מבוא למערכות לומדות – תרגיל 2

תרגיל מחייב – נקודות ראויות לציון

- היות וסדר הפעולות חשוב, מצוינות פה הפעולות לפי סדר הופעתן בקוד:

1. **Data Splitting:**
   a. בצענו פיצול לשם שמירת הסטים המקוריים בקובץ נפרד.
   b. שמרנו את אינדקסי הפיצול על-מנת שנוכל להמשיך לעבוד על ה-dataFrame המלא (ללא פיצול) ולפצל לאותן הקבוצות בעתיד.

2. **Type/Value modification:**
   a. הפכנו את כל הערכים למיספריים בצורה אוטומטית (על-גבי כל ה-data, ללא חלוקה לסטים).

3. **Identify and set the correct type of each attribute:**
   a. הסרנו ערכים שליליים מהמערכת על-ידי הפיכתם ל-NaN, מתוך הנחה כי כל הערכים אמורים להיות אי-שליליים.

4. **Imputation:**
   a. פצלנו את ה-data ל-train/validation/test בשנית (עם אותם אינדקסים, כך שהדוגמאות בכל סט נשארו זהות).
   b. השלמנו בסט האימון את הערכים החסרים בצורה הבאה:
      i. לערך float השתמשנו ב-median של אותו הפיצ'ר בסט.
      ii. לערך str השתמשנו ב-most_common של אותו הפיצ'ר בסט.
   c. השלמנו בסטים הנותרים (validation/test) את הערכים באופן דומה, אך באמצעות ערכי ה-median וה-most common שנלקחו מתוך סט האימון.

5. **Outlier Detection:**
   a. השתמשנו ב-LocalOutlierFactor מהחבילה sklearn על-מנת לזהות outliers ב-train ויצרנו קבוצה חדשה של train ללא ה-outliers בשם train_without_outliers בה השתמשנו בשלבים הבאים.

6. **Normalization:**
   a. לכל עמודה ב-train_without_outliers, השתמשנו ב-NormalTest מהחבילה SciPy על מנת לבדוק האם איברי העמודה בעלי התפלגות נורמלית.
   b. עמודות שמתפלגות נורמלית נורמלו על-ידי z-score, והאחרות על-ידי min-max לכל אחד מהסטים הדרושים.

7. **Feature Selection:**
   1. הרצנו VarianceThreshold מ-sklearn (?) בצורה שמרנית על train_without_outliers, ושמרנו ב-train_without_outliers רק את הפיצ'רים הנבחרים. האלגוריתם סינן 2 פיצ'רים מיותרים:

Voting_Time
Financial_balance_score_(0-1)

   2. בצורה דומה, הרצנו על train_without_outliers את Relief ולאחריו את SFS. הרחבה על הנ"ל בהמשך. Relief סינן פיצ'ר יחיד:

Number_of_differnt_parties_voted_for

אחריו, SFS בחר חמישה פיצ'רים לקבוצה הסופית:

Avg_education_importance
Number_of_valued_Kneset_members
Avg_monthly_income_all_years
Avg_government_satisfaction
Last_school_grades

3. כעת, הורדנו 8 תכונות כאשר 5 מתוכן סווגו כטובות וה-3 הנותרות סווגו כמיותרות. יצרנו וקטור עם ערכי 0 לכל שאר התכונות והרצנו Feature Importance Forest, SelectFwe לפי chi2 ו-SelectKBest לפי mutual_info_classifier, כולם מ-sklearn על-גבי מה שנשאר מ-train_without_outliers . כל אחד מה-wrappers הללו החזיר 1 עבור ה-features שהיו חשובים לדעתו. לבסוף בחרנו ב-features שקבלו לפחות 2 קולות. נבחרו 13 ה-features הבאים:

Avg_monthly_expense_when_under_age_21
AVG_lottary_expanses
Most_Important_Issue
Avg_monthly_expense_on_pets_or_plants
Avg_environmental_importance
Married
Avg_Residancy_Altitude
Avg_Satisfaction_with_previous_vote
Avg_monthly_household_cost
Phone_minutes_10_years
Avg_size_per_room
Weighted_education_rank
Political_interest_Total_Score

