Modelling

חיזוי התפלגות ההצבעות וחיזוי המפלגה המנצחת

בחירת המודל

ראשית נציין כי ניתן להסיק מי המפלגה מנצח מתוך התפלגות ההצבעות בין המלפגות, זו כמובן המפלגה בעלת אחוז הקולות הגבוה ביותר. (אחרת, לפחות אחד מהחיזויים שגוי)

המודל הפשוט וכנראה המדוייק ביותר לחיזוי התפלגות הוא בדיקת יחס ההצבעות בדוגמאות הנתונות.

נכונות המודל נובעת מכך שאנו מניחים כי התכונות של הדוגמאות התוייגות מייצגות את התפלגות התפלגות התכונות של כלל האוכלוסיה, ולכן ההצבעות שלהם מייצגות את התפלגות ההצבעות של כלל האוכלוסיה.

למשל, אם הנחה זו לא היתה נכונה, והיה בידנו סט מתוייג אשר לא מייצג וסט מייצג שהוא לא מתוייג אז על מנת להשיג את התפלגות ההצבעות היינו מאמנים מסווג על הסט המתוייג, בעזרתו מתייגים את הסט המייצג ומוציאים ממנו את התפלגות ההצבעות.

תוצאות

```
##########################################################################################
# Predict distribution and winner
##########################################################################################
The predicted distribution:
{
'Greys': 0.0626,
'Purples': 0.255,
'Yellows': 0.0581,
'Pinks': 0.0922,
'Browns': 0.1718,
'Whites': 0.0324,
'Blues': 0.1003,
'Oranges': 0.0487,
'Greens': 0.1068,
'Turquoises': 0.0405,
'Reds': 0.0316
}
The predicted winner: Purples
##########################################################################################
```

חיזוי הצבעות

בחירת המודל

לשם נוחות נשתמש במונח דיוק עבור precision.

המודלים אותם בחרנו

קודם כל ביצענו כיוונון פרמטרים עבור המסווגים: id3, knn, forest, svm.

מהמודלים האלו יצרנו מודל מורכב שהוא ועדה (voting) והפרמטרים שכיוונו הם הרכב הועדה וגודלה.

באופן זה אנו מחפשים היפוטזה מתאימה במרחב גדול ומגוון של מסווגים אפשריים.

אופן מדידת ביצועי המסווגים בתהליך התבצע באופן זהה כפי שמותאר בהמשך.

הערכת הביצועים של המודלים

מכיוון שחיזוי ההצבעות משמש את המפלגות לביצוע שירותי הסעות, נרצה כי הדיוק לכל מפלגה יהיה גבוה. כלומר, למזער את הסעות השוא עבור כל מפלגה.

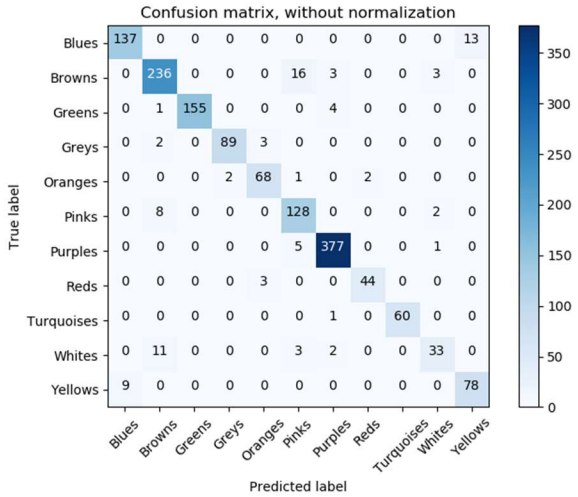
על מנת לבצע זאת באופן שיוויוני, כך שמספר הסעות השווא של כל מפלגה יהיה יחסי לגודלה, מדדנו את איכות המסווגים ע"פ הדיוק המינימלי שהתקבל מכל אחת מהמפלגות (עמודות במטריצת הבלבול).

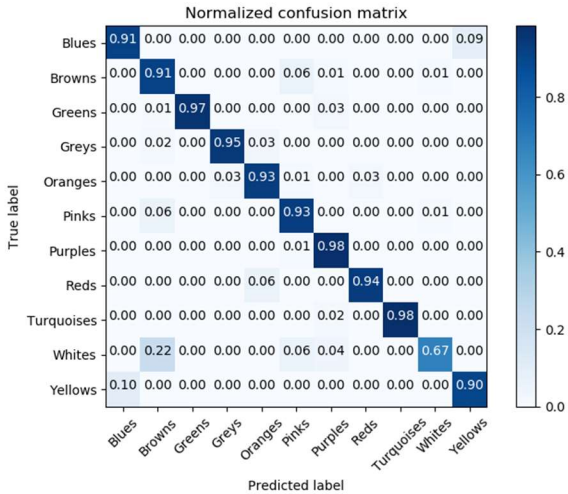
אופן בחירת המודל המנצח

בחרנו את המודל בעל הביצועים הטובים ביותר כמו שתואר מעלה, ובנוסף בדקנו כי למסווג זה ביצועים טובים גם בפרמטרים אחרים ע"פ מטריצת הבלבול שהתקבלה.

תוצאות

מטריצת הבלבול





השגיאה

```
Error: 0.06333333333333335
```

בונוסים

A

מתבצע על פי התהליך שהוסבר לעיל וממומש ומתועד בקובץ המצורף בסעיף 3 בקבצים.

B

השתמשנו בשני מודלים ע"מ לחזות את ההתפלגות (וממנה להסיק את המנצח)

- מודל א: כפי שתואר מעלה
- מודל ב: שימוש במסווג שנבחר על מנת לחזות את ההצבעות, ובדיקת התפלגות ההצבעות בקבוצת המבחן.

