אורן אופזהר, ת"ז 205633423
יבגני פורוסט, ת"ז 314391558

AdaBoost – Documentation

את האלגוריתם מימשנו ב – Python 3, על בסיס מחלקת AdaBoost המכילה 2 פונקציות (למידה ובדיקה) ופונקציית main אשר דרכה מתבצעת קריאת הקובץ ושליטה בפרמטרים השונים.

מחלקת AdaBoost –

שדות המחלקה AdaBoost ושימושן:

1. data_set ו – test_data – מקבלים את בסיס הנתונים המלא למטריצה באמצעות pandas.
2. T מציין את מספר ה – decision stumps שברצוננו ליצור על בסיס הנתונים. יצירת ה – stumps מתבצעת ע"ב ספריית DecisionTreeClassifier ושימוש בעצי בחירה בעומק 1.
3. מערכי stumps, alphas לשמירת המסווג שביצע הכי טוב בכל איטרציה והמשקל שלו.
4. מערך accuracy לתיעוד ביצועים וחישוב דיוק של הלמידה.
5. train_accuracy לחישוב דיוק של ההיפותזה הסופית.
6. Target_acc הוא שדה שניתן לבחירה בפונקציית ה - main ובו ניתן לבחור מתי פונקציית predict תעצור.
7. מערך predictions – טבלת החיזויים של המודל.
8. num_of_stumps_used ו – train_num_of_stumps_used – מספר stumps שהשתמשנו בו לחישוב ההיפותזה ולחיזוי בהתאמה.
9. Same – משתנה בוליאני המייצג האם הנתונים ב – test_data יהיו מהנתונים שב – data_set.

פונקציית fit –
הפונקציה נקראת פעם אחת מפונקציית ה – main לאחר יצירת אובייקט של המחלקה, מחלקת את בסיס הנתונים לשדה הנתונים ושדה תיוגים (0 או 1). לשם ביצוע AdaBoost כל תיוג 0 של דוגמה מסוימת הומר למינוס 1.
הפונקציה פועלת בצורה טריוויאלית ע"פ האלגוריתם כפי שהצגנו אותו במצגת, כשלאחר שיצרנו stump, ביצענו חיזוי עליו ושמרנו את החיזויים והשגיאות במערך Process תחת התוויות המתאימות לכל נתון.
בשורה 44 מתבצע חישוב השגיאה של ה – stump הנוכחי כשלבסוף ה – stump בעל השגיאה המינימלית נבחר ומתבצעים החישובים הבאים:

1. שורה 58 – חישוב משקל המסווג הנבחר, ובעצם מידת ההשפעה שתהיה לו בהצבעה הסופית.
2. שורה 61 – חישוב מחדש של משקלי הנתונים בהתאם לאם החיזוי שלהם היה נכון או שגוי.

כך מתבצעת יצירת המודל הסופי (בחירת מסווג אחר מסווג במשך T פעמים).

בנוסף, מתבצע חישוב דיוק של המסווגים שחושבו עד כה בכל איתרציה, במקרה שבסיס הנתונים לבדיקה שונה מבסיס הנתונים לחישוב ההיפותזה הסופית וקבוצת המסווגים באיתרציה כלשהי השיגה את הדיוק המירבי הלולאה תעצור והדיוק ישמר.

הערות:

- הערך T הוא ערך שניתן לשינוי בפונקציית ה – main בשדה number_of_base_learners.
  - ככלל, הגדלת מספר המסווגים הבסיסיים יעלה בסופו של דבר את הדיוק הסופי של המודל, אך עם זאת, יש בסיסי נתונים קטנים שלא יהיה צורך ביותר מ10-20 מסווגים כדיי להגיע לאחוזי דיוק גבוהים מאד. (בבסיס הנתונים הקטן שצירפנו, כ-41 מסווגים השיגו 100% דיוק)

- מנגד, בבסיסי נתונים גדולים מאד וקשים (בעלי מספר דוגמאות ומספר מאפיינים גבוה מאד), AdaBoost עלול להתקשות בהגעה ל – 100% דיוק ויידרשו לכך אלפי ואפילו עשרות אלפי מסווגים, דבר שלא יכולנו גם לבדוק במחשבנו הביתי מבחינת זמן ריצה וכוח חישוב דרוש.

