי מספר הדוגמאות בכל אלה מתפלג באותה צורה כמו שמתפלג הדוגמאות אימון שלנו) :

*הסתברות אחרי השינוי :*

*ולכן תוחלת מספר הדוגמאות שיסווגו בצורה שגוייה הוא :*

*כעת נרצה כי השגיאה שלנו תהיה קטנה יותר כאשר נשתמש באלפא :*

*לכן נדרוש :*

*13. כעת נתון לנו כי מספר הדוגמאות השליליות גדול בהרבה ממספר הדוגמאות החיוביות בקבוצת האימון.  
לכן, נוכל להסיק כי ככל ש k יגדל, כך גם ההסתברות שדוגמת מבחן כלשהי, תגדל בהתאם. כאשר עבור k ששווה לפי 2 מגודל הקבוצה החיובית, נקבל כי ההסתברות לסיווג דוגמת מבחן כשלילית היא 1. (מספר הדוגמאות החיוביות הקרובות יהיה במקסימום גודל הקבוצה החיובית, לכן תמיד יהיה יותר דוגמאות שליליות קרובות)  
כאשר k קטן (שווה ל1), בגלל אי תלות בין הדוגמאות, נוכל להסיק כי ישנה הסתברות של p שדוגמת מבחן x כלשהי, תהיה קרובה יותר לדוגמת אימון שלילית, ולכן תסווג כשלילית בהסתברות שווה לp.*

*כאשר k קטן אך גדול מ1, ההסתברות שדוגמת מבחן תסווג כשלילית גדולה מ-p (מפני שההסתברות שדוגמת המבחן תהיה קרובה ליותר מדוגמת מבחן אחת חיובית הינה קטנה יותר מאשר q ).*