התוצאות השונות שקיבלנו:

```
##########################################################################################
# Predict distribution and winner
##########################################################################################
The predicted distribution:
{
'Greys': 0.0626,
'Purples': 0.255,
'Yellows': 0.0581,
'Pinks': 0.0922,
'Browns': 0.1718,
'Whites': 0.0324,
'Blues': 0.1003,
'Oranges': 0.0487,
'Greens': 0.1068,
'Turquoises': 0.0405,
'Reds': 0.0316
}
The predicted winner: Purples
##########################################################################################
# Using the selected model to predict distribution
##########################################################################################
The predicted distribution:
{
'Greys': 0.061,
'Greens': 0.103,
'Purples': 0.258,
'Yellows': 0.061,
'Pinks': 0.102,
'Browns': 0.172,
'Oranges': 0.049,
'Reds': 0.031,
'Blues': 0.097,
'Turquoises': 0.04,
'Whites': 0.026
}
Using predict votes classifier to predict distribution accuracy: 0.936
##########################################################################################
```

כפי שניתן לראות אין שינוי משמעותי בין ההתפלגויות השונות והמנצחת היא אותה המפלגה.

לכן נסיק כי מודל א' עדיף מכיוון שהוא פשוט יותר והשגיאה בו קטנה ותלויה אך ורק בשגיאת איסוף הנתונים ובהנחה שדוגמאות אלו אכן מייצגות את הכלל האוכלסיה, ושגיאות אלו משפיעות גם מודל ב' וזאת בנוסף לשגיאה בזיהוי שהיא 0.064

הסקריפט ממומש ומתועד בסוף הקובץ המצורף בסעיף 3 בקבצים.

C

רצינו לבדוק איזה שינוי של תכונה או קבוצת תכונות יגרום לשינוי תוצאות הבחירות.

לצורך כך חילקנו את התכונות ל-3 קבוצות:

- תכונות המתפלגות בין 1- ל-1 (נכנה אותן מתוחמות).
- תכונות המתפלגות בין מינוס אינסוף לאינסוף עם תוחלת אפס (נכנה אותן נורמליות).
- תכונות המקבלות או את הערך 1 או את הערך 1- (נכנה אותן בינאריות).

יצרנו כל תת קבוצה אפשרית של התכונות בגדלים 1 עד 3 וביצענו מניפוליות על קבוצת המבחן כדי לבדוק את ההשפעה על הבחירות.

מניפולציה על תכונה יכולה להיות העלאת או הורדת ערכה לפי סוגה:

- עבור תכונות מתוחמות – רצינו להשאר בתחום הקבוע לכן עבור מצביע מסוים ביצענו ממוצע של ערך התכונה עבורו ו-1 (או עם 1- עבור הורדת ערך תכונה).

- עבור תכונות נורמליות – הוספנו או החסרנו 0.5 בהתאמה עבור העלאה והורדה מערך התכונה עבור הבוחר.
- עבור בינאריות – על מנת להעלות תכונה בינארית נרצה לשים בה ערך אחד וערך מינוס אחד בכל התכונות שלא ייתכן שיחזיקו גם בערך 1 (כיוון שבפועל הן מייצגות את אותה התכונה – למשל, את התכונה MostImportantIssue פיצלנו למספר תכונות שונות שכל אחת מייצגת תשובה אפשרית, כיוון שרק תשובה אחת אפשרית לא ייתכן שלשתי תכונות יהיה ערך 1). על מנת להוריד ערך תכונה נשים 1- עבור המצביע ונגריל תכונה אחרת מהתכונות האפשריות שתקבל ערך 1. עבור העלאה והורדה בהתאמה ביצענו פעולות אלו עבור כל מצביע בהסתברות 0.8, כלומר, בהנתן שנרצה להעלות נבצע בהסתברות 0.8 את הפעולה הראשונה ובהנתן שנרצה להוריד נבצע בהסתברות 0.8 את הפעולה השניה.

יצרנו שני עותקים של קבוצת המבחן, אחד מיועד להעלאת ערכים והשני מיועד להורדת ערכים ועבור כל קבוצת ערכים ביצענו העלאה והורדה עבור כל עותק מתאים, בדקנו מחדש את תוצאות הבחירות ומצאנו מספר שינויים שישפיעו על הבחירות, דוגמאות נבחרות מתוך אלו שקיבלנו: (שאר השינויים בקובץ הפלט)

```
##########################################################################################################
# Find Factor - most likely to change which party will win the elections
##########################################################################################################
Features:
 *  Number_of_valued_Kneset_members
Were scaled DOWN and the new winner prediction is Browns
-----------------------------------------------------------------------
Features:
 *  Will_vote_only_large_party
Were scaled UP and the new winner prediction is Oranges
-----------------------------------------------------------------------
Features:
 *  Number_of_valued_Kneset_members
 *  Garden_sqr_meter_per_person_in_residancy_area
Were scaled DOWN and the new winner prediction is Pinks
-----------------------------------------------------------------------
```

קבצים

1. The Python script file that implements the data preparation: \prepossessing\preprocessing.py
2. CSV files of the prepared train, validation and test data sets: \csv_dada_sets\
3. The Python script file that implements the modelling: \models_selection\models_selection.py
4. A CSV file that contain the voting predictions: \csv_dada_sets\test_predicted_df.csv
5. A short documentation: this file 😊

בנוסף צירפנו את קובץ הפלט של models_selection.py המכיל גם את הבונוסים: models_selection_output.txt

מגישים

אביאל שמחי, 305376063

עמית סולומון, 305025785