Data Preparation

Prepossessing

Data split and Filling missing values

חילקנו את המידע באופן הבא:

- 70% אימון, 15% ולידציה, 15% מבחן. (stratify)

בבחינת המידע ראינו כי המידע החסר בכל תכונה הוא קטן יחסית למידע הקיים ולכל התכונות מספר הערכים החברים הוא מאותו סדר גודל.

בחרנו בקבוצת אימון גדולה מכיוון שרצינו להשלים את המידע החסר בה בשיטת Closest Fit תוך התייחסות לתיוג של הדוגמאות, לדעתנו עדיף לעשות זאת על קבוצה גדולה יחסית.

את המידע החסר בקבוצות הולידציה והאימון השלמנו בשיטת החציון בתקווה שזה לא ישפיע על שונות התכונות בצורה ניכרת מכיוון שהן קטנות יחסית. את החציון לקחנו מקבוצת האימון בלבד, עשינו זאת לפני שביצענו את השלמת המידע החסר, כדי שנקבל מידע מדוייק יותר במקרה ושיטת השלמת המידע החסר לא טובה מספיק.

Outliers

