Report

רקע

בתרגיל זה התבקשנו לממש שלושה אלגוריתמי למידה: KNN, Perceptron ו- PA. הדאתא עליו מימשתי את האלגוריתמים הנ"ל הינו דאתא על איכויות של יינות. בסט האימון התקבלו 355 דוגמאות עם הלייבלים 0,1,2 (עבור כל דוגמא) ובסט המבחן התקבלו 89 דוגמאות חדשות. כל דוגמא מכילה 12 פיצ'רים, כאשר 11 פיצ'רים בעלי ערכים מספריים ופיצ'ר נוסף המתאר את סוג היין מטיפוס מחרוזת (W- עבור יין לבן ו- R- עבור יין אדום).

Pre-Processing

לפני הפעלת אלגוריתמי הלמידה על סט האימון נרצה לבצע עיבוד מקדים לדאתא. כחלק משלב זה עלינו להמיר את העמודה הקטגוריאלית שמציינת את סוג היין (אדום/לבן) לעמודה עם ערכים מספריים (1/0 בהתאמה) וזאת נעשה ע"י שיטת label encoding ידועה ללא שימוש בספריית sklearn. לאחר ביצוע שיטה זו, נקבל סט אימון של 12 פיצ'רים שכולם מכילים ערכים מספריים. לאחר מכן, נרצה לנרמל את הדאתא כדי שכל הערכים שלנו ינועו בתחום 1-0. נבדקו שתי שיטות נרמול שנלמדו: z-score ו- min-man. עבור שיטת הנרמול min-max normalization התקבלו ביצועים קרובים לביצועים שהתקבלו עבור השימוש ב- z-score normalization בהרצת האלגוריתמים השונים. יחד עם זאת, החלטתי להשתמש ב- min-max normalization מאחר ושיטה זו שומרת על היחסים בין הערכים המקוריים.

קביעת משקלים והיפר פרמטרים

קביעת המשקלים הראשונית עבור שני האלגוריתמים Perceptron, PA זהה והינה שתי מטריצות משקלים של אפסים אשר מתעדכנות תוך כדי ריצה בהתאם לסט האימון שלנו.

כחלק מהפעלת אלגוריתמי הלמידה יש צורך בקביעת היפר פרמטרים קבועים אשר ייצגו את קצב הלמידה ומידת הענישה במקרה של קלסיפיקציה לא נכונה. אלו פרמטרים שלא נלמדים/משתנים בתהליך הלמידה. ערכים אלו נקבעו עבור כל אלגוריתם בנפרד. פרמטרים אלו נבחרו באמצעות ריצת האלגוריתם על מספר סופי של ערכים ובחינת הערכים בעלי אחוזי ההצלחה הגבוהים ביותר.

הפרמטרים שנקבעו עבור כל אחד מהאלגוריתמים:

KNN Algorithm- עבור אלגוריתם זה נבחר K = 7. בבדיקה באמצעות k-fold cross validation (עם k = n) התקבל שהדיוק הגבוה ביותר מתקבל עבור K = 7 ומגיע לסדר גודל של לפחות 80%.

Perceptron Algorithm- עבור אלגוריתם זה נבחרו 100 אפוקים ו- eta = 0.1 המבטאת את ה- learning rate של האלגוריתם.

גם עבור היפר פרמטרים אלו נשתמש ב- k-fold cross validation. בבדיקה עבור ערכי eta שונים בטווח של [0.0,1.0] התקבל שהדיוק הגבוה ביותר של האלגוריתם מתקבל עבור eta = 0.1. באותו אופן, נבדקו ערכי אפוקים שונים בטווח של [0,1000] ונמצא כי רמת הדיוק הגבוהה ביותר מתקבלת עבור סביבות 100 אפוקים. אציין שרמת הדיוק שהתקבלה באלגוריתם מגיעה לסדר גודל של לפחות 60%.

Passive Aggressive Algorithm- עבור אלגוריתם זה נבחר לבצע אפוק יחיד. האלגוריתם נותן משקל מאוד גדול לדוגמאות האחרונות שראה, ולכן לא משנה אם נבצע אפוק יחיד או 100. בסופו של דבר, הדוגמאות האחרונות יהיו בעלות המשקל הגדול ביותר. כלומר, סדר הדוגמאות משפיע הרבה יותר מאשר כמות האפוקים. בנוסף לכך, רמת הדיוק שהתקבלה באלגוריתם מגיעה לסדר גודל של לפחות 60%.

<u>k-fold cross validation</u>

על מנת לקבוע את ההיפר פרמטרים של KNN ו- Perceptron השתמשתי בשיטה זו עם k = n. למעשה, בכל ריצה מוציאים נקודה אחת מסט האימון ומאמנים את המודל על יתר הדאתא, ואז מנבאים את התיוג לנקודה שהוצאנו. באופן דומה עושים את התהליך עבור כל הנקודות בסט האימון. לבסוף, רמת הדיוק שחושבה עבור ביצועי כל אלגוריתם היא שיעור הפעמים שהמודל צדק בחיזוי.