4. לסיכום, נבחרו בסך הכל 18 features מתוך 37 שהתקבלו, כאשר הסט הסופי הוא:

Avg_education_importance
Number_of_valued_Kneset_members
Avg_monthly_income_all_years
Avg_government_satisfaction
Last_school_grades
Avg_monthly_expense_when_under_age_21
AVG_lottary_expanses
Most_Important_Issue
Avg_monthly_expense_on_pets_or_plants
Avg_environmental_importance
Married
Avg_Residancy_Altitude
Avg_Satisfaction_with_previous_vote
Avg_monthly_household_cost
Phone_minutes_10_years
Avg_size_per_room
Weighted_education_rank
Political_interest_Total_Score

תרגילי בונוס - הרחבה

<u>משימת בונוס לזוגות:</u>

A. בחלק זה ביצענו מספר פעולות. בהתחלה, יצרנו מטריצת קורלציות של ה-features ביחס ל-label. קבענו סף של 0.1 ומצאנו את כל ה-features שעבורם הקורלציה הנ"ל גבוהה מ-0.1 בערך מוחלט, לאחר השלמת ערכים חסרים, הסרת ערכים שליליים והתעלמות מ-outliers. ה-features שהתקבלו הם:

Avg_monthly_income_all_years
AVG_lottary_expanses
Number_of_valued_Kneset_members
Avg_monthly_expense_when_under_age_21
Avg_monthly_expense_on_pets_or_plants
Weighted_education_rank

את כל הפיצ'רים הנ"ל שמרנו בתור בקרה על פעולות ההמשך ולא הסרנו אותם מסט האימון. לאחר ביצוע feature selection, נבחרו ע"י שקלול הפעולות שפירטנו 18 features. כבקרה, בדקנו את הימצאותם של הפיצ'רים בעלי הקורלציה הגבוהה ל-Vote ואכן נוכחנו לדעת שכל השישה שנמצאו אכן נמצאים בסט שהתקבל כאשר מתוך השישה הנ"ל, שני features נבחרו ע"י SFS והארבעה הנותרים נבחרו בשלב האחרון ע"י לפחות שני אלגוריתמים.

בשלב השני,

B. מימוש אלגוריתם Relief מופיע בקוד.
שיטת הפעולה של האלגוריתם היא ביצוע איטרציות, כאשר בכל איטרציה נדגמת אקראית דוגמה מסט האימון. לאחר מכן, האלגוריתם מחפש את הדוגמה הקרובה ביותר לדוגמה שנדגמה המסווגת עם אותו label (nearest hit) ואת הדוגמה הקרובה ביותר לדוגמה שנדגמה המסווגת עם label שונה (nearest miss). כעת, האלגוריתם מחשב משקלים עבור הפיצ'רים לפי שקלול המרחק מהדוגמאות. לבסוף, מתקבל וקטור משקלים עם משקל משוקלל לכל פיצ'ר וניתן לבחור אותם לפי סף מסוים.

מעצם שיטת פעולתו של האלגוריתם, יש לו חסרון הנובע מאקראיות בחירת הדוגמאות. בפרט, ככל שמספר האיטרציות נמוך, כך חיסרון זה בולט יותר עבור מדגם דוגמאות מפוזר.
חסרון נוסף של האלגוריתם כ-filter method הוא שהוא אינו מתחשב במודלים אחרים אלא רק בחישוב המבוצע על ידיו.
לעומת זאת, יתרון של Relief, כמו רוב שיטות ה-filter, הוא בזמן ריצתו שהוא יחסית מהיר לאיטרציה בודדת ועל כן מאפשר לבצע מספר איטרציות רב ולקבל תוצאות יותר אמינות סטטיסטית (בעיקר בהתחשב בעובדה שהוא בוחר דוגמאות באופן אקראי).

את האלגוריתם הרצנו על סט האימון שהוצאנו ממנו את ה-outliers לאחר שעבר נרמול. הרצנו את האלגוריתם עשר פעמים כאשר בכל פעם האלגוריתם ביצע 100 איטרציות. בכל ריצה יצרנו קבוצה המכילה את הפיצ'רים שאותם האלגוריתם הציע לסנן לפי סף שהוגדר כ-0.1. לבסוף, מצאנו את הפיצ'רים המופיעים בחיתוך תוצאות כל הריצות והורדנו אותם מ-train_without_outliers.
הפיצ'ר אותו בחר Relief להוריד הוא:
Number_of_differnt_parties_voted_for

Triplets Mandatory Assignment

A. מימוש האלגוריתם SFS מופיע בקוד.
שיטת הפעולה של האלגוריתם היא חמדנית ופועלת בצורת forward selection. כלומר, האלגוריתם מתחיל את פעולתו עם סט features ריק ובאופן חמדני בכל פעם מבצע הוספת features ובודק את התוצאה לפי המסווג שאיתו הוא עובד.

באופן ישיר מאופן פעולתו החמדני, אחד החסרונות הבולטים של האלגוריתם הוא שלאחר בחירת תכונה מסוימת לא ניתן להסירה מהקבוצה הנבחרת. בנוסף, כאלגוריתם חמדני האלגוריתם מסתכל אך ורק על הצעד הנוכחי ולכן לא מאפשר צעדים מרעים שיתכן ובתוצאה הסופית היו משפיעים לטובה.
נוסף על כך, בדומה לרוב אלגוריתמי ה-wrapper, זמן ריצת האלגוריתם הינו ארוך.

לעומת זאת, לאלגוריתם קיימים גם יתרונות, ביניהם התממשקותו עם אלגוריתמי סיווג. כך ניתן לדעת ממה מושפעת תוצאת אלגוריתם ה-SFS בהתאם לאופן פעולת המסווג.

את האלגוריתם הפעלנו עם מספר מסווגים: KNN, Random Forest, עץ החלטה, ExtraTree ו-SVM. עבור כל אחד מהמסווגים הנ"ל, יצרנו בעת ריצת האלגוריתם קבוצה המכילה את כל הפיצ'רים שעברו את הסף שהוגדר כ-0.000001. לבסוף, מצאנו את ה-features שנבחרו על ידי לפחות ארבעה מתוך חמשת המסווגים, בחרנו אותם לקבוצה הסופית והסרנו אותם מ-train_without_outliers.

מסווג Decision Tree בחר את שמונת הפיצ'רים הבאים:

AVG_lottary_expanses
Last_school_grades
Avg_education_importance
Avg_monthly_household_cost
Phone_minutes_10_years
Avg_size_per_room
Number_of_valued_Kneset_members
Avg_monthly_income_all_years

מסווג Extra Tree בחר את ששת הפיצ'רים הבאים:

Last_school_grades
Avg_education_importance
Avg_Satisfaction_with_previous_vote
Number_of_valued_Kneset_members
Avg_government_satisfaction
Avg_monthly_income_all_years

KNN בחר את ששת הפיצ'רים הבאים:

Married
Avg_education_importance
Avg_Satisfaction_with_previous_vote
Most_Important_Issue
Avg_government_satisfaction
Avg_monthly_income_all_years

SVM בחר את שמונת הפיצ'רים הבאים:

AVG_lottary_expanses
Last_school_grades
Avg_monthly_expense_when_under_age_21
Avg_environmental_importance
Avg_education_importance
Phone_minutes_10_years
Number_of_valued_Kneset_members
Avg_government_satisfaction

Random Forest בחר את תשעת הפיצ'רים הבאים:

Last_school_grades
Avg_environmental_importance
Avg_education_importance
Avg_monthly_household_cost
%Of_Household_Income
Avg_size_per_room
Number_of_valued_Kneset_members
Avg_government_satisfaction
Avg_monthly_income_all_years

כאמור, החלטנו לבחור את כל הפיצ'רים שנבחרו ע"י לפחות ארבעה מתוך חמשת המסווגים שנבדקו.
פיצ'ר יחיד נבחר ע"י כל המסווגים, והוא:

Avg_education_importance

ארבעה פיצ'רים נבחרו ע"י ארבעה מתוך חמשת המסווגים, והם:

Number_of_valued_Kneset_members
Avg_monthly_income_all_years
Avg_government_satisfaction
Last_school_grades

ולכן הורדנו את חמשת הפיצ'רים הנ"ל מהקבוצה וסימנו אותם בתור נבחרים.