מבוא לבינה מלאכותית - תרגיל בית 3

מגיש: **אלמוג צברי, 312433576.**

תאריך הגשה: 26/01/2021

מספר גיליון: **3**

שאלה 1

לאחר אימון המסווג על קבוצת האימון ובדיקה על ידי קבוצת הבוחן הדיוק המתקבל הינו 0.9469026548672567.

```
C:\Users\Almog\miniconda3\envs\intro-to-ai-hw3\python.exe
0.9469026548672567

Process finished with exit code 0
```

איור 1: דיוק עבור קבוצת הבוחן

שאלה 2

הטענה נכונה. תחילה נשים לב כי נורמליזציית $MinMax$ היא למעשה הטרנספורמציה הבאה:

$$g(v) = \frac{v - v_{min}}{v_{max} - v_{min}} = \frac{1}{v_{max} - v_{min}} \cdot v - \frac{v_{min}}{v_{max} - v_{min}}$$

כאשר v_{max} ו־ v_{min} הם קבועים הנובעים מהדאטה המתקבל. נגזור ונקבל $g'(v) = \frac{1}{v_{max} - v_{min}} > 0$. כלומר נורמליזציית $MinMax$ היא **טרנספורמציה מונוטונית עולה ממש** ולפיכך מתקיים כי $v_1 > v_2$ אם ורק אם $g(v_1) > g(v_2)$.
מכאן נשים לב כי אם $threshold = \frac{v_1 + v_2}{2}$ ואם נסמן ב־ $(threshold)'$ את ערך הסף המתקבל מהערכים המנורמלים אז מתקיים:

$$(threshold)' = \frac{g(v_1) + g(v_2)}{2} = \frac{\frac{v_1 - v_{min}}{v_{max} - v_{min}} + \frac{v_2 - v_{min}}{v_{max} - v_{min}}}{2} = \frac{\frac{v_1 + v_2}{2} - v_{min}}{v_{max} - v_{min}} = \frac{theshold - v_{min}}{v_{max} - v_{min}} = g(threshold)$$

כלומר ערך הסף המתקבל מהערכים שעברו נרמול זהה להפעלת הנרמול על ערך הסף, ולכן מתקיים:

$$v_1 < threshold < v_2 \iff g(v_1) < g(threshold) < g(v_2)$$

כעת, יהי T העץ המתקבל על ידי הפעלת $ID3$ על הדאטה הלא מנורמל ויהי T' העץ המתקבל על ידי הפעלת $ID3$ על הדאטה המנורמל. נראה באינדוקציה על עומק העץ כי T' זהה ל־ T כלומר יבחר בכל צומת את אותו מאפיין המתאים ב־ T ואת אותו ערך סף (עד כדי נרמול):
<u>בסיס:</u> עומק 0, כלומר u, שהוא הצומת בעומק זה, הינו השורש: אם כמות הדוגמאות המתקבלת בצומת זו היא 0 אז לא תבחר תכונה ולכן הטענה נכונה באופן ריק. אחרת, כמות הדוגמאות גדולה מ־ 0: תהי f התכונה שנבחרה בצומת זה בעץ T על פי פונקציית בחירת המאפיינים ויהיו v_1, v_2 שתי ערכים שונים שהתכונה f יכולה לקבל כך שערך הסף $threshold$ הוא הממוצע שלהם. נשים לב כי כיוון שמדובר בשורש אז הדוגמאות שמגיעות לצומת זה זהות בשני העצים. עוד נשים לב כי ה־ $threshold$ המתאים בעץ T' הוא למעשה הפעלת הטרנספורמציה על ה־ $threshold$ ב־ T, ומשיקולים שהוסברו קודם נסיק כי אותן דוגמאות שהיו קטנות מ־ $threshold$ ב־ T יהיו קטנות מ־ $(threshold)'$ ב־ T' ולכן פיצול הדוגמאות בעצים, כאשר רוצים לחשב את תוספת המידע, יהיה זהה (ולכן גם ההסתברויות). כמו כן, מכיוון שתוספת המידע תלויה גם בסיווג, **שאינו עובר נירמול**, נובע כי תוספת המידע תהיה זהה ולכן אותה תכונה f תבחר עם $(threshold)' = g(threshold)$.
<u>צעד:</u> נניח כי לכל צומת עד עומק k העץ T זהה לעץ T'. יהי u צומת כלשהו בעומק $k+1$ ב־ T'. מהנחת האינדוקציה נובע שעד צומת האב העץ T זהה לעץ T' עד כדי נרמול של ערך הסף, ולכן הדוגמאות המגיעות ל־ u זהות לאלו שמגיעות למקבילו בעץ T. לפיכך נוכל לחזור על שהוסבר בבסיס ולהסיק כי המאפיין שייבחר בשני העצים זהה עם ערכי סף זהים עד כדי נרמול, ובכך סיימנו את ההוכחה.
<u>מסקנה:</u> נרמול של הדאטה אינו משפיע על בניית העץ ולכן גם לא משפיע על הדיוק על קבוצת הבוחן (כמובן בתנאי שקבוצת הבוחן עוברת את **אותה טרנספורמציה מונוטונית עולה ממש**).

