דוח פרויקט סוף – למידה חישובית

מגישים:

ניר ממן – 313446841

מאי משה - 208832873

קישור למחברת הפיתון: <https://drive.google.com/drive/u/0/folders/1JMZq32AenpbvHVNxSfmv3mNjYsvr5F3w>

על הדאטא:

מאגר המידע אותו בחרנו עוסק בחיזוי סימון שפת הסימנים. המידע מסווג ל24 מחלקות כאשר כל אחד מבטא אות שונה בשפה האנגלית (לא כולל J ו Z הכוללות "תנועות מחווה" כלומר הסימון שלהן כולל תנועה ולא תמונה אחידה).

במאגר המידע ישנם 7173 שורות אשר הן תמונות שונות עבור הסימונים השונים, כאשר כל אחת מייצגת אות שונה בשפה.

בעזרת עיבוד תמונה התמונות מומרות ל 28 x 28 פיקסלים, כך שיש 784 סוגים של פיקסלים. ערכי הפיקסלים נעים בין 0-255.

כל תמונה בדאטא נחתכה ב"אזור המעניין" של היד, צבע הרקע שונה לאפור בכדי ליצור אחידות, והוספו סוגים שונים של בהירות\ניגודיות ולבסוף גם סיבוב של 3 מעלות על מנת לשנות ביעלות את הרזולוציה וליצור דאטא נרחב יותר עם מחלקות מעניינות.

בטבלת האקסל נוכל לראות 7173 שורות ו785 עמודות, כלומר 7173 דגימות ו 784 סוגי פיקסלים עבור כל דגימה.

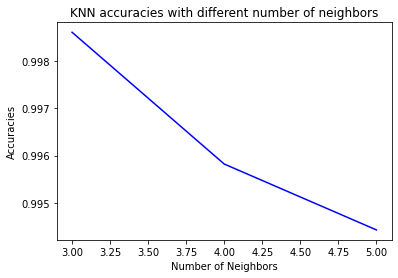
באמצעות חקר אלגוריתמי למידה נוכל לסייע לחירשים ולכבדי שמיעה לתקשר בצורה טובה יותר על ידי הנגשה של השפה לציבור כולו. באופן כזה יוכל כבד שמיעה לשוחח עם עמיתו לעבודה באמצעות כלי חישובי אשר יתרגם את הסימנים למילים ולמשפטים שלמים.

בנוסף, יוכלו כבדי שמיעה ברחבי העולם לעסוק במקצועות אשר עד היום נבצר מהם לעסוק בהם, כמו מגישי חדשות, שדרני רדיו, ובעצם כל מקצוע אשר נדרש דיבור שוטף.

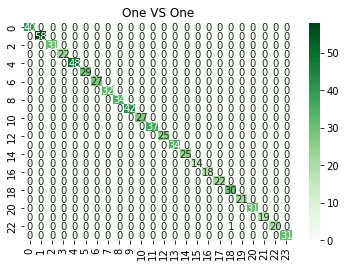
על הקוד:

במהלך הפרויקט בחרנו לעסוק בין היתר ב 3 סוגים של אלגוריתמי למידה:

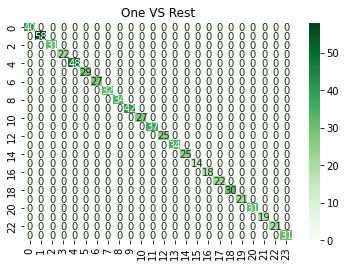
* K Nearest Neighbors – אלגוריתם זה מקבל קלט מסוים ומשייך אותו למחלקה הרלוונטית על בסיס שכנים קרובים. הקלט משויך למחלקה הנפוצה ביותר בקרב K השכנים הקרובים. בפרויקט שלנו רצינו להגדיר את K להיות כמספר התכונות (784) אך זמן הריצה היה ארוך מאוד ולכן עשינו הרצות עם מספר שכנים שונים והבנו כי ניתן להסיק מסקנות לגבי מספר השכנים עם 3 שכנים, 4 שכנים ו – 5 שכנים. קיבלנו אחוז הצלחה גבוה מאוד (0.9986). על פי הגרף הנ"ל ניתן לראות כי מספר השכנים האופטימלי שבאמצעותו ניתן לסווג את הדאטא בצורה הכי טובה הוא 4.

