



3

68

30/05/19

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 – Modeling:

ראשית כמתבקש טענו את הקובץ הדוגמאות שוב, והחלנו עליו את מניפולציות העיבוד המקדים להכנת המידע כפי שביצענו בתרגיל הקודם, מניפולציות אלה כוללות:

1. חלוקת סט הדוגמאות הכולל ( על די StratifiedShuffleSplit) לשלושה סטים:

|  |  |
| --- | --- |
| Percentage from Original Data Set | Data Set |
| 65% | Train set |
| 10% | Validating set |
| 25% | Test set |

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

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

* מה היא המפלגה המנצחת?
* חלוקת הקולות בין המצביעים
* אספקת שירותי הסעה למצביעים של כל מפלגה.

על מנת להתמודד עם משימות אלו נרצה למצוא את המסווג הטוב ביותר מבין קבוצת מסווגים ולכן ביצענו את התהליך הבא:

1. בחירת סוגי המסווגים איתם נרצה לבצע את התחזיות:
   1. Random Forest Classifier- הרכב של כמה עצי החלטה אשר הסיווג נקבע על ידי סיווג הרוב של עצי החלטה, בחרנו בו מכיוון שהוא מכליל את עקרון אצי החלטה ומכיוון שעצי החלטה הינו מסווג שאנו מכירים והינו מסווג שתמיד כדאי להיעזר בו.
   2. Stochastic Gradient Descent Classifier – בחרנו במודל זה מכיוון שאנו מכירים אותו ואת שיטת הפעולה שלו ומכיוון שרצינו לדעת האם קיים הפרדה לינארית לבעיה.
   3. K- Nearest Neighbors – בחרנו במודל זה מכיוון שאנו מכירים אותו ואת שיטת הפעולה שלו, בנוסף מודל זה פשוט ביחס למודלים אחרים וקל יותר להבנה. בנוסף מתוך הנחה כי בני אדם בעלי אותם מאפיינים נוטים לבחור באופן דומה הרי שכדאי לבדוק את ביצועיו על המידע.
   4. Decision Tree- עץ החלטה רגיל כפי שנלמד בהרצאה, בחרנו במודל זה למרות שהשתמשנו ביער של עצים מכיוון שרצינו להציג עץ החלט הוויזואלי ולקבל אינטואיציה כיצד עץ החלטה מתמודד עם הסט הדוגמאות.
2. בחירת הפרמטרים הטובים ביותר עבור כל סוג מסווג שנבחר. עבור כל אחד מסוגי מהמסווגים שהוזכר לעיל בנינו כמה דוגמאות ממנו עם סטים של היפר-פרמטרים שונים והערכנו את ביצועיו לפי מדד הדיוק.
   1. random\_forest\_tuple **=** **(  
        
      RandomForestClassifier(***random\_state***=**0**,** *criterion***=**'entropy'**,** *min\_samples\_split***=**5**,** *min\_samples\_leaf***=**3**,** *n\_estimators***=**50**),  
        
      RandomForestClassifier(***random\_state***=**0**,** *criterion***=**'entropy'**,** *min\_samples\_split***=**3**,** *min\_samples\_leaf***=**1**,** *n\_estimators***=**500**),  
        
      RandomForestClassifier(***random\_state***=**0**,** *criterion***=**'gini'**,** *min\_samples\_split***=**3**,** *min\_samples\_leaf***=**1**,** *n\_estimators***=**500**)**

**)**

* 1. sgd\_tuple **=** **(**

**SGDClassifier(***random\_state***=**0**,** *max\_iter***=**1000**,** *tol***=**1e-3**),**

**SGDClassifier(***random\_state***=**0**,** *max\_iter***=**1000**,** *tol***=**1e-2**),**

**SGDClassifier(***random\_state***=**0**,** *max\_iter***=**1500**,** *tol***=**1e-4**),**

**)**

* 1. knn\_tuple **=** **(**

**KNeighborsClassifier(***n\_neighbors***=**3**,** *algorithm***=**'auto'**),**

**KNeighborsClassifier(***n\_neighbors***=**5**,** *algorithm***=**'auto'**),**

**)**

* 1. tree\_tuple **=** **(**

**DecisionTreeClassifier(***random\_state***=**0**,** *criterion***=**'entropy'**,** *min\_samples\_split***=**5**,** *min\_samples\_leaf***=**3**),**

**DecisionTreeClassifier(***random\_state***=**0**,** *criterion***=**'entropy'**,** *min\_samples\_split***=**3**,** *min\_samples\_leaf***=**1**)**

**)**

הערכת כל מודל התבצעה באמצעות שיטת k- fold cross validation על סט האימון כאשר k = 5.

1. לאחר בחירת סט הפרמטרים הטוב ביותר לכל סוג מסווג, בחירת סוג המסווג הטוב ביותר התבצעה באמצעות בחינת הביצועים של המסווג על פני סט הוולידציה.  
   נדגיש כי מבחינתו הביצועים של המסווג הינם אחוז הדיוק (accuracy) על פני סט המבחן.  
   בחרנו דווקא בדיוק כמדד להערכת ביצועים מכיוון שהתחזיות המפלגה המנצחת וחלוקת הקולות בין המפלגות אנו זקוקים לדיוק גבוהה על מנת לספק תחזיות תואמות לסט הדוגמאות.  
   התוצאות שקבלנו:

Random Forest Classifier accuracy score on validation set is: 91.0 %

SGD Classifier accuracy score on validation set is: 72.3 %

KNN Classifier accuracy score on validation set is: 77.8 %

Decision Tree Classifier accuracy score on validation set is: 86.5 %

ולכן המסווג הנבחר הינו Random Forest Classifier, עם סט הפרמטרים הבאים:

**RandomForestClassifier(***random\_state***=**0**,** *criterion***=**'entropy'**,** *n\_samples\_split***=**3**,** *min\_samples\_leaf***=**1**,** *n\_estimators***=**500**)**

1. ביצוע התחזיות התבצע על ידי המסווג הנבחר כאשר הוא אימנו אותו מחדש על סט האימון וסט הוולידציה ביחד. המוטיבציה לבצע זאת היא מכיוון שכבר השתמשנו בסט הוולידציה על מנת להעריך באיזה מסווג להשתמש ואנו לא מתכוונים להשתמש בו שוב וכמובן שסט הוולידציה מכיל דוגמאות שנרצה שהמסווג שלנו יתאמן, באופן כללי ככל שסט האימון גדול יותר כך המסווג טוב יותר.  
   **להלן ביצועי המסווג על סט המבחן:**

The accuracy score on TEST set is: 92.28%

The error score on TEST set is: 7.72%

כפי שניתן לשים לב קיבלנו דיוק יותר גבוה על סט המבחן מאשר בבחינת ביצועי המסווג ב k- fold cross validation וגם מאשר בחינת ביצועי המסווג על סט הוולידציה דבר המעיד כי המסווג אינו סובל מoverfitting על סט האימון שלו.

1. כעת נסביר כיצד פעלנו עבור כל אחת מהתחזיות שנתבקשנו לבצע:
   1. עבור תחזית המפלגה המנצחת ביצענו חיזוי על סט המבחן ובחרנו במפלגה שקיבלה הכי הרבה קולות, ולפי המסווג שלנו המפלגה המנצחת הינה: הסגולים 😊.
   2. עבור

First Bonus – Relief Algorithm: