עיבוד שפות טבעיות - תרגיל בית שלישי

פרטי הגשה

מגישים

אורי אריאל ניפומניאשצי	318261468
ישי גרוניך	208989186

מערכת הפעלה

Windows 7

שימוש בכלים של הספרייה scikit-learn

שימוש במסווגים

בספרייה scikit-learn ישנו מגוון של מסווגים, ובפרט ארבעת המסווגים SVM, Naive Bayes, Decision Tree, KNN. הסיפרייה תוכננה כך שכל המסווגים יממשו את הפונקציה `train` שמקבלת אוסף וקטורים והמחלקה אליה כל וקטור שייך, והפונקציה `predict` לאחר מכן שמקבלת אוסף וקטורים ומשערת מה המחלקה המתאימה, על בסיס החישובים השונים של כל מסווג. הממשק המשותף לכל המסווגים איפשר לנו לבנות פונקציה כללית בשם `build_classifier` שמקבלת מחלקה למשל `sklearn.svm.SVC`, מייצרת אובייקט מהמחלקה ומאמנת אותו על ידי הפונקציה train. כלומר, כל הקוד שלנו כללי לחלוטין וניתן להשתמש בו גם עם מסווגים אחרים.

בדיקת האיכות של מסווג

על מנת לבדוק את האיכות של המסווגים השונים, ביצענו tenfold cross validation באמצעות שני כלים של הספרייה scikit-learn. הפונקציה `cross_validation.cross_val_score` מקבלת מסווג, וקטורים, והתוצאות הרצויות שלהם, מבצעת עליהם cross validation ומחזיקה מערך של תוצאות, אחת לכל שלב בשיטת cross validation. לפונקציה זו היינו צריכים להעביר שני פרמטרים נוספים - האחד הוא העובדה שאנו מתעניינים בערך accuracy של המסווג, והשני הוא האופן בו לבצע cross validation. על מנת לבצע tenfold cross validation יצרנו אובייקט מטיפוס KFold, שיש להעביר לו את: מספר הדגימות, מספר folds רצויים (10), והעובדה שאנו מעוניינים בערבוב הוקטורים לפני תחילת הבדיקה. הקוד עבור חלק זה נראה כך:

```
tenfold_validation = cross_validation.KFold(arr_length(all_samples),
      n_folds=cross_val_folds, shuffle=True)
accuracies = cross_validation.cross_val_score(classifier, all_samples,
      expected_results, scoring='accuracy', cv=tenfold_validation)
```

שאלה 1

סעיף א

בחרו כמה עשרות (עד 50) מילים אשר עשויות לדעתכם לתפוס את ההבדלים בין ביקורות חיוביות לשליליות. תוכלו להעזר ברשימות מילים קיימות ברשת למטלות מסוג זה (שתמצאו בעצמכם). פרטו את הרשימה שבחרתם.

כתבנו שתי רשימות, האחת עם מילים העשויות להופיע הרבה בביקורות חיוביות, והשניה עם מילים העשויות להופיע הרבה בביקורות שליליות. חלק מהמילים נלקחו מהאתר הזה אך את רובן זיהינו תוך קריאה של חלק מביקורות הקלט.
להלן שתי רשימות המילים (סכום גדלי הרשימות הוא 50):

entertaining, fascinating, beautiful, powerful, fun, funny, recommend, recommended, brilliant, incredible, enjoyed, talented, colorful, best, amazing, wonderful, charming, touching, powerful, love, loved, interesting, excellent, again, great, favorite, perfect, perfectly

poor, poorly, terrible, terribly, worst, disappointing, disappointed, wasted, dislike, awful, boring, predictable, annoying, hate, hated, avoid, failed, stupid, bad, ruined, pathetic, horrible

סעיף ד

יש להעריך את הביצועים של כל מסווג בעזרת ten-fold-cross-validation ולדווח את הדיוק (accuracy) הממוצע של כל ה folds. פרטו את התוצאות שקבלתם ודונו בהם: האם הן כפי שציפיתם? איזה מסווג עבד טוב יותר מאחרים?

