מבוא לבינה מלאכותית – 236501

תרגיל בית 3: מבוא ללמידה

מגישים:

מאי ליינר –
לידור סורקין –

דו"ח תרגיל בית 3

ID3

2. הוכחה/ הפרך: "בהינתן דאטה כלשהו עם תכונות רציפות ותיוגים בינאריים המחולק לקבוצת אימון ומבחן, הפעלה של פונקציית נירמול MinMax הנלמד בתרגול על הדאטה אינה משפיעה על דיוק מסווג ID3 הנלמד על קבוצת האימון והנבחן על קבוצת המבחן" (20 שורות

תשובה: **הטענה נכונה.**

פונקציית הנירמול MinMax מוגדרת באופן הבא: $MiniMaxNorm(X) = \frac{X-X_{min}}{X_{max}-X_{min}}$.

בהנחה שהחסם העליון והתחתון אכן מוגדרים כמינימום ומקסימום גם עבור קבוצת המבחן וגם עבור קבוצת הבדיקה. פונקציה זו היא חח"ע, על [0,1] ו**מונוטונית עולה.**

באלגוריתם ID3, בכל צומת נבחר את אופן הפיצול בה כתלות ב-IG מקסימלי. נראה את השפעת פונקציית הנרמול על IG: תהי n צומת המועמדת לפיצול ותהי fi תכונה הרלוונטית לצומת אותה אנו בוחנים. נסמן ב-n_i^0 את קבוצת הדוגמאות ב-n עבורה fi מחזירה 0 וב-n_i^1 את קבוצת הדוגמאות עבורה fi מחזירה 1.
איחוד הקבוצות האלה הוא n מכיוון ש-fi היא תיוג בינארי (בה"כ מחזירה 0 ו-1)

$$IG\left(f_i, n\right) = entropy(n) - \left(entropy\left(n_i^0\right) \bullet \frac{\left|n_i^0\right|}{|n|} + entropy\left(n_i^1\right) \bullet \frac{\left|n_i^1\right|}{|n|}\right)$$

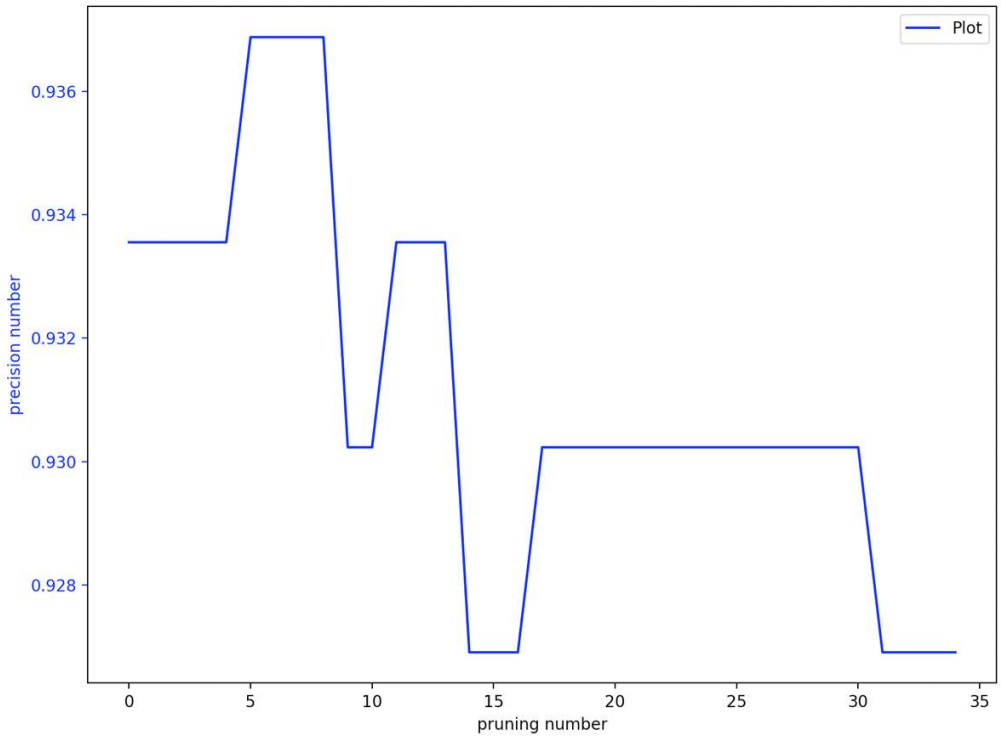
נראה כי הפעלת פונקציית נרמול על הדאטה לא תשפיע על בחירת המסווג לצומת (כאשר מסווג מוגדר ע"י feature- fi וערך סף- t).