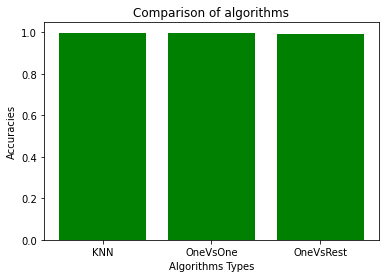


* One vs One – שיטה זו היא שיטה לסיווג הדאטא ל K מחלקות. בכל פעם האלגוריתם מסתכל על 2 מחלקות מתוך K ומנסה להפריד את הדאטא ל 2 מחלקות אלו. אלגוריתם זה עושה שימוש ביסודות הlogistic regression . כאשר הפעלנו את האלגוריתם על הדאטא קיבלנו אחוז הצלחה גבוה מאוד (0.9986). מפת החום הנ"ל מציגה לנו את הconfusion matrix של האלגוריתם. ניתן לראות כי מספר הטעויות הוא מאוד קטן אזי הוא אינו מתבלבל בין מחלקות. מכאן שאין אנו נמצאים במצב של overfitting כיוון שהאלגוריתם מצליח ללמוד על קבוצת הtrain ולחזות על קבוצת הtest עם אחוזי הצלחה גבוהים.



* One vs Rest – שיטה נוספת לסיווג הדאטא ל-K מחלקות. במודל זה האלגוריתם מסתכל על כל K המחלקות ובכל פעם מנסה להפריד מחלקה אחת מהשאר. גם בעת הפעלת אלגוריתם זה קיבלנו אחוזי הצלחה גבוהים מאוד (0.9944). מפת החום הנ"ל מציגה לנו את הconfusion matrix של האלגוריתם. גם במקרה זה אנחנו לא נמצאים במצב של overfitting כיוון שאחוזי הדיוק מאוד גבוהים והאלגוריתם כמעט ואינו מתבלבל בין מחלקות.



בגרף הבא ניתן לראות את ההשוואה שביצענו בין האלגוריתמים:

על פי הגרף ניתן לראות כי שלושת האלגוריתמים נותנים אחוזי דיוק גבוהים למדי ולכן קשה לראות איזה אלגוריתם מתעלה על כולם. לפי השוואה מספרית של אחוזי ההצלחה ניתן לראות כי שני האלגוריתמים – KNN ו – One Vs One הם בעלי אחוזי הצלחה זהים והגבוהים יותר מהאלגוריתם השלישי. לכן ניתן להסיק כי האלגוריתמים הנ"ל הם האופטימליים על הדאטא שלנו.

השוואה בין אלגוריתמים נוספים לבין KNN:

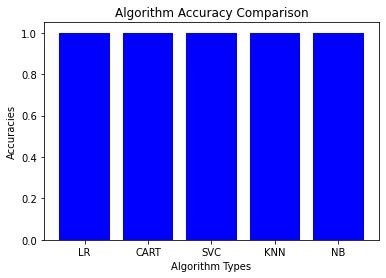
בחלקו האחרון של הפרויקט בחרנו להמשיך הלאה עם האלגוריתם KNN לבצע השוואות נוספות בין אלגוריתמים נוספים על הדאטא שלנו.

ביצענו השוואה בין 4 אלגוריתמים שונים:

* Logistic Regression
* Decision Tree Classifier
* Support Vector Clustering (SVC)
* Gussian Naive Bayes

לבין האלגוריתם KNN (עם פרמטרים שונים) על הדאטא שלנו בכדי להבין מהו האלגוריתם הכי אופטימלי ביניהם על הדאטא. ההשוואות התבצעו עם פרמטרים שמוגדרים כברירת מחדל של הספרייה Sklearn.

מדדנו על פי המדד Accuracy שמטרתו לעריך את אחוז הניחושים המוצלח שלנו. מדד זה נותן אינדיקציה טובה לגבי טיב האלגוריתם הנמדד.

על פי הגרף הנ"ל ניתן לראות את ההשוואה שביצענו בין האלגוריתמים וכי אחוזי ההצלחה של כולם היו מאוד גבוהים על הדאטה שלנו.

על פי הגרף קשה לראות מהם האלגוריתמים האופטימליים ביותר אך לפי השוואה מספרית הגענו למסקנה כי האלגוריתמים האופטימליים ביותר לשימוש על הדאטא שלנו הם: Logistic Regression, Gussian Naive Bayes.

מסקנות:

לאחר הרצת שלושת האלגוריתמים KNN, One Vs One, One Vs Rest, הגענו למסקנה כי עבור הדאטא שלנו האלגוריתמים האופטימליים לשימוש הם KNN ו – One Vs One, הם אלה שמביאים את אחוז הדיוק הגבוה ביותר. ניתן להבין כי הדאטא שבחרנו ניתן לסיווג ל – 24 מחלקות עם אחוזי דיוק גבוהים מאוד ונוכל להשתמש בו כדי לחזות אותיות בשפת הסימנים. בנוסף גילינו כי אלגוריתמים שונים מוצלחים במדד ה – Accuracy והאלגוריתמים האופטימליים ביותר לדאטא שלנו הם: Logistic Regression, Gussian Naive Bayes.