**אז אילו אלגוריתמים יש ל"תחרות"**

שילוב הפרמטרים והתכונות הטוב ביותר(אולי גם המקום השני במקרים מסוימים) ילכו לניסוי בקבוצת המבחן עצמה.

1. עץ החלטה בודד-לא ממש מיועד לשימוש ישיר אלא יותר עבור שימוש במסווגי boost. יש כמה עצים שהולכים סביר אבל כנראה overfit גדול. אין חשיבות לנרמול.

הולך טוב יותר משמעותית כאשר ממושקל.

הכי טוב בלי סינון תכונות:

0.011465824869612944

סינון תכונות מביא ליתרון קל ב75% הטובות ו50% הטובות. הכי טוב נצפה ב50% תכונות:

0.015529765155652629

1. יער עצי החלטה אקראיים-האלגוריתם לפי הספר. זמני ריצה גדולים, נראה שיחסית לא שמרן... אין חשיבות לנרמול. נבדק עם sampler בגלל זמני הריצה וhot spot.

הכי טוב בינתיים:

0.032822939866369735

ייתכן שיש תועלת בצמצום ל90% תכונות, אם כי עוד לא נבדק ממש עם hotspot. נראה עם זאת שיותר תכונות יותר טוב גם הפעם.

1. עצים אקראיים מאוד-האלגוריתם לפי הספר. מצליח פחות טוב מיער אקראי רגיל, אין חשיבות לנרמול, לא שמרן. זמני ריצה גדולים.

הכי טוב:

0.012198287765810521

הורדת תכונות פוגעת מאוד.

1. רגרסיה לינארית: האלגוריתם הפשוט עובד לא רע, אין משמעות רבה לנרמול –כנראה כי אפשר גם בתוך האלגוריתם. הכי טוב בינתיים עם 76 פיצ'רים(נקודת המקסימום אם כי אין הבדל גדול)

הכי טוב בינתיים:

-0.021077762619372484

ridge-מצליח טיפה יותר טוב, עובד יותר טוב עם נרמול(שלהם-לא שלנו) וfit. נראה שאין לרוב חשיבות גדולה לsolver מבחינת דיוק אלא אולי מבחינת זמן ריצה אבל כולם די מהירים. נראה שאלפא בינוני(0.4-0.6) נותן לרוב תוצאות טובות.

באופן כללי די יציב. אין קורולציה ברורה בין אם מנרמלים נרמול שלנו או לא מבחינת ביצועים.

הכי טוב בינתיים:

על 76 תכונות:

-0.015139116202946033

יש לציין שאין הבדלים גדולים מבחינת התכונות.

elastic net-מצליח יפה, נרמול שלנו משפר מאוד את הביצועים. לא מעט פרמטרים, אבל כולם נבדקו. נראה שהולך יחסית טוב עם נרמול שלהם בנוסף, ערכי האלפא והl1\_ratio האופטימליים תלויים במספר התכונות.

מספר בינוני של תכונות מביא לשיפור גדול עד כדי רווח:

הכי טוב בינתיים:

על 76 תכונות:

0.00959749796029372

\*לא\* ניתן להשתמש בהם לboosterים בגלל המימוש הייחודי.

1. רגרסיה לוגיסטית: יש 6 תת אלגוריתמים. משקול עוזר מאוד, ונרמול חובה. משום מה דווקא המקרים שאין התכנסות נותנים תוצאות עדיפות-אולי העלאת מספר האיטרציות תעזור למרות שזה מגדיל מאוד זמני ריצה. לא יציב ותלוי מאוד בפרמטרים.

רק עם הפותרים sag, saga יש תוצאות טובות-השאר נוראי בכל פרמטר או מספר תכונות.

בפותרים הטובים התוצאות עדיפות עם יותר תכונות.

הכי טוב בינתיים-:

כל התכונות:

0.028667776852622815

1. רגרסיה לוגיסטית לא מולטי לייבל. \*לא\* ניתן להשתמש בהם לboosterים בגלל המימוש הייחודי. משקול \*מזיק\* מאוד, נרמול הכרחי.

מאוד יציב ונותן תוצאות עדיפות בדרך כלל מאשר מולטי, אבל כחלק מהיציבות-גם המקום הראשון פחות טוב...

הכי טוב בינתיים:

0.0026582625376815514

1. Adaboost-על עץ מאוד טיפש, ועל העץ המוביל שלנו, ועל רגרסיה לוגיסטית שלנו. אין הרבה פרמטרים לשנות, נראה שמוביל לשיפור ביחס לעץ בודד-אבל לא בהרבה.

אין חשיבות לנרמול למעט עבור רגרסיות שם המצב נורא גם ככה(יפורט בהמשך).

על עץ שלנו הוא לא יציב ויש הרבה חשיבות לבחירת פרמטרים(למרות שאין הרבה). על המסווג הדפולטי הוא מאוד יציב לרעה...

יש לציין שהפרמטר score שלו כמובן אינו בדיוק מה שאנחנו מחפשים(הוא מחפש דיוק בסיווג נטו).

הכי טוב בינתיים: על עץ משלנו,

0.001530386740331486

האפשרות של עץ טיפש או רגרסיה לוגיסטית נכשלו קשות ואפילו אין טעם לבדוק להם הבדלים במספר התכונות.

1. Gradient boost-על העץ המוביל שלנו, על מסווג טיפש ועל לא מסווג בכלל. אפשר לנסות על המוביל של רגרסיה לוגיסטית. המון פרמטרים שלא כולם נבדקו, תוצאות לא יציבות שמאוד תלויות פרמטרים. מביא לתוצאות מרשימות!

זמני ריצה עצומים לכן בוצע sampler כמעט ללא hotspot. עם זאת זה משתלם-אחלה תוצאות!

נראה שמהמר מאוד "שמרני".

יש לציין שהפרמטר score שלו כמובן אינו בדיוק מה שאנחנו מחפשים(הוא מחפש דיוק בסיווג נטו).

הכי טוב בינתיים על עץ משלנו:

76 תכונות:

0.04706637530729311

על מסווג אפס:

25 תכונות:

0.02485008328706272

על מסווג טיפש:

36 תכונות:

0.036390616167326174