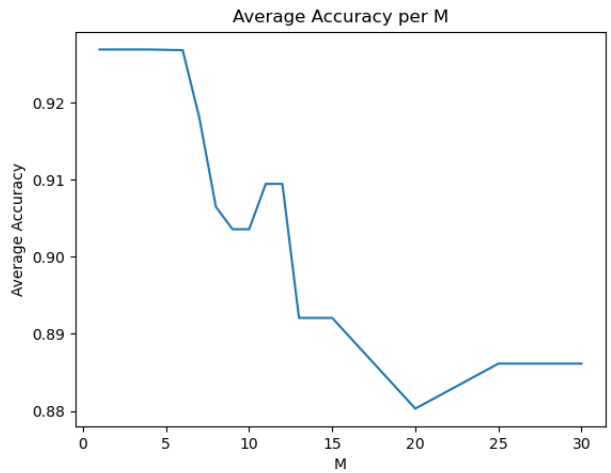
* הערה: ניתן אישור מהמתרגל האחראי על התרגיל לחרוג במעט מ־ 20 שורות *

שאלה 3

סעיף 1

גיזום של עץ החלטה מאפשר לקבל עץ סיווג קטן יותר, אשר בתקווה יוכל למזער את תופעת ה־ $Over-Fitting$. מטרתו של גיזום עץ החלטה היא למנוע התאמת יתר של המסווג לדאטה עליו אומן (סט האימון), כלומר תיתכן שגיאה גדולה יותר עבור דאטה זה, אך נוכל לגרום לכך שכל החלטה, אשר מתבצעת בעלים, תסתמך על יותר דאטה ובכך אולי תגרום לדיוק גבוה יותר על **דוגמה חדשה שלא נראתה קודם לכן**.

סעיף 3



איור 2: דיוק ממוצע כתלות בערך של M

ניתן לראות שהגרף הנ"ל הינו במגמה יורדת, כלומר **באופן כללי** ככל ש־ M יותר קטן כך הדיוק הממוצע גבוה יותר. בפרט ניתן לראות כי הדיוק הממוצע המקסימלי מתקבל עבור $M = 1$ והוא 0.9469026548672567. מהתבוננות בגרף אנו רואים כי הגיזום למעשה רק פוגע ביכולת ההכללה של המסווג, אולי מכיוון שסט האימון קטן יחסית (כ־ 350 דוגמאות בלבד) ולכן העץ "קטן מספיק" גם ללא גיזום.

סעיף 4

כפי שניתן לראות בסעיף הקודם, הערך של M הנותן דיוק מקסימלי הינו $M = 1$, כלומר ללא גיזום כלל, ולכן הדיוק ישאר 0.9469026548672567 כפי שהיה מקודם.

שאלה 4

סעיף 1

עבור הרצת אלגוריתם $ID3$ עם גיזום מוקדם ועם פונקציית ההפסד המוזכרת בשאלה ערך ה־ M המתקבל הינו $M = 1$.
לאחר אימון המודל עם ערך ה־ M שנמצא על כל סט האימון, השגיאה המתקבלת על סט הבוחן הינה 0.02123893805309735.