כפי שראינו בהרצאה עבור תכונה רציפה, אנו נמיין את הדאטה שאותו fi בודק בסדרה עולה: V_j^i (כאשר, j הוא אינדקס הסדרה שמתקבל מבחינת תכונת fi). כעת, לאחר הנרמול הסדרה שלנו תהיה: $MiniMaxNorm\left(V_j^i\right)$. מכיוון שפונקציית הנרמול שלנו מונוטונית עולה נקבל כי סידור איברי הסדרה המקורית נשמר. כלומר האיבר ה-$MiniMaxNorm\left(V_j^i\right)$ יהיה עדיין האיבר ה-j-י בסדרה החדשה המנורמלת.

נסמן ב-ti את הסף שנותן IG מקסימלי עבור תכונה fi. נקבל כי הוא שווה לממוצע ערכי: $V_{j^*}^i$ ו-$V_{j^*+1}^i$ עבור j* כלשהו. ממוצע ערכי: $MiniMaxNorm\left(V_{j^*}^i\right)$ ו-$MiniMaxNorm\left(V_{j^*+1}^i\right)$ יביא לנו את אותה החלוקה של n ולכן ה-IG יהיה זהה (כי ערך זה תלוי רק במבנה של צומת n ובמבנה של החלוקה של 2 הבנים של הצומת) ולכן הפיצולים האפשריים לכל fi יהיו זהים לפיצולים האפשריים לפני הנרמול.
הראינו כי לכל תכונה ולכל סף, ה-IG נשמר תחת פונקציית נרמול זו ולכן דיוק המסווג לא תפגע.

3.

א. להלן גרף המציג את השפעת הגיזום המוקדם על דיוק העץ, המחושב בעזרת K-fold cross validation:



ביצענו כ-35 ניסויים (pruning number בין 0 ל-34)
ציר pruning number (X) מייצג את המספר המינימלי של דוגמאות בעלים על מנת להמשיך לפצל.
ציר precision number (Y) מייצג את אחוז הדיוק של העץ עם הגיזום בעזרת K-fold cross validation.

ב. כפי שראינו בשיעורים, גיזום העצים מנסה למנוע מצב של overfitting בעצי החלטה.
באלגוריתם TDIDT, אנו יוצרים עץ החלטה עקבי לכל הדוגמאות בקבוצת האימון. כל רעש בדוגמאות (סיווג לא נכון, מקרה יוצא דופן שלא מלמד על הכלל) יתבטא בעץ. גיזום העץ נותן לנו עץ עם שגיאת אימון גדולה יותר אבל יכול להקטין את שגיאת המבחן. לאחר גיזום מתאים (החלפת תת עץ בעלה), אנו עשויים לקבל עץ החלטות המייצג את הבעיה באופן טוב יותר.
יתרון נוסף, הוא שהעץ ההחלטות שלנו עלול להיות קטן יותר.

ג. באופן כללי אחוז הדיוק הממוצע של 14 ההרצות הראשונות (כאשר pruning number בין 0 ל-13) גדול יותר מאחוז הדיוק הממוצע של שאר ההרצות. תוצאה זו אינה מפתיעה מכיוון שככל ש-pruning number קטן יותר, אנו נהפוך פחות תתי עצים לצמתים וכתוצאה מכך, העץ שקיבלנו ייצג יותר טוב את קבוצת האימון.
לעומת זאת, קיבלנו שעבור ערך pruning number בין 5 ל-8 אחוז הדיוק גבוהה יותר מאשר ללא גיזום. גם תוצאה זו מתאימה לאינטואיציה שלנו, מפני שבקבוצת האימון יכול להיות רעש (תוצאות סיווג לא תקינות או תוצאות שאינן מראות על מגמה כללית אלא על מקרה יוצא דופן). בעזרת גיזום, יכולנו להפחית את ההשפעה של הרעש על בניית עץ ההחלטות שלנו.
עבור גיזום משמעותי יותר (ערכים הגדולים מ—11), קיבלנו תוצאות גרועות יותר מעץ ללא גיזום. זאת, מכיוון שאנו עושים הכללה רחבה מדי עבור הבעיה.

ד. קיבלנו את התוצאה הטובה ביותר עבור גיזום עם pruning number השווה ל-8.
אחוז הדיוק במקרה זה הוא: 0.9368770764119602.
<u>הערה:</u> קיבלנו כי עבור pruning number בין 5 ל-8 את אותו אחוז הדיוק. בחרנו ב-8 מכיוון שהעץ יכול להיות קצת יותר קטן במקרה זה.

4.

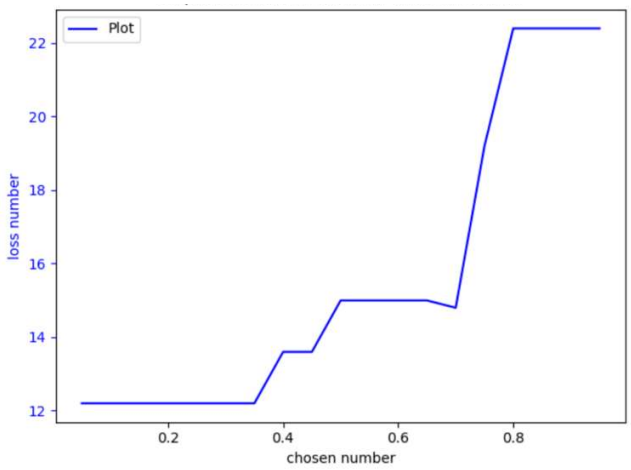
$$loss(T) := FP + 8FN$$

א. ערך הloss- הממוצע עם ערך הגיזום האופטימלי (8) הוא: 15.0

ב. ערך הloss- הממוצע שמתקבל אם מסווגים את כל הדוגמאות כ-"חולים" הוא: 30.8.
נשים לב כי ישנה עלייה של יותר מפי 2 בערך ה-loss בין 2 הדוגמאות ולכן סיווג כל הדוגמאות כ-"חולים" אינה מורידה את ערך ה-loss.
עץ ההחלטות שבנינו בסעיף קודם, סיווג אנשים גם כ"בריאים" וכ"חולים" עפ"י עץ ההחלטות שנבנה- האלגוריתם ניסה למצוא הבדל בין קבוצת החולים והבריאים ולנסות להבדיל ביניהם.
עכשיו, אנו מגדירים כל אדם כחולה. כתוצאה מזה, אנו למעשה, מבטלים את סוג השגיאות מסוג FN (שהיו פי 8 חמורות יותר מ-FP), אבל משלמים ביותר טעויות FP באופן משמעותי (אף אדם בריא לא יסווג כבריא) ולכן ה-loss גדול יותר.

ג. ערכי loss גבוהים מתקבל בעקבות טעויות בסיווג. אנו נשאף להוריד את מספר המצבים בהם יש טעויות שבהם אנו מסווגים אדם חולה כבריא מבלי לגרום לעלייה דרמטית במספר התיוגים של אדם בריא כחולה (שגם הוא משפיע על ערך ה-loss אבל פי 8 פחות). בסיס האלגוריתם שלנו הוא: ID3 עם גיזום מוקדם שבו מספר מינימלי של דוגמאות בעלים הוא 8 (שנתן לנו בסעיפים הקודמים את אחוז הדיוק הגבוהה ביותר).

