תרגיל בית 3 – מבוא ללמידה

**מגיש: טל רוזנצוויג**

**ת.ז:**

**שאלה 1:**

תוצאת הדיוק שקיבלתי עבור ביצוע אימון האלגוריתם על קבוצת האימון וביצוע בדיקה שלו על קבוצת המבחן היא:

**צירוף תצלום של תוצאת הדיוק**

**שאלה 2:**

הוכחה: זה העתק הדבק משני – לשנות פורמלי כמו שרצית

נניח כי קיים דאטה כלשהו עם תכונות רציפות ותיוגים בינאריים המחולק לקבוצת אימון ומבחן. נסמן ב- מסווג ID3 שיקבל את הדאטה הנתון. כעת, נשתמש בפונקציית הנרמול כדי לנרמל את הדאטה הנ"ל וניצור מסווג ID3 חדש שיקבל את הדאטה המנורמל, נסמנו ב-. נראה כי תוספת המידע(IG) המתקבלת עבור תכונה כלשהי וערך סף עבור ערכים שאינם מנורמלים, עבור קבוצת דוגמאות אימון בצומת מסוים של מסווג ID3 *, שווה לתוספת המידע המתקבלת עבור אותה תכונה ואותם הערכים לאחר הפעלה של פונקציית נרמול :* תהי תכונה כלשהי תכונת הסף המחלקת את דוגמאות האימון כך שהיא מספקת תוספת אינפורמציה כלשהי. יהא צומת כלשהו ונגדיר: *כקבוצת דוגמאות האימון הממוינות לפי ערכי שהצומת מכיל, כאשר הן אינן מנורמלות.* נסמן את ערך הסף ב- כך ש- . נניח בה"כ עבור *כאשר הינו ערך של הדוגמאות בבן ה-, מתקיים , נקבל כי:*

*קיבלנו כי הערכים שהיו גדולים מערך הסף לפני הנרמול נשארים גדולים מערך זה גם אחרי הנרמול, וכך באופן דומה יחס זה מתקיים גם עבור ערכים שהיו קטנים מערך הסף. כמו כן, נקבל כי דרך חלוקת קבוצת הדוגמאות לפי ערך הסף תישאר זהה גם לאחר ביצוע הנרמול, לכן מאחר וחישוב תוספת המידע תלוי רק ביחס בין תיוגי הדוגמאות שיש בכל בן של הצומת וכך כפי שהחלוקה נשארה זהה, כך גם תוספת המידע, כלומר נקבל כי גם לאחר ביצוע נרמול תוספת האינפורמציה תישאר זהה עבור תכונה וערך סף אלו.*

*מכאן נוכל להסיק כי עבור השורש של המסווג המנורמל תיבחר תכונת f להיות התכונה המפרידה ומאחר והראינו כי ערך ה-IG אינו משתנה בין המסווג המנורמל למסווג הרגיל, תיבחר אותה תכונה f גם עבור המסווג הרגיל. ערכי הסף אומנם יהיו שונים, אך הם ישרו אותה חלוקה של דוגמאות עבור הצומת. מכאן, ניתן להסיק כי יבנה עץ דומה עבור שני המסווגים. אם כן, נקבל כי באופן אינדוקטיבי נקבל מסווג זהה למסווג של וכך נקבל כי עבור כל קבוצת מבחן יתקבל אותו הדיוק המבוקש ואותם הסיווגים.*

**שאלה 3:**

סעיף 3.1:

*גיזום עצי החלטה נעשה כדי להקטין את העץ ולהחליש את התאמת היתר. כדי להתמודד עם בעיית מבוצע תהליך גיזום שבו מוחלפים חלק מהצמתים בעץ בעלים עם סיווג שנקבע ע"פ הסיווג של רוב הדוגמאות ששייכות לתת העץ שתחת אותו הצומת.*

*סעיף 3.3:*

הגרף המבוקש בסעיף זה:

*לצרף גרף*

גרף זה מתאר כיצד הפרמטר M משפיע על הדיוק.

לפי הגרף, ניתן לראות כי עבור *תשים את הערך שלך, היה זכור לי אפס ותמחק את הערה* מתקבל ערך הדיוק הטוב ביותר. *כמו כן, ניתן לראות בגרף כי ערכים גדולים של M גורמים לירידה בדיוק מפני שהם גוזמים ענפים בעלי חשיבות לצורך חישוב תוצאת הדיוק.*

אני הוספתי תצלום שמראה מה ה-M הטוב ביותר לפי הערך המודפס ב- experiment אבל לא חייב זה סתם תוספת.

סעיף 3.4:

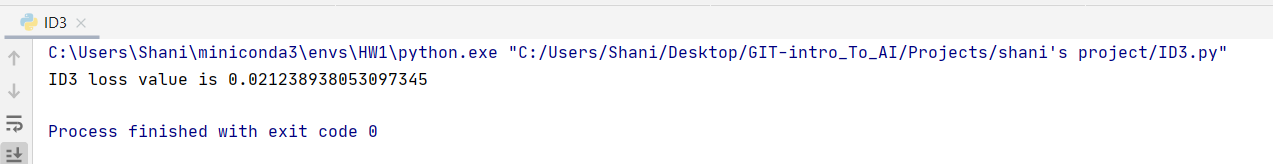
הגיזום לא שיפר את הביצועים ביחס להרצה ללא גיזום בשאלה 1 וזאת ניתן לראות בתצלום הבא:

*להוסיף תצלום שבו מוצגות 2 התוצאות ומראות שאין שיפור*

**שאלה 4:**

סעיף 4.1:

ערך ה- של הוא . *לצרף תמונה של ערך ה-loss*



סעיף 4.2:

על מנת לצמצם את פונקציית ה-loss בחרתי באלגוריתם המשפר את ערך ה-loss מפני שהצמתים שנגזמו היו צמתים שהגיזום שלהם שיפר את ערך ה-loss עבור עץ זה בהשוואה לאלגוריתם הרגיל.

אתאר את האלגוריתם שבו השתמשתי:

תחילה, ביצעתי חלוקה של ה- ל- עבור אימון ו- עבור וולידציה. לאחר מכן, בניתי עץ ID3 רגיל מ- האימון. כעת, ביצעתי מעבר על העץ מהעלים כלפי מעלה ובכל צומת ביצעתי את בדיקה עבור ערכי ה-loss באופן הבא:

* נשמור את ערכו של ה- loss עבור עץ זה ללא גיזום הצומת.
* נשמור את ערכו של ה- loss עבור עץ זה כאשר נבצע גיזום של צומת זה ונהפוך אותו לעלה כך שהסיווג שלו ייבחר לפי פונקציית ה-loss.

כעת, נבדוק האם קיים שיפור ב-loss עבור גיזום הצומת, ובמידה וכן נבצע גיזום.

נשים לב כי בשלב זה קיבלנו עץ ID3 גזום המשפר פונקציית loss.

*ככה זה אצלי העתק הדבק שתראה*

על מנת לצמצם את פונקציית ה-loss בחרתי באלגוריתם הבא:

1. חלוקת ה- לשני חלקים: לאימון ו- לוולידציה.
2. בניית עץ ID3 רגיל מ- האימון.
3. כעת, עוברים על העץ מהעלים כלפי מעלה ובכל צומת נבצע את ההחלטה הבאה:

* נמדוד loss עבור עץ זה ללא גיזום הצומת.
* נמדוד loss עבור עץ זה כאשר נגזום את צומת זה ונהפוך אותו לעלה כאשר הסיווג שלו ייבחר לפי פונקציית ה-loss.
* במידה וקיים שיפור ב-loss עבור גיזום הצומת, נבצע גיזום.

1. בשלב זה קיבלנו עץ ID3 גזום המשפר פונקציית loss.

האלגוריתם הנ"ל משפר את ערך ה-loss אל מול האלגוריתם הרגיל מאחר והצמתים שנגזמו היו צמתים שהגיזום שלהם שיפר את ערך ה-loss עבור עץ זה אל מול האלגוריתם הרגיל.

סעיף 4.3: *אצלך עשית והשארת פונקציות שעושות את זה אז תפרט על זה ב2 שורות ותוסיף תצלום*

הפרמטר שהייתי צריכה לכוונן באלגוריתם הנ"ל הוא בחירת גודל קבוצת הוולידציה.

ניסיתי באופן ידני ערכים בין 0.2 ל-0.8 בקפיצות של 0.01 עד שהגעתי לערך 0.56 שממנו הייתי מרוצה מהתוצאה. – לא רלוונטי אליך אבל הוספתי לך שתראה, אצלך תפרט

**שאלה 5:**

סעיף א':

מקרא:

מסווג מסומן בגרף בקו מקווקוו ומסווג המטרה מסומן בגרף בקו רציף בצבע כחול.

הסבר

ניתן לראות כי בגרף קיימת דוגמא המסווגת כחיוביות ומסומנת בכחול, וקיימות שתי נקודות המסווגות כשליליות והן מסומנות באדום. דוגמת המבחן נמצאת בנקודת  *והיא מסומנת בצהוב.*

עבור דוגמת המבחן זו ניתן לראות בבירור כי מתקיים לכל ערך K שייבחר כי למידת עץ ID3 תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית, אך למידת KNN מניבה מסווג שעבורו דוגמת מבחן זו הוא יטעה לכל ערך שייבחר.

*בדוגמת מבחן זו אנו רואים כי לכל K שייבחר, מתקיים כי מסווג יחזיר תשובה שלילית עבור דוגמת המבחן כאשר הסיווג הנכון עבור דוגמה זו. לעומתו, מסווג יסווג נכון לכל ערך של שיבחר לכל דוגמת מבחן אפשרית(התקבל מסווג המטרה).*

*למשל, עבור k=1 מסווג KNN יחזיר תשובה שלילית ומסווג ID3 יחזיר תוצאה חיובית, בהתאם למסווג המטרה, כלומר הוא יסווג את הערך הנכון.*

**סעיף ב'**

מקרא:

מסווג מסומן בגרף בקו מקווקוו ומסווג המטרה מסומן בגרף בקו רציף בצבע כחול.

הסבר

ניתן לראות כי בגרף קיימות שלוש דוגמאות המסווגת כחיוביות ומסומנת בכחול, וקיימות שלוש דוגמאות המסווגות כשליליות והן מסומנות באדום. דוגמת המבחן נמצאת בנקודת  *והיא מסומנת בצהוב. כמו כן, נשים לב כי עבור הרבעון המסומן ב-*  מבטא שערכו של מסווג ID3 ברבעון זה הוא חיובי ובכל שאר שלושת הרבעונים הוא שלילי.

עבור דוגמת המבחן זו ניתן לראות בבירור כי מתקיים עבור כי למידת מסווג תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית, אך למידת עץ ID3 מניבה מסווג שעבורו דוגמת מבחן זו הוא יטעה*. בדוגמת מבחן זו אנו רואים כי עבור ערך מתקיים כי מסווג ID3 יחזיר ערך חיובי בהתאם לרבעון אליו הוא שייך. לעומתו, מסווג KNN יחזיר ערך שלילי ואכן זהו הערך הנכון, בהתאם לתוצאת מסווג המטרה שהגדרנו שגם יחזיר ערך שלילי.*

**סעיף ג'**

מקרא:

מסווג מסומן בגרף בקו מקווקוו ומסווג המטרה מסומן בגרף בקו רציף בצבע כחול.

הסבר

ניתן לראות כי בגרף קיימות שלוש דוגמאות המסווגת כחיוביות ומסומנת בכחול, וקיימות שלוש דוגמאות המסווגות כשליליות והן מסומנות באדום.

קיימות שתי דוגמאות מבחן:

1. דוגמת מבחן ראשונה הנמצאת בנקודת  *והיא מסומנת בצהוב.*
2. דוגמת מבחן שנייה הנמצאת בנקודת  *והיא מסומנת בירוק.*

*כמו כן, נשים לב כי עבור הרבעון המסומן ב-*  מבטא שערכו של מסווג ID3 ברבעון זה הוא שלילי ובכל שאר שלושת הרבעונים הוא חיובי.

*עבור הדוגמא הראשונה נקבל:*

|  |  |  |  |
| --- | --- | --- | --- |
| *מסווג מטרה* | *מסווג KNN* | *מסווג ID3* | *ערך K* |
| *-* | *-* | *+* | *1* |
| *-* | *-* | *+* | *3* |

*עבור הדוגמא השנייה נקבל:*

|  |  |  |  |
| --- | --- | --- | --- |
| *מסווג מטרה* | *מסווג KNN* | *מסווג ID3* | *ערך K* |
| *+* | *-* | *+* | *1* |
| *+* | *-* | *+* | *3* |

כאשר בדוגמאות אלו הראיתי את הנדרש והצגתי כי מתקיים עבור שניהם ערכי K שונים.

**סעיף ד'**

מקרא:

מסווג מסומן בגרף בקו מקווקוו ומסווג המטרה מסומן בגרף בקו רציף בצבע כחול.

הסבר

ניתן לראות כי בגרף קיימת דוגמא אחת המסווגת כחיובית ומסומנת בכחול, ודוגמא אחת המסווגת כשלילית ומסומנת באדום.

מסווג ID3 יסווג כערך שלילי עבור החלק שקטן שווה ל- ועבור מסווג ID3 יסווג כערך חיובי.

עבור וקבוצת האימון שהוצגה, למידת מסווג מסוים תניב מסווג אשר יחזיר תוצאה נכונה עבור כל דוגמת מבחן אפשרית וכך גם למידת עץ תניב מסווג העונה נכונה עבור כל דוגמת מבחן אפשרית כנדרש.

**שאלה 6:**

סעיף 6.1

*לפרט על הניסויים שעשית – אצלך זה בפונקציה אז לפרט מעט כי אתה לא נדרש.*

הדיוק המקסימלי שקיבלתי על קבוצת המבחן הוא:

*לצרף תצלום*

**שאלה 7:**

סעיף 7.1

אסביר בקצרה את שני השיפורים שביצעתי בחלק זה:

שיפור ראשון:

*זהה לגמרי לשלי – לשנות אבל זה הרעיון קצר קולע*

עד כה, המסווג קבע את תוצאת החיזוי שלו לפי התוצאה שחזרה הכי הרבה פעמים מתוך K העצים שנבחרו באמצעות מרחק אוקלידי מ- האימון של העצים ל- המבחן.

בשיפור זה, לכל עץ ישנו מרחק מה-של המבחן כך שהחיזוי שהוא נותן מקבל ציון כפונקציה יורדת של המרחק שלו מ-המבחן כך שנקבל שככל שהמרחק גדול יותר, הציון יהיה נמוך יותר. לכן, המסווג יקבע את תוצאת החיזוי על פי החיזוי שקיבל את התוצאה הגבוהה ביותר. כמו כן, לצורך כך בחרתי את הפונקציה היורדת  *כך ש- מייצג את המרחק.*

שיפור שני:

*לעטוף את החלק הזה ולשנות מעט*

כל תכונה נבדקת פעמים לפי ה-kfold. בכל נבצע בדיקה מהו דיוק החישוב ללא פרומטציה של התכונה ולאחר מכן נבדוק מהו ערך דיוק החישוב עם פרמוטציה של התכונה. ההפרש בין דיוקים אלו מצביע על חשיבות התכונה לצורכי דיוק – כך שככל שההפרש גדול יותר, התכונה חשובה יותר, ועל פי כך נקבעו המשקלים.

**סעיף 7.2:**

הפרמטרים שהשתמשתי בהם לצורך קביעת הפרמטרים לאלגוריתם הינם זהים לפרמטרים של סעיף 6.1. אצלך אתה לא צריך לפרט כי יש לך פונקציות שעושות את זה אז תוכל לרשום גם את המשפט הזה שזה זהה לסעיף 6.1

*לצרף תצלום של הדיוק ועם תוצאה אחת של הפונקציה experiment כמו שמצוין מטה*