להלן התוצאות עבור סיווג באמצעות features שנבחרו ידנית:

```
- SVM: 0.7608793969849247
- Naive Bayes: 0.7638743718592964
- DecisionTree: 0.6883291457286432
- KNN: 0.6408291457286432
```

התוצאות הפתיעו אותנו בכמה מובנים:

- ציפינו לקבל תוצאות יותר גבוהות בכל המסווגים. חשבנו שחמישים מילים משמעותיות הם אוסף מספיק גדול על מנת להבחין בין ביקורות טובות לביקורות רעות. הופתענו לגלות שגם המסווג הטוב ביותר מבין ארבעתם טועה פעם בכל ארבע ביקורות. ייתכן שהסיבה היא שישנן הרבה ביקורות שלא מכילות אף מילה מבין 50 המילים שבחרנו, או שמכילות רק מילה אחת.
- במיוחד הפתיעה אותנו התוצאה הנמוכה של המסווג KNN. השיטה האינטואיטיבית שבה עובד המסווג גרמה לנו לחשוב שהוא יעבוד היטב. הסבר אפשרי לתוצאה המפתיעה הוא דלילות המידע - ייתכן שהמסווג KNN היה עובד טוב יותר אילו היו הרבה יותר ביקורות.

תוצאה שלא הפתיעה אותנו היא שהמסווג Naive Bayes הגיע לרמת accuracy הגבוהה ביותר מבין ארבעת המסווגים (עם תוצאה דומה לזו של SVM). בקורס "למידה ממוכנת" למדנו על המסווג Naive Bayes, ועל ההצלחה הנרחבת שיש לו בסיווג טקסטים, אף על פי ההנחות ההסתברויות שבהכרח אינן נכונות.

שאלה 2

סעיף א

מספר המילים השונות שיש בטקסטים, ושאינן stop words על פי ההגדרה של scikit, הוא 22878.

סעיף ג

בנו feature vectors מהטקסטים (למשל בעזרת CountVectorizer). סווגו כעת את ה feature vectors בעזרת המסווגים מסעיף 1.ג. העריכו את ביצועי המסווגים בעזרת ten-fold-cross-validation כפי שעשיתם ב 1.ד. האם תוצאות טובות יותר כעת? דונו בהבדלים מהסעיף הקודם (1.ד.).

להלן התוצאות עבור סיווג באמצעות Bag Of Words:

```
- SVM: 0.46824874371859293
- Naive Bayes: 0.8969447236180905
- DecisionTree: 0.6953417085427136
- KNN: 0.8594497487437185
```

השוואת התוצאות לתוצאות עם features ידניים:

- באופן מפתיע, תוצאות המסווג SVM ירדו משמעותית. ערך הדיוק 0.47 מצביע על כשלון מוחלט של המסווג - כרמת הדיוק של ניחוש אקראי. ניסינו לבדוק מדוע זה קורה, וגילינו שהמסווג החזיר את אותה התשובה על כל הדגימות. אנו מנחשים שהמקור לבעייה זו הוא חוסר יכולת לבצע הפרדה לינארית של המידע לאחר ההטלה באמצעות kernel.
- רמת accuracy של Naive Bayes השתפרה בצורה ניכרת. אחוז הצלחה של כתשעים אחוזים היה הציפייה הראשונה שלנו כשהתחלנו לעבוד על תרגיל הבית. ניתן להבין את ההשתפרות המשמעותית של מסווג זה מאחר שהוא כעת יכול לקחת בחשבון הרבה יותר מידע בחישוב ההסתברותי ולהתחשב גם במילים שהן נדירות אז תורמות לסיווג.
- לא הבחנו בהבדלים משמעותיים בתוצאות של המסווג Decision Tree. מאחר שהעץ נבנה באופן כזה שהעומק שלו יהיה נמוך, סביר שהוא ישתמש רק בחלק מהמידע גם אם נספק לו וקטורי מידע מלאים. סיבה אפשרית נוספת לכך שהמסווג Decision Tree הוא לאו דווקא יותר טוב כאשר משתמשים במידע Bag Of Words מלא - הוא עלול ליצור overfitting.
- הביצועים של המסווג KNN השתפרו משמעותית. כנראה שרשימה מוגבלת של 50 מילים אינה מספיקה על מנת שהמסווג KNN יפעל היטב מאחר שישנן ביקורות שהחיתוך של אוצר המילים שלהן, עם רשימת המילים, היא קטנה. ייתכן גם שהביצועים הנמוכים של KNN בסעיף הקודם קשורים לעובדה שהאינדיקציה בסעיף הקודם היתה בינארית ולכן המרחק האוקלידי בין שני feature vectors הוא פחות אינפורמטיבי מאשר בסעיף זה.

שאלה 3

סעיף א

השתמשו ב SelectKBest על מנת לבחור את 50 המילים בעלות התרומה הגבוהה ביותר לסיווג. התרשמו מרשימת המילים והשוו אותה לרשימה שבחרתם באופן ידני בסעיף 1.א. האם כל המילים שקבלתם כעת הן צפויות? פרטו את התוצאות (50 המילים).