פונקציית predict –
הפונקציה נקראת לאחר סיום הרצת fit ומבצעת בדיקה על בסיס הנתונים אותו למדנו, או על בסיס נתונים אחר, על מנת לקבל את אחוזי הדיוק של המודל. באמצעות המסווגים וכוח ההצבעה שלהם (alpha) אשר שמורים בשדות המחלקה, ביצענו בלולאה בדיקה על בסיס הנתונים, כאשר כל מסווג חלש שנבחר בתורו מבצע חיזוי לכל דוגמה בבסיס הנתונים ועבור כל מסווג נשמרים במערך predictions חיזויים (1 או 1-) בהתאם למיקום הדוגמאות בבסיס הנתונים המקורי, כאשר בסיום החישוב נשמרים החיזויים הסופיים.
בנוסף, מתבצע חישוב דיוק של המסווגים שחושבו עד כה בכל איתרציה, ואם קבוצת המסווגים באיתרציה כלשהי השיגה את הדיוק הרצוי שהגדרנו הלולאה תעצור והדיוק ישמר.

פונקציית ה – main –
בפונקציה יש שורה לקבלת הקובץ וקריאתו (נדרש כמובן מיקום יחסי במחשב, במקרה שלנו בסיסי הנתונים נמצאים בתיקייה יחד עם AdaBoost.py ו – Main.py). השתמשנו ב – pandas על מנת לקרוא את קבצי ה – csv.
לאחר מכן נתנו שם label לעמודת התיוגים לשם ההפרדה בין העמודה לבסיס הנתונים לאחר מכן.
בשורות 17 ו – 18 ניתן לבחור את מספר המסווגים שנרצה לייצר (T), ואת אחוז הדיוק הרצוי בו נרצה לעצור.
בשורות 19 ו – 20 ניתן לבחור את הגדלים של בסיס הנתונים ללמידה ובסיס הנתונים לבדיקה. בשורה 21 ניתן לבחור האם בבדיקה להשתמש בנתונים מבסיס הנתונים ללמידה או לא.
השתמשנו ב – plot על מנת להציג גרף המראה את רמת הדיוק ביחס למספר המסווגים שהגדרנו מראש. לבסוף אנחנו מדפיסים את טבלת החיזויים של הבדיקה, את אחוזי הדיוק של הלמידה והבדיקה אליהם הגענו ואת מספר המסווגים שנדרשו לכל אחד מהם.

שינויים בקוד:

- התבצעו מספר שינויים בקוד בהגשה הסופית של העבודה, השינויים לא שינו את מהות הקוד, את סדר החישובים בקוד או את החישובים עצמם.
  - התבצעו שינויים בנראות הקוד (גודל שורות, רווחים וכו').
  - התווספו מספר משתנים על מנת ליעל את בחירות הערכים שנשלחים לחישוב (דיוק החישוב ומספר המסווגים החלשים בהם השתמשנו פוצלו ל – 2 לבדיקה וללמידה בהתאמה, נוסף משתנה בוליאני לבחירה – האם להתמש לבדיקה בנתונים מהלמידה או להשתמש בנתונים שונים)
  - התווספו מספר חישובים קטנים שלא היו, על מנת להדפיס כמה נתונים נוספים בסוף הריצה שיראה יותר מסודר (חישוב הדיוק בלמידה ולא רק בבדיקה, חישוב תנאי העצירה גם בלמידה ולא רק בבדיקה)
  - ושינוי הייצוג של טבלת החיזויים (ממערך לטבלה עם מיספור – כדי שיהיה אפשר להשוות לנתונים בקווב לפי מספר שורה)