בס"ד

עיבוד שפות טבעיות – דו"ח סיווג טקסטים

מגישים:

פורטנוי ליאור

אפראימוב אורן

מערכת הפעלה:

נבדק על כמה מערכות הפעלה: windows 7, windows 10 ו-Ubuntu 15.4.
התוכנית רצה בסביבות 39 דקות, מתוכם SVM שאלה 2 רץ סביבות 31 דקות!

סיווג ו"ערבוב" המסווג:

הסיווג שאנו בנינו, נעשה על רשימה של וקטורים מעורבבת, כך ששלב הtest ו-learn יעשו על שני המחלקות ולא רק על מחלקה אחד!
בנוסף, כדי לשמור על "הוגנות" ועל מנת שנוכל להשוות בין שני המסווגים, שמרנו על אותו סדר של ווקטורים, כך שנוכל באמת לראות את ההבדלים.
לצערנו, בגלל מבנה הווקטורים, היה לנו קשה להשתמש ב-shuffle, היו שגיאות די רבות, שלא ידעתנו מה לעשות ולכן בחרנו בשיטה הידנית!

הערה חשובה: לאחר התייעצות עם קבוצות אחרות, והשוואת תוצאות המסווגים בשאלה, אנו חושבים שתוצאות המסווגים Decision-Tree ו-knn שונים בשאלה 2, עקב עוצמת החישוב של המחשב, בשאלה 4 קיבלנו אותם תוצאות.

בס"ד

שאלה 1 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.

תשובה:

לינק לרשימת המילים (לחץ עליי)

רשימת המילים "החיוביות" (בסך הכול 24 מילים):

first-rate, insightful, clever, charming, comical, charismatic, enjoyable, uproarious, original, tender, hilarious, absorbing, sensitive, riveting, intriguing, powerful, fascinating ,pleasant, surprising, dazzling, thought provoking, imaginative, legendary, unpretentious

רשימת המילים "השליליות" (בסך הכול 26 מילים):

violent, moronic, third rate, flawed, juvenile, boring, distasteful, ordinary, disgusting, senseless, static, brutal, confused, disappointing, bloody, silly, tired, predictable, stupid ,uninteresting, weak, trite, uneven, outdated, dreadful, bland

בחירת המילים נעשתה על ידי חיפוש בגוגל, ללא התעמקות ברשימה זו או אחרת. למרות, שאחוזי הזיהוי לא מאוד משמעותיים, והיה ניתן לשפר בעזרת חיפוש רשימה טובה יותר, או על ידי הפונקציה SelectKBest, בחרנו להשאיר אותם על מנת "לחוש" את טיב המסווגים, אפילו על מלים לא כל כך משמעותיות!

שאלה 1 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?

תשובה:

~~~~Question 1~~~~

SVM classifier - the accuracy is of is 0.584500 it's take 5.9584 sec

Navie-Bayes classifier - the accuracy is of is 0.576000 it's take 0.6446 sec

Decision-Tree classifier - the accuracy is of is 0.621500 it's take 0.6019 sec

KNN classifier - the accuracy is of is 0.559500 it's take 1.2512 sec

ערכת התוצאות:

- באופן כללי, וזאת לאחר סעיף ג', התוצאות לא הפתיעו אותנו, שכן מילות הסיווג שנבחרו לא היו בעלות משמעויות, ולכן ערכת המסווגים כה נמוכה!
- ארבעת המסווגים, הגיעו לאותו תוצאה, לכן קשה לנו לעמוד את טיב המסווגים.

בגלל, שמבט לאחור המילים שנבחרו לא היו משמעותיים, קשה לנו להבחין ולעמוד את ארבעת המסווגים, כי כולם השיגו אותו תוצאה, מלבד decision tree, שהשיגה שיפור קטן יחסית משאר המסווגים, אך לא באופן משמעותי.
~~~~

שאלה 2 א':

בנו feature vectors מהטקסטים למשל בעזרת (CountVectorizer).
תשתמשו **בכל המילים של הטקסט** (לא מילון מוגבל), **ללא stop words** של אנגלית. המשיכו וחשבו שכיחויות (לא אינדיקציה בינארית) של כל המילים (השתמשו ב tf-idf) בעזרת TfidfTransformer. כמה מילים שונות ישנן בטקסטים (במילים אחרות, מה אורך ה feature vectors שנוצרו)?

תשובה:

נוצרו 22,878 מילים (טוקנים) שונים זה מזה.
צריך להשים לב, שלפי הגדרת התרגיל, לא התחשבנו במילים שהם stop words.

שאלה 2 ג':

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.).

תשובה:

~~~~Question 2~~~~
SVM classifier - the accuracy is of is 0.508500 it's take 30.8459 min
Navie-Bayes classifier - the accuracy is of is 0.781000 it's take 0.1330 min
Decision-Tree classifier - the accuracy is of is 0.696500 it's take 2.5994 min
KNN classifier - the accuracy is of is 0.637000 it's take 4.5206 min

תשובה:

פרט למסווג SVM, תוצאות כל המסווגים ישתפרו ובחלק אף שיפור ניכר (Navie-Bayes).

<u>ערכת התוצאות:</u>

- באופן לא מפתיע לדעתנו, תוצאת המסווג SVM קרובה מאוד לאחוז דיוק 50%. דבר זה מעיד, שמסווג לא יידע לסווג בצורה טובה ואולי החליט בצורה רנדומלית מה ערך הסיווג. אנו חושבים, שמקור הבעיה היא עודף פ'יצרים ובגלל זה, היה למסווג קשה להבדיל בין ביקורת "חיובית" ל"שלילית", שכן מעודף "פ'יצרים" לא יכול לזהות בין הביקורות ולכן החליט בצורה רנדומלית.
~~~~

- רמת הסיווג של המסווג Navie bayes, השתפרה בצורה ניכרת מסעיף הקודם. לדעתנו, ההשתפרות הניכרת שלו היא שכעת הוא יכול לקחת בחשבון הרבה יותר מידע בחישוב ההסתברותי שלו ולהתחשב במילים שלא מופיעות באופן די נפוץ וכך לשפר את אחוזי הדיוק שלו.
- רמת הסיווג של המסווג Decision tree השתפרה, אך לא בשיפור די גדול. לדעתנו, בגלל שאלגוריתם בונה עץ בצורה כזאת שעומקו יהיה די נמוך, אנו בטוחים אפילו אם נשלח לו את כל הווקטור (22878 פיצרים), הוא ישתמש בחלק מן המידע לפי היוריסטיקות שלו, שכן הבעיה היא NP.
דבר נוסף, לפי ווקיפדיה, מתבצע תהליך "גיזום" על מנת להתמודד עם בעיית תאימות היתר (overfitting), יכול להיות שאלגוריתם שחברנו לא עושה את זה, או לא עושה את זה בצורה טובה בגלל כמות הפיצרים.
- רמת הסיווג של המסווג knn השפרה בצורה די קטנה, וזה באופן מוזר למה שחשבנו שנקבל. לדעתנו, היות ובסעיף הקודם נבחרו רק 50 מילים, זה לא היה מספיק למסווג, שכן יש ביקורות שחיתוך שלהם עם המילון שבחרנו די קטן. לכן, בגלל שכעת המילון מכיל את כל המילים, בגלל צורת החישוב של המסווג, ביקרות כאלו יזכו למענה ולכן זה התבטא גם באחוזי הסיווג.

<u>שאלה 3 א':</u>

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).

<u>תשובה:</u>

']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 אכן היו צפויות, לאדם אשר מבין ויודע איך נראים ביקורות, ואפילו לאדם עם מעט היגיון בריא.
רק חלק מן המילים שבחרנו באופן ידני מופיעות פה.

אך, יש מילים כמו, portman או 'hitchcock, שמסמלות את שמות השחקנים, והיה די קשה לפני הסיווג לעלות עליהם. זה מאוד הגיוני ששמות של שחקנים יעלו בכתיבת הביקורת, שכן אנשים נוהגים לציין את השחקן/ית שאהבו וכך לצרף לביקורת. אך, מילים אלו קשורות לסרטים מסויימים ולכן נראה שאולי רוב הביקורות שנבחרו, הגיעו מאותו סרט או הגיעו מסרטים בהם שני השחקים הופיעו.

מנגד, יש את המילים האלו, שרק אדם עם ידע בסיווג ביקורות היה יכול לחשוב: hour, war,minutes , thing וכד'.
לדעתנו, קשה לנו להבין כיצד מילים אלו תורמות להבדל בין ביקורת חיובית לשלילית.

לכן, אפשר להגיד שאת רוב המילים היו צפוי לדעת שייבחרו שכן הן מסמנים ביקורת "חיובית" או "שלילית", אך יש כמה מילים שהופיעו בגלל שם הסרט, שחקנים או כי הם נפוצות בביקורות האלו שלא ממש תרמו לסיווג!

שאלה 4 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.
השווה את התוצאות למספרים שקבלתם עבור 2.ג.עם(bag-of-words). האם ההבדלים משמעותיים?

תשובה:

~~~~Question 4~~~~
SVM classifier - the accuracy is of is 0.795500 it's take 5.2653 sec
Navie-Bayes classifier - the accuracy is of is 0.805000 it's take 0.1098 sec
Decision-Tree classifier - the accuracy is of is 0.687500 it's take 0.4597 sec
KNN classifier - the accuracy is of is 0.743000 it's take 1.1882 sec

ההבדלים אכן ניכרים בין התוצאות שקיבלו עכשיו לאלו שהיו בסעיף 2 ג', אך השיפור המשמעותי ביותר הינו בשתי מסווגים שאחוזי הדיוק השתפרו בצורה ניכרת.
המסווגים האלו הם svm ו-knn, בשאר הסווגים, השיפור יחסית זניח.

- רמת הסיווג של המסווג SVM השתפרה בצורה ניכרת, זאת בגלל העובדה שציינו קודם, שכעת המילון קטן, וכך בגלל שמילון מספק פיצרים איכותיים, המסווג יכל לסווג בצורה טובה, ללא צורך לנחש, בין ביקורת חיובית לשלילית.
- רמת הסיווג של המסווג knn השתפרה בצורה ניכרת, לדעתנו זה בדיוק כמו הסיבה ש-SVM השתפר, וזאת בגלל המילון שנבחר קטן מספיק.
לכן, חישוב המרחק האוקלידי, אכן יותר משמעותי על מיליון אשר מספק הבחנה טובה בין 2 סוגי הביקורת ולכן לדעתנו היה שיפור יחסית די ניכר.

בחרנו להתעלם משווה של שאר המסווגים, שכן בגלל שאחוזי הזיהוי יחסית דומים, לא היה לנו מה להוסיף!
~~~~