המילים שהתקבלו הן:

amazing, annoying, avoid, awful, bad, badly, beautiful, best, boring, brilliant, effects, excellent, great, highly, hitchcock, horrible, hour, idea, just, lame, life, like, lives, looks, love, loved, make, masterpiece, minutes, money, mother, perfect, performance, plot, poor, poorly, portman, ridiculous, script, strong, stupid, superb, terrible, thing, war, waste, wasted, wonderful, worse, worst

רוב המילים שהתקבלו הן אכן צפויות. בפרט, 22 מהמילים הנ"ל הן מילים אשר בחרנו בתחילת התרגיל כמילים אשר לדעתנו תעזורנה להבדיל בין ביקורות טובות לרעות. לעומת זאת, ישנן מילים שמפתיע למצוא ברשימה הנ"ל.

למשל, ברשימה ישנם שני שמות של שחקנים, hitchcock וגם portman. לכאורה אילו מילים נטרליות לחלוטין, אך בפועל אנשים מזכירים שחקנים ספציפים בסרטים (ובמיוחד שחקנים מפורסמים) כאשר הם מתרשמים מביצועיהם.

מילים נוספות שהופתענו לגלות ברשימה הנ"ל: thing, war, idea, money, effects, minutes.

לגבי חלק מהמילים ניתן בכל זאת לנחש באיזה סוג של ביקורות הן מופיעות - למשל, ייתכן שהמילה money מופיעה בביקורות של אנשים שמעוניינים לקבל את כספם חזרה, או שחבל להם על ביזבוז הכסף בצפייה בסרט, או שהם חושבים שלסרט לא היה תקציב. בכל אופן, העיסוק בכסף בביקורת על סרט (אלא כחלק מעלילת הסרט) כנראה מרמז על קונוטציה שלילית.

לעומת זאת, קשה לנו לנחש כיצד המילים war, thing מבדילות בין ביקורת חיובית לביקורת שלילית.

שאלה 4

סעיף ב

חיזרו על סעיף 2 (עם כל תת-סעיפיו) כאשר כעת אתם משתמשים ברשימת מילים סגורה ל CountVectorizer (יש לו אופציה לקבל מילון, ראו תיעוד). השתמשו ברשימת 50 המילים בעלות התרומה הגבוהה ביותר לסיווג מסעיף 3.א. השוו את התוצאות למספרים שקבלתם עבור 2.ג. (עם bag-of-words). האם ההבדלים משמעותיים?

להלן התוצאות עבור סיווג באמצעות המילים המשמעותיות מהסעיף הקודם:

```
- SVM: 0.7944045226130653
- Naive Bayes: 0.8084070351758795
- DecisionTree: 0.693856783919598
- KNN: 0.7593793969849246
```

ההבדלים בין סיווג זה לבין הסיווג עם Bag Of Words הם אכן הבדלים משמעותיים. התוצאות מסעיף זה דומות יותר לתוצאות שקיבלנו עם אוצר המילים שכתבנו בעצמנו, אם כי הן בכל זאת מעט יותר טובות.

- ניתן לראות שהתוצאות של המסווג SVM הן גבוהות, כמו בסעיף הראשון, ובשונה מהסעיף הקודם. הסיבה לכך היא כנראה המילון הקטן, בדומה למילון שכתבנו בעצמנו.
- המסווג Naive Bayes השתפר מעט לעומת הסעיף הראשון, אבל ביצועיו אינם קרובים לביצועים שלו עם Bag Of Words מלא. ניתן לייחס זאת לעובדה שהמילים שהתקבלו הן אכן יותר שימושיות לצורך הסיווג, אבל מספר המילים הקטן בכל זאת מגביל את המסווג.
- המסווג Decision Tree לא הושפע מאוד מהשינוי באוצר המילים, וניתן היה לחזות זאת מאחר שביצועיו לא השתנו הרבה גם בין הסעיפים הקודמים.
- התוצאה של המסווג KNN היא מעניינת. למרות שחל בו ירידה ברמת accuracy, היא עדיין גבוהה בהרבה מאשר רמת accuracy בסעיף הראשון. יכולות להיות לכך שתי סיבות - האחת, כפי שנאמר בסעיף הראשון, שכעת הוקטורים אינם בינארים, ולכן הם יותר אינפורמטיבים בחישוב מרחק אוקלידי; והשנייה, שהמילים שנבחרו על ידי אלגוריתם SelectKBest נמצות ביותר ביקורות מאשר המילים שאנחנו בחרנו ידנית, ולכן ישנן ביקורות שהוקטור המתאר אותן הוא כעת הרבה יותר אינפורמטיבי.