



5

68

30/06/19

עידו יחזקאל

אמיר אביבי

idoye

saavivi

2 0 4 3 9 7 3 6 8

3 0 5 1 8 3 8 7 3

מבוא למערכות לומדות

236756

סמסטר אביב תשע"ט

Mandatory Part – Loading and Preparing the Data:

ראשית כמתבקש טענו את הקובץ הדוגמאות שקיבלנו בתחילת הקורס שוב, והחלנו עליו את מניפולציות העיבוד המקדים להכנת המידע כפי שביצענו בתרגילים הקודמים, מניפולציות אלה כוללות:

1. חלוקת סט הדוגמאות הכולל ( על די Stratified Shuffle Split) לשני סטים:

|  |  |
| --- | --- |
| Percentage from Original Data Set | Data Set |
| 80% | Train set |
| 20% | Test set |

1. הוצאות ערכים שהינם outliers, לדוגמא ערכים שליליים.
2. השלמת ערכים חסרים לפי השיטות המקובלות: feature correlation, mean and majority.
3. בחירת סט הפיצ'רים הנכון כפי שנבחר בתרגיל מספר 3.
4. ביצוע נורמליזציה לערכים קטיגוריאליים וZ-scale לערכים נומינליים.
5. ייצוא המידע ל 2X2 קבצי CSV לפני ואחרי השנויים.

כחלק מהתאמת סט המבחן החדש בתרגיל זה ביצענו את אותן מניפולציות שביצענו על סט האימון בדיוק גם על סט המבחן וזאת על מנת שהמסווג שלנו יתמודד עם סט המבחן כפי שביצענו לאורך כל הסמסטר.

לאחר מכן, ניגשנו למשימת החיזוי כאשר בחלק החובה נדרשנו לחזות:

* מה היא המפלגה המנצחת לפי סט המבחן.
* מה הוא פילוח הקולות לפי סט המבחן.
* עבור כל מצביע מסט המבחן לחזות את הצבעתו.
* לחזות קואליציה יציבה והומוגנית הכוללת לפחות 51% מסך כל הקולות.

Mandatory Part – Voting Predictions:

כדי לבצע את שלושת התחזיות הרכבנו עטפנו 3 מסווגים שונים אשר הניבו מדד דיוק הכי טוב מתרגלי בית קודמים, המסווגים הם:

1. RandomForestClassifier**(**random\_state**=**0**,** criterion**=**'entropy'**,** min\_samples\_split**=**3**,** min\_samples\_leaf**=**1**,** n\_estimators**=**450**)**
2. MLPClassifier**(**hidden\_layer\_sizes**=(**100**,** 10**),**activation**=**'relu'**,**

solver**=**'lbfgs'**,** alpha**=**0.0001**,**batch\_size**=**'auto'**,**

learning\_rate**=**'adaptive'**,** learning\_rate\_init**=**0.001**,**

power\_t**=**0.5**,** max\_iter**=**1000**,** shuffle**=True,**

random\_state**=**0**,** tol**=**0.0001**,**

verbose**=False,** warm\_start**=False,**

momentum**=**0.9**,** nesterovs\_momentum**=True,**

early\_stopping**=True,** validation\_fraction**=**0.1**,**

beta\_1**=**0.9**,** beta\_2**=**0.999**,**

epsilon**=**1e-08**,** n\_iter\_no\_change**=**10**)**

1. RandomForestClassifier**(**random\_state**=**0**,** criterion**=**'gini'**,** min\_samples\_split**=**3**,** min\_samples\_leaf**=**1**,** n\_estimators**=**450**)**

את החלטת המסווג הכולל מימשנו על פי החלטת הרוב וכאשר כל אחד מהמסווגים סיווג בצורה שונה את הדוגמא אזי נלקח הסיווג של מסווג מספר 1 אשר בעל הדיוק המקסימלי.

