פרויקט למידת מכונה -זיהוי פריט לבוש בתמונה

מגיש:עוז קלינגל

תיאור של המאגר: הדטה סט מכיל תמונות של 10000 פרטי לבוש,1000 תמונות מכל פריט.

https://www.kaggle.com/zalando-research/fashionmnist

הבעיה : בהנתן תמונה כלשהי ,איזה סוג של פריט-לבוש מופיע בתמונה.

טכניקות בהם השתמשתי:

1.pca
2.knn
3. LogisticRegression
4 .DecisionTreeClassifier
5 .Random Forest

תיאור תהליך המחקר:

Pre-proccesing : ביצעתי סטנדרטיזציה של הדטה סט וחלוקה לרשומות אימון ורשומות מבחן

בחינת מודלים כוללת: הרצתי על הרשומות של האימון את המודלים לעיל (2-5) עם הפרמטרים הדיפולטיבים של פייטון.

התוצאות הראו ש knn ו LogisticRegression בעלות התוצאות הטובות ביותר עם דיוק העומד על כ 80-81 אחוז.

Dimentionality Reduction: כדי ליעל את המודל על ידי הורדת "רעש" ולהוריד את הסיבוכיות שלו ,השתמשתי בPCA על מנת ליצג את הDATA במרחב בעל מימד קטן יותר,אשר שומר את השונות הקיימת בDATA המקורי בצורה מקסימאלית.אחר PCA נשארו 243 פיצ'רים עיקריים.

בחינת מודלים כוללת שנייה: : הרצתי על הרשומות של האימון את המודלים לעיל (2-5) עם הפרמטרים הדיפולטיבים של פייטון. התוצאות הראו ש knn ו LogisticRegression בעלות התוצאות הטובות ביותר עם דיוק העומד על כ 82-83 אחוז.

Prameters tunning: בשלב זה בתמקדתי בבחינה של הפרמטרים בשני המודלים הטובים ביותר-KNN ו LR.

ב KNN עשיתי אופטימיזציה של P (המטריקה על פיה נמדד המרחק מנקודה אחת לשניה) עם הערכים P=1,2,infinity, ושל מספר השכנים (..) עם הערכים 1,10,20 .

התוצאה- כאשר p בעל ערך 2 ומספר השכנים הוא 20 התקבלת תוצאה אופטימאלית למודל עם דיוק של 82 אחוז.

ב LR עשיתי אופטימיזציה של C אשר בעת ריצת המודל "מעניש" אותו כאשר יש שונות גדולה.

התוצאה-כאשר C בעל ערך 0.01 מתקבלת תוצאה אופטימאלית למודל עם דיוק של 86.875 אחוז.

הרצה על רשומות המבחן: בהרצה של LR עם הפרמטרים האופטימאלים מתקבלת התוצאה הטובה ביותר שהגעתי אליה-85.25 אחוז דיוק .

הסברים:

-תמונות, ובמיוחד פרצופים נשענים במידה רבה על קשרים מקומיים בין תכונות (כלומר פיקסלים קרובים זה לזה). עצי

ההחלטה (גם עצים רנדומים) אינם לוקחים זאת בחשבון, ולכן התוצאות לא יכולות מדויקות, או עשויות להיות מושפעות

מאוד מרעש.כמו כן, עצים הם רבי עוצמה, אך בדרך כלל הם שימושיים כאשר הם תמציתיים זאת אומרת יש פיצ'רים רבי

משמעות אך לתמונות בהן הפיקסלים הם הפיצ'רים קשה לתפוס פיקסל שהוא יחשב כדומיננטי אלא צריך מקבץ של

פיקסלים שיחד הם משמעותיים.

-ב KNN שהורץ על המודל שלנו נמצא שבהנתן רשומה חדשה מומלץ להתחשב ב20 "נקודות" הכי קרובות אליה (במטריקה

אוקלידית) לייבלינג ע"פ הCLASS הכי מצוי ב20 נקודות אלו.
-ב LR הקצב למידה האופטימאלי יחסית קטן (0.01) אך לא מידי כי ב100 איטרציות הירידה של הגרדיאנט צריכה להיות יחסית מהירה כדי שתהיה משמעותית לכיוון הנקודת מינימום.

-הייצוג של התמונה במימד קטן יותר על ידי PCA שיפר משמעותית את התוצאות ואת סיבוכיות הריצה של המודל.להערכתי דבר זה קרה בשל הורדת ה"רעש" שהיה בתמונות.

אתגרים:

1.בבחינה של במודלים adaboost ו- SVM לא הגעתי לתוצאה משום שכל פעם שניסיתי להריץ אותם על הדטה סט המחשב החל להתחמם ולעשות רעש מטריד שלא חלף גם אחרי 5 דקות אז הפסקתיאת ההרצה והסתפקתי ב4 המודלים לעיל.

2.היה קשה לי להבין מה בדיוק הPCA עושה. בתחילה חשבתי שהוא מוריד מימדים אך אחר חזרה על החומר הבנתי שהוא לא "זורק" את הפיצ'רים החלשים לפח אלא פשוט נותן לאטה סט יצוג במרחב בעל מימד נמוך יותר עם מתן משקל חזק יותר לפיצ'רים החזקים.

3.קשה להבין מדוע בדיוק LR נתן את הדיוק המירבי.