ניסיון השיפור הראשון שעשינו הוא: הורדת סף החולים לתיוג בצומת כצומת חולה.
לפני השינוי, אם רוב האיברים בצומת הם חולים, נתייג את הצומת כחולה וכך גם ההיפך.
מצב זה גרם לכך שמספר האנשים בקבוצה הקטנה יותר לא היה פקטור בתיוג הצומת.
נניח כי בצומת מעורבת (שבה יש גם דוגמאות של חולים ובריאים יחד) יש 7 דוגמאות (צומת בגודל זה תהפוך לעלה) אז במצב שבו יש 3 חולים ו-4 בריאים הצומת תתויג כבריאה, למרות שמספר החולים גבוה יחסית. טעות בזיהוי חולה כבריא עולה לנו פי 8 יותר מטעות בזיהוי בריא כחולה ולכן רצינו להוריד את סף תיוג הצומת מ-0.5 (באלגוריתם המקורי, צומת תתויג כחולה אם החלק היחסי של מספר החולים הוא חצי ממספר הדוגמאות בצומת) למספר אחר. לשם כך, הרצנו את האלגוריתם הזה, עם מספרים בין 0.05 ל-0.95 בקפיצות של 0.05 המציינים את סף התיוג ובדקנו את ערך ה-loss. התוצאות מופיעות בגרף הבא:

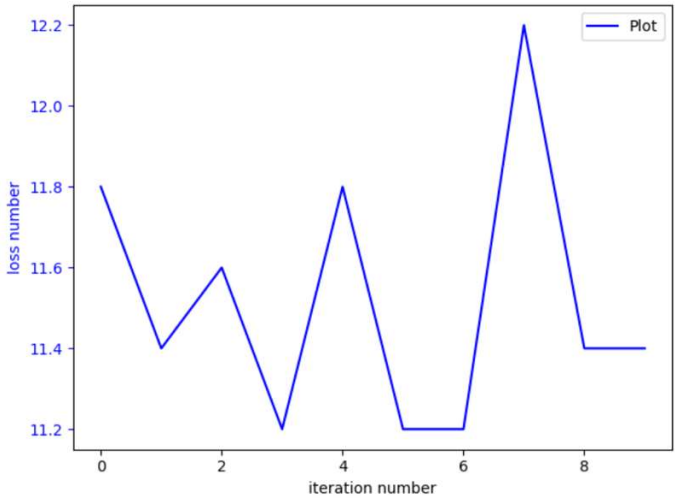


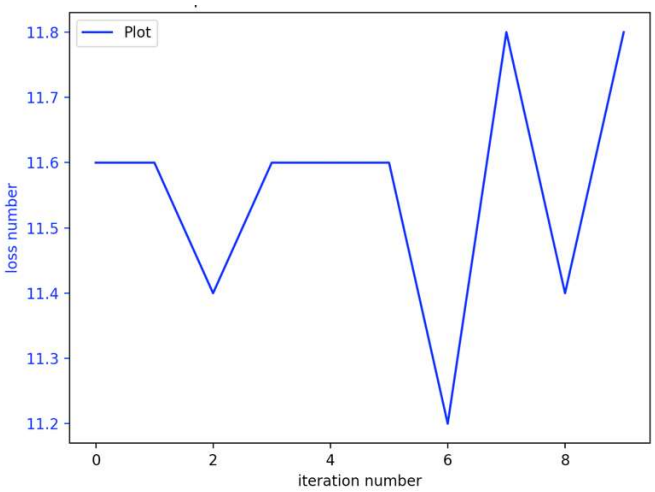
קיבלנו שערך ה-loss הנמוך ביותר מתקבל הוא 12.2. ערך זה מתקבל בין 0.05 ל-0.3. בפתרון שלנו נבחר בערך סף של 0.3 מפני שעדיין נרצה לתייג צמתים שיש להם רוב מוחלט של דוגמאות בריאות כצומת בריאה. כלומר, מעכשיו כאשר נבחר label לצומת הערך יהיה "חולה" אם מספר החולים בצומת גדול מ-0.3 כפול מספר הדוגמאות בצומת ו"בריא" אחרת.
נקבל בזכות זאת שיפור של כ-18.6% מערך ה-loss המקורי.

לאחר שינוי זה, ניסינו להוסיף שינוי נוסף: שינוי הדיסקרטיזציה של ה-feature ממסווג בינארי למסווג טרינארי.
באלגוריתם המקורי, המסווג היה בינארי וחילק כל צומת לשני בנים. כעת, המסווג הטרינארי יחלק כל צומת לשלושה בנים ע"י בחירת 2 ערכים t_1 ו-t_2, ומיון ערכי הדוגמאות לפי מספרים אלו.
המחשבה, מאחורי השינוי הייתה, שבעזרת פיצול כל צומת ל-3 בנים, נוכל לקבל דיוק יותר טוב בסיווג. שינוי זה גרם לגידול אקספוננציאלי הן במקום והן בזמן ריצת האלגוריתם. עקב מגבלת הזמן, לא בחרנו בשיפור זה לאלגוריתם, למרות שקיבלנו שיפור של 28% בערך ה-loss בעזרתו.

לאחר בחינה נוספת של הבעיה, שמנו לב שלאחר השיפור הראשון, בחלק מן הטסטים מספר טעויות התיוג של אדם חולה כבריא ירד, אך מספר הטעויות של אדם בריא כחולה עלה. בשל כך החלטנו לנסות לשנות את השיפור הראשון ע"י תיוג צומת לפי התפלגות.
בכל עלה מעורב (תת עץ שהפך לעלה ע"י אלג' הגיזום המוקדם) יש לכל היותר 7 דוגמאות.
נבחר תיוג לעלה ע"י שימוש בפונקציה רנדומלית, המגרילה מספר בין 0 למספר הדוגמאות שבעלה. אנו ניקח בחשבון את החלק היחסי של מספר החולים ביחס לגודל הצומת באופן הבא:
נניח כי גודל הצומת הוא n ומספר החולים בצומת הוא k. ההסתברות לתיוג הצומת כחולה הוא: $\frac{k}{n}$
וההסתברות לתיוג הצומת כבריאה הוא $\frac{n-k}{n}$.
לדוגמא בצומת שבה יש 7 דוגמאות כאשר 3 איברים מוגדרים כחולים ו-4 כבריאים. בעת, בניית העץ נגריל מספר בין 0 ל-6. אם ערך המספר קטן מ-3 נגדיר את הצומת כחולה ואחרת כבריאה.

מכיוון שתיוג צומת הינה בעיה הסתברותית, הרצנו את הניסוי מספר פעמים כדי לראות את השפעת שיפור זה על ערך ה-loss.





שני הגרפים שבנינו מתארים 10 איטרציות של בדיקת גודל ה-loss עבור הרצת האלגוריתם החדש.
הערך המקסימלי שקיבלנו, לא עולה על 12.2 ולכן החלטנו ששיפור זה עדיף בהשוואה לשיפור הראשון שלנו.
בממוצע כל ההרצות שלנו נתנו לנו ערך loss של: 11.58. כלומר שיפור ממוצע של 23% מערך ה-loss המקורי

ניסינו לשפר גם את האלגוריתם הזה, ע"י קביעת תיוג העלה רק בזמן ריצת ה-test.
כלומר, במצב בו נרצה לבצע pruning נסמן את העלה ב-label מיוחד שיציין שאת תיוגו יש לקבל בזמן המעבר על עץ ההחלטות בשלב ה-testing.
כאשר, נגיע לעלה מסוג זה בשלב ה-testing, נבצע את האלגוריתם מהשיפור הקודם. כלומר, עבור כל בדיקה המגיע לעלה זה, נקבע את תוצאת הבדיקה באופן רנדומלי כתלות בהתפלגות בדוגמאות בעלה.
עבור טסט שרק בדיקה אחת מגיעה לעלה זה, התוצאה לא תהיה שונה מהשיפור הקודם. אך עבור מספר בדיקות גדול יותר המגיעים לאותו העלה, נוכל לקבל עבור כל שורה ב-test תיוג שונה באופן הסתברותי. בניגוד לאינטואיציה הראשונית שלנו, שיפור זה גרם לערך ה-loss להיות גדול

יותר מערך ה-loss המקורי. ואכן, מבחינה הסתברותית, הסיכוי לבחור כל פעם מספר התואם את התיוג קטן יותר מהסיכוי לבחור מספר אחד שיתאים לכל הטסטים שלנו.

בסופו של דבר, בחרנו בשיפור השלישי שלנו, בו בעת בניית עץ ההחלטות אנו מתייגים את ערך הצומת לפי התפלגות הדוגמאות שבו מפני שהוא שיפר את גודל ה-loss באופן המשמעותי ביותר.