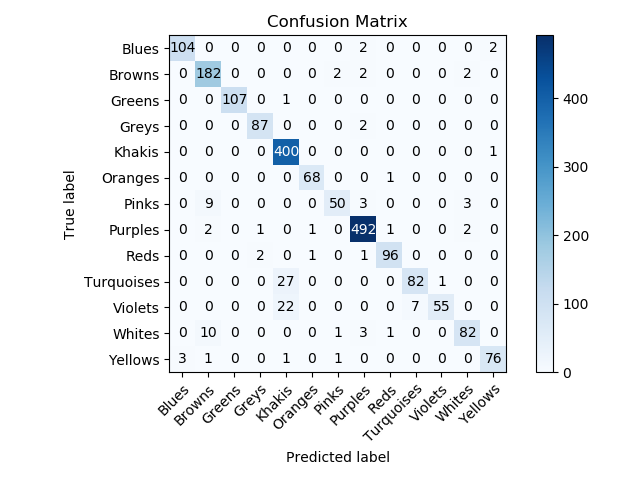
**השוואת תוצאות המסווג על סט מבחן מתויג מתוך סט האימון אל מול סט המבחן לא מתויג:**

**סט מבחן מתוייג:**

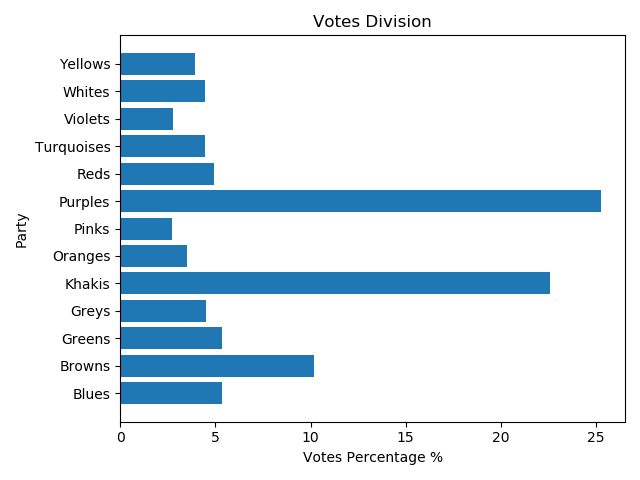
עבור סט מבחן מתויג אשר המסווג הכולל לא האתמן עליו כלל ולא הסתמך עליו כלל הצלחנו להגיע לאחוז דיוק של 94.05% .  
על מנת לחזות את המפלגת המנצחת לפי סט זה חזינו את כל ההצבעות והמפלגה המנצחת היא בעלת רוב הקולות, לפי סט מבחן זה המפלגה המנצחת הינה הסגולים.

התפלגות הקולות לפי סט מבחן זה:

|  |  |
| --- | --- |
| Vote Percentage | Color |
| 5.35% | Blues |
| 10.2% | Browns |
| 5.35% | Greens |
| 4.5% | Greys |
| **22.5%** | **Khakis** |
| 3.5% | Oranges |
| 2.7% | Pinks |
| **25.25%** | **Purples** |
| 4.95% | Reds |
| 4.45% | Turquoises |
| 2.8% | Violets |
| 4.45% | Whites |
| 3.95% | Yellows |



המסווג שלנו בעיקר מתבלבל בחיזוי צבעי טורקיז וסיגל לצבע החאקי ולכן אנו נרצה לשים דגש על צבעים אלו בסט המבחן הלא מתויג.

****

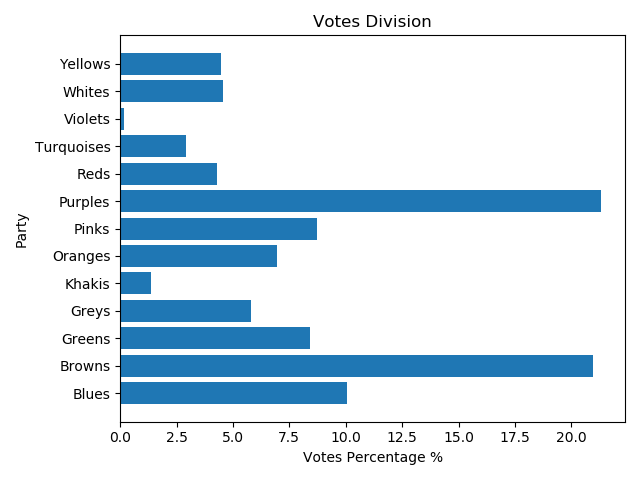
**סט מבחן לא מתויג (תחזיות אלה הן התחזיות להגשה):**

כעת כאשר בידנו סט מבחן לא מתויג אימנו את המסווג בשנית כל סט האימון מתחילת הקורס מכיוון שכעת סט המבחן אינו תלוי בסט האימון כלל.

על מנת לחזות את המפלגת המנצחת לפי סט זה חזינו את כל ההצבעות והמפלגה המנצחת היא בעלת רוב הקולות, לפי סט מבחן זה המפלגה המנצחת הינה הסגולים, בדומה לסט הקודם.

התפלגות הקולות לפי סט מבחן זה:

|  |  |  |
| --- | --- | --- |
| Vote Percentage | Votes | Color |
| 10.05% | 1005 | Blues |
| **20.96%** | **2096** | **Browns** |
| 8.43% | 843 | Greens |
| 5.78% | 578 | Greys |
| 1.38% | 138 | Khakis |
| 6.94% | 694 | Oranges |
| 8.72% | 872 | Pinks |
| **21.31%** | **2131** | **Purples** |
| 4.31% | 431 | Reds |
| 2.91% | 291 | Turquoises |
| 0.17% | 17 | Violets |
| 4.55% | 455 | Whites |
| 4.49% | 449 | Yellows |



כפי שניתן לשים לב מהחיזוי עבור סט המבחן הלא מתויג המפלגה המנצחת היא עדיין הסגולים אמנם המפלגה שכעת במקום השני היא לא החאקי כמו שראינו בסט המבחן הקודם אלא דווקא החומים. מכיוון שלפי סט המבחן הקודם אחוז הדיוק של המסווג שלנו מוגדר כטוב מאוד ולא ראינו נטייה ברורה של המסווג שלנו להתבלבל בין קולות של המפלגה החומה לבין קולות של מפלגות אחרות הרי שאנו מניחים כי תוצאות אלו אמינות.

לבסוף ייצאנו תוצאות אלו לקובץ CSV בשם “test\_predictions.csv” כנדרש בתרגיל.

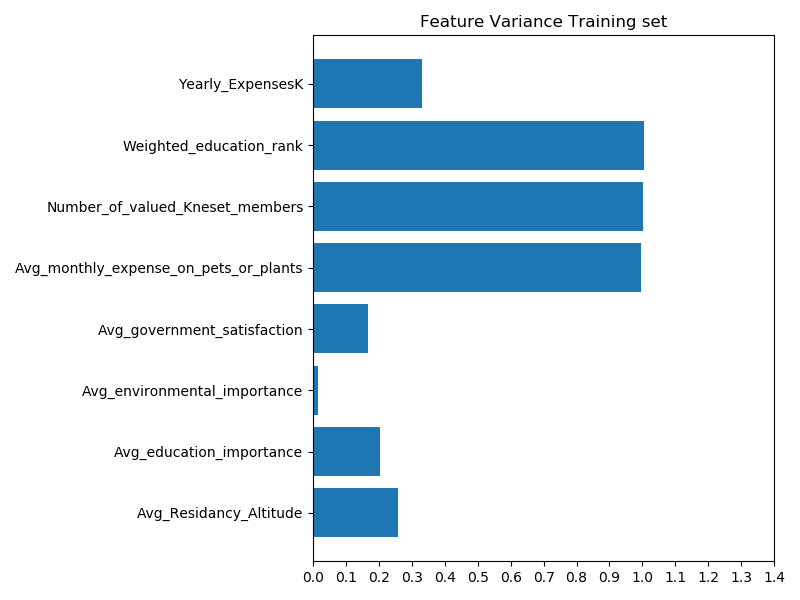
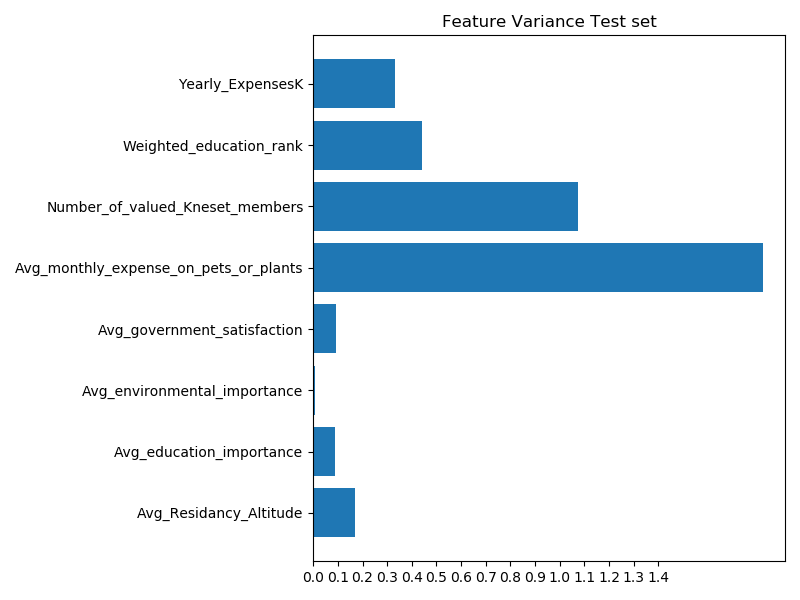
Building Steady Coalition Using Clustering Model:

לצורך בניית קואליציה יציבה לפי הגדרתה בתרגיל אנו נדרשים למצוא קבוצת מפלגות אשר קיים דמיון בין המאפיינים של המצביעים שלהם. מכיוון שבבעיה זו רוב מאפייני המצביעים (לאחר בחירת הפיצ'רים) הם מאפיינים בעלי ערך מספרי רציף הרי שנוכל למדוד בין שני דגימות על ידי מרחק בין הערכים המספריים של מאפייני הדגימות.  
על מנת לקבץ את המפלגות לכדי קואליציה הומוגנית פעלנו בשיטה הבאה על בסיס תרגיל בית 4:

1. Clustering Model- כאשר אנו משתמשים במודל מסוג זה אנו למעשה "מקבלים בחינם"   
   את היכולת לזהות את הדמיון והשוני בין המצביעים באמצעות היכולת להתייחס למצביעים אשר **נמצאים** באותו אשכול כבעלי מאפיינים דומים ואילו למצביעים אשר **אינם נמצאים** באותו אשכול כחסרי דמיון.  
   בצורה זו אנו ננסה ליצור **קבוצת מצביעים** השייכת לאשכול מסוים אשר יחסית הומוגנית והיא זאת שתהווה **בסיס להרכבת הקואליציה**.

לשם בניית קואליציה יציבה על ידי שימוש ב Clustering Model אימנו המודל Gaussian Mixture עם 2 אשכולות לאימון, מכיוון שהניב את תוצאות הכי טובות יותר בתרגיל בית 4.   
העדפנו מודל זה מכיוון שהדרך פעולה שלנו היא לנסות לבנות קואליציה על בסיס אשכול שהמודל יצר, לשם כך נרצה לרכז כל מצביעי מפלגה מסוימת באשכול מסוים ולא לפזר אותם על פני כמה אשכולות בצורה שלא ניתנת להבחנה לאיזה אשכול שייכת המפלגה. נגדיר כי מפלגה שייכת לאשכול מסוים אם אחוז סף ממבצעי המפלגה שייכים לאשכול זה. אשכול אשר מכיל מפלגות בעל רוב של 51% מצביעים יהווה בסיס להרכבת קואליציה.  
השלבים להרכבת הקואליציה הינם:

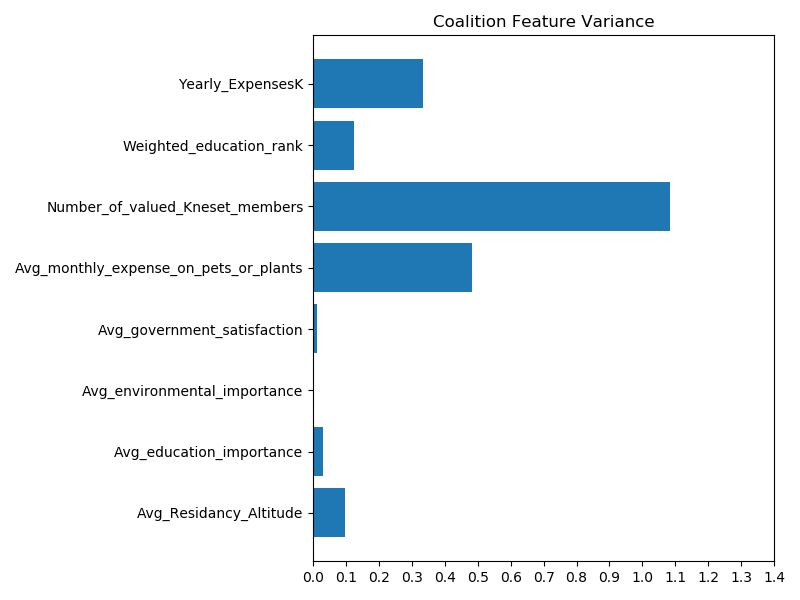
1. אימון המודל Gaussian Mixtureעל כל סט האימון.
2. הרכבת קואליציה בעזרת סט המבחן הלא מתויג אשר את התיוגים הכין עבורנו המסווג מהבעיה הקודמת בעל דיוק של 94.05% על סט מבחן מתויג.
3. כל אשכול שהמודל יצר יכול להוות בסיס להרכבת קואליציה, נגדיר שמפלגה שייכת לאשכול מסוים אם ורק אם לפחות 95% מסך כל הקולות שהצביעו לה שייכים לאשכול זה ובנוסף סך כל הקולות של המפלגות ששייכות לאשכול עולה על 51% מסך כל הקולות. בצורה זו בנינו קואליציה עם רוב קולות, קואליציה יחסית הומוגנית מכיוון שעבור מפלגות ששייכות לאשכול הרי שהמודל בחר לשים את רוב המצביעים שלהם באותו אשכול ולכל קיים דמיון בין המאפיינים שלהם, לעומת מפלגות שלא שייכות לאשכול הרי שלא קיים רוב מובהק של המצביעים שלהן באשכול ולכן הם יהיו באופוזיציה.
4. סינון קואליציות זהות שהורכבו ובחירת הקואליציה הכי הומוגנית, קואליציה שבה השונות בין מאפייני המצביעים של הקואליציה כולה היא הקטנה ביותר.

**תוצאות**:  
נשים לב כי לפני הרכבת קואליציה ניתן להבחין כי זוהי השונות בין מאפייני המצביעים השונים על סט האימון:  
(הפיצ'רים לאחר scaling)  
  
נשים לב כי לפני הרכבת קואליציה ניתן להבחין כי זוהי השונות בין מאפייני המצביעים השונים על סט המבחן:  


ניתן לשים כי סט האימון וסט המבחן בעלי מגמות זהות מבחינת השוני בין מאפייני המצביעים.

הקואליציה נבנתה על ידי המודל Gaussian Mixture בעזרת **סט המבחן** (הבסיס להרכבתה הוא אשכול 1). הקואליציה כוללת את המפלגות הבאות וכוללת 55.25% מסך כל הקולות.

Browns, Greens, Purples, Whites

לאחר הרכבת קואליציה ניתן להבחין כי ישנה ירידה בשונות עבור מאפייני המצביעים השייכים קואליציה בסט המבחן דבר המעיד כי קבוצת מצביעים זו יותר הומוגנית מהקבוצה הכוללת:  


בתרגיל הקוד ם בו התבקשנו למצוא קואליציה וחזינו כי הקואליציה תהיה:   
Browns, Greens, Greys, Oranges, Pinks, Purples, Reds, Whites.  
ניתן לשים לב כי הקואליציה הנוכחית מוכלת בקואליציה שחזינו בתרגיל קודם אמנם כעת קיבלנו קואליציה יותר מצומצמת וזאת מכיוון שיחס ההצבעות בין המפלגות השתנה והחומים קיבלו אחוז קולות גבוה יותר מאשר שצפינו על פני סט הוולידציה מתרגיל קודם שהורכב מסט האימון הכולל.  
מכיוון שקואליציה זו יותר מצומצמת הרי שהיא יותר הומוגנית ולכן בפרט יותר יציבה (לפי הגדרה) מהקואליציה שחזינו בתרגיל קודם.