ניסוי ועד העצים בבינה מלאכותית

ת.ז : 318489200 שם : מחמוד נסאר

<u>תיאור הניסוי :</u>

הניסוי שיש בידינו בודק מספר דרכים ליצירת מסווג בעזרת מספר מסווגים של עצי החלטה. מספר מסווגי עצי ההחלטה הנ"ל הוא 101. כל דרך של יצירת מסווג מבוססת על קרטריון מסוים (שנזכור בהמשך).

קודם כל אנו בוחנים שלושה דאסטים עבור כל דאטסט אנחנו מחלקים את הדאטאסט לחמישה פולדים, ולכן יהיה 5 איטרציות, כל פעם פולד אחר יהיה הוא קבוצה המבחן ושאר הפולדים יהיו קבוצת הלמידה, הדיוק מוגדר להיות ממוצע הדיוקים שהתקבלו בחמישה ההרצות.

עבור כל איטרציה,כל עץ מיוצר בצורה הבאה, מחלקים רנדומלית את הפולד של קבוצה הלמידה לשתי קבוצות, אחת שווה ל-$\frac{2}{3}$ מהגודל המקורי של קבוצה הלמידה, ניעזר בה כדי ללמד את העץ (קבוצת הלמידה) , והשניה שווה ל-$\frac{1}{3}$ מהגודל המקורי של קבוצה הלמידה (קבוצת ההערכה) וניעזר בה כדי להעריך את דיוק העצים.

אנחנו בודקים חמישה דרכים אפשריים ליצרית מסווג והם :

מלמדים 101 עצים כמו שתיארנו למעלה, ואז

1- המסווג שנוצר הוא בעצם מסווג שמחליט כמו שמחליטים רוב ה- 101 עצים
2- נבחר אקראית 21 עצים מבין ה- 101, המסווג שמוגדר הוא מסווג שמחליט כמו שמחליטים ה- 21 עצים.
3- אחרי שלמדנו את העצים השונים ע קבוצת המבחן, נעריך את הדיוק שלהם בעזרת קבוצת ההערכה. נבחר את 21 עצים הטובים ביות מבחינת הדיוק הנ"ל, המסווג ש- מוגדר הוא מסווג שמחליט כמו שמחליטים ה- 21 עצים.
4- בהינתן דוגמא לסיווג מקבוצת המבחן, נבדק המרחק שלה ממרחק של קבוצת הלמידה של כל אחד מהעצים (בשבילנו מרחק בין קבוצה הלמידה לדוגמא שווה לממוצע המרחקים האוקלידים בין כל דוגמא בלמידה לבין הדוגמא לסיווג), נבחר את 21 העצים עם המרחקים הנ"ל הכי קטנים, המסווג מוגדר להיות המסווג שמחליט על כל דוגמא לסיווג כמו שמחליטים ה- 21 עצים הנ"ל (נשים לב כי לכל דוגמא לסיווג יכול להיות 21 עצים אחרים).

תיאור אוספי הנתונים :

1- האוסף נתונים הראשון הוא בשם $heart\,attack\,analysis$ והוא אוסף שכל שורה בו מהווה בן אדם מסוים עם מספר מהתכונות שיש לו, ושדה מטרה שהערך שלו קובע האם לבן אדם עם התכונות הנ"ל יש סיכוי מוגבר להתקף לב או לא.

יש באוסף הנ"ל 304 דוגמאות, כל דוגמא יש לה 13 תכונות, התכונות הם תיאורים של המאפיינים הרפאוים של הבן אדם, כמו כמות הסוכר בדם, דופק מקסמאלי שנמדד וכו"..

2- האוסף השני הוא בשם $voice\,gender\,Recognition$ והוא אוסף שכל שורה בו מהווה בן אדם, עם מספר מהתכונות של הקול שלו, ושדה מטרה שהערך שלו קובע אם הבן אדם הוא שכר או נקבה.

יש באוסף הנ"ל 1168 דוגמאות, לכל דוגמא יש 20 תכונות, התכונות הם תיאורים של גלי הקול של הבן אדם, לדוגמא סטייה של התדר, אנטרופיה ספקטרלית וכו" ...

3- האוסף השלישי הוא בשם $water\,potability$ והוא אוסף שכל שורה בו היא דגימה מסויימת של מים, עם מספר מהתכונות של המים שיש בדגימה, ושדה מטרה שהערך שלו קובע האם המים (שנלקחה ממנו הדגימה) ראוי לשתיה או לא.

יש באוסף הנ"ל 548 דוגמאות, לכל דוגמא יש 9 תכונות, התכונות הם תיאור של האמפיניים השונים של המים בדגימה, למשל סך המוצקים שהמתומססים במים, כמות פחמן אורגני וכו" ...

תיאור של האלגוריתם :

האלגוריתם משתמש בעצי החלטה כדי לממש את הניסוי שתיארנו, כל אחד מ-101 המסווגים הוא עץ החלטה. הנה $pseudo-code$ של המימוש :

Input : K = 101 , C = 21

- For dataset in datasets:
    - Divide the dataset into 5 folds, iterate 5 times, each time one of fold is test set and the other 4 is training set :
    - Repeat K times:

        - Divied the training set randomly into 2 sets, s.t. the new training set is 2/3 of the original one, and the remaining is evaluation set.
        - Train a classifier on the training set
        - Predict the classification of evaluation set
        - Get the precision of the prediction
        - Save the classifier, it`s training and evaluation set and precision.
    - Get the prediction of the saved K classifiers on the test set
    - print the precision of the majority prediction of the classifiers
    - print the precision of the majority prediction of C random classifiers from the K
    - sort the K classifiers according to their saved evaluation precision, get the best C classifiers in this sort, and print out their majority prediction.
    - For every example in test set:
        - Calculate its distance from the training of every classifier
        - Get the C classifiers that their set is closer to example
        - Set the prediction of this example to be the majority prediction of the C classifiers
    - Print the precision of the predictions stated in the last loop.

- ההגיון מאחורי שימוש ביותר מאוסף נתונים הוא שיהיה לנו תוצאות תלויות פחות באוספי הנתונים עצמם, ושתהיה התוצאת תלויות רק בדרך בחירת המסווגים.
- השתמשנו ב- k-cross validation, כי היא ידועה להיות השיטה הכי טובה להעריך את הדיוק של המסווג על אוסף נתונים ולנטרל רעשים על ההערכה, כמו סוג החלוקה של סטים האימון והבדיקה.
- אפשר לראות שלמרות חישוב הדיוק ב 4 גישות שונות, בכל איטרציה של פולד מסויים, מאמנים כל עץ פעם אחת, ועושים חיזוי פעמיים, כלומר ביצוע רק הפעולות הנחוצות, בלי חזרות על הפעולות, וזה כדי להוריד מסובוכיות הזמן, דבר שמאוד חשוב בתחום הזה בדרך כלל, וספציפית פה מאחר ואלגוריתם זה מאמן הרבה עצים ולוקח הרבה זמן יחסית.

תוצאות :

נגדיר שמות :

מסווג הרוב : המסווג שנבחר בעזרת לקיחת ההחלטה של רוב העצים מבין ה-101.

המסווג הרנדומלי : המסווג שנבחר בעזרת לקיחת ההחלטה של 21 עצים רנדומלים מביו ה- 101.

המסווג מדויק ההערכה : המסווג שנבחר בעזרת לקיחת ההחלטה של 21 העצים עם דיוק הכי גבוהה על קבוצת ההערכה מבין ה-101.

המסווג הקרוב : המסווג שנבחר בעזרת לקיחת 21 העצים שיש להם סט דגומואות למיד הכי קרוב לדוגמת הטסט (עבור כל דוגמת טסט)

שם אוסף הנתונים	דיוק מסווג הרוב	דיוק המסווג הרנדומלי	דיוק המסווג מדויק ההערכה	דיוק המסווג הקרוב
heart attack analysis	94.4%	92.1%	94.7%	92.4%
voice gender Recognition	99.4%	99.5%	99.4%	99.1%
water potability	93%	89.2%	91.7%	89.25%

סיכום ומסקנות:

אפשר לרראות בתוצאות לפי כל אוספי הנתונים השונים, ש**דיוק מסווג הרוב ו- דיוק מסווג מדויק ההערכה** הם הדיוקים הכי גבוהים מבין ה- 4 נסויים, בין שני הניסויים האלה יש קרב בתוצאות וקשה לקבוע בוודאות מי עדדיף על השני. אפשר לראות למשל באוסף *heart attack analysis* שדיוק מסווג מדויק ההערכה יותר טוב מ- דיוק מסווג הרוב ולהפך באוסף הנתונים water potability .

לגבי המסווגים הפחות טובים המסווג הקרוב ו- המסווג הרנדומלי, אפשר להגיד שהם פחות טובים באותה רמה, מאחר וגם בינם יש קרב בתוצאות עבור אוספי הנתונים השונים.

גם יש שוני בטווחים הדיוקים בין אוספי הנתונים השונים, וזה מחזק את הטענה שטענו בהתחלה, והיא שהוספת אוסף נתונים מנטרל את השפעת אוסף הנתונים עצמו על הניסוי, והשוני בטווחים מעיד על כך שדפוס התוצאות שקיבלנו הוא כללי.

לפני המסקנות אני רוצה להעיר, שלמרות הדיוקים בכיוונים השונים עבור אותו אוסף נתונים נבדלים רק באחוזים בודדים, זה בשביל הניסוי שלנו לא הבדל קטן, מאחר וכפי שהסברנו אנו משתמשים במספר טכניקות (כמו k-cross validation , מספר של אוסף נתונים ועוד...) שמנסות לנטרל את הרעשים השונים, כדי שזה יהיה הבדל נובע אך ורק משינוי שיטת בחירת המסווג.

אנחנו מסיקים מהתוצאות שדיוק הרוב הוא אחד הטובים מבין ארבעת הגישות שבחנו, אפשר להסביר את זה בכך ששיתוף מספר עצים יותר גדול יכול להביא לתוצאות יותר טובות, מאחר וזאת היא הגישה היחידה, שהמסווג שלה לא מוגבל בבחירת 21 עצים, בנוסף אפשר להסיק שעם כל החוכמה שהפעלנו בשלושת הגישות האחרות, בחירת הרוב היא שיטה טוב ומחזיקה מעמד מול שיטות מסובוכות יותר.

דיוק המסווג מדויק ההערכה לא פחות טוב ממסווג הרוב, זה מעיד על כך שמסווג טוב היא לא תכונה מאוד ספציפית לדוגמאות מסויימות, זאת אומרת שמסווג טוב עבור קבוצת טסטים מסויימת (שהתוצאות שלו יציבות ולא תלויות באופן ישיר באוסף הנתונים) סביר להניח שהיה מסווג טוב עבור קבוצת טסטים אחרת (עבור אותה בעיה).

דיוק המסווג הרנדומלי הוא פחות טוב, וזה מצופה, כי בפועל בחירה רנדומלית יכולה להיות טובה במקומות אחרים עבור מטרות שונות, למשל הורדת רעשים, אבל בעניין של בחירת העצים שחקרנו, לבחור עצים רנדומלים זורק את היתרון הדגול שלנו בזה שיש הרבה עצים, והוא היתרון של בחירה מושכלת מסין העצים שיכולה לנצל את הטיב של עצים מסויימים ולנטרל את הרוע של עצים אחרים.

דיוק המסווג הקרוב הוא לא טוב יחסית, זה אולי נובע מהעובדה שאנו **רק** מסתמכים על המרחק של דוגמאת הטסט מאוסף דוגמאות הלמידה, כלומר יכול מאוד להיות שהעצים "שלמדו הכי קרוב" לדוגמאת הטסט, הם עדיין עצים לא טובים, כי שום דבר לא מבטיח שהעצים האלה מדויקם אפילו על הדוגמאות שהם למדנו מהם, מאחר ובכיוון הזה אנו בחרנו רק לפי הקריטריון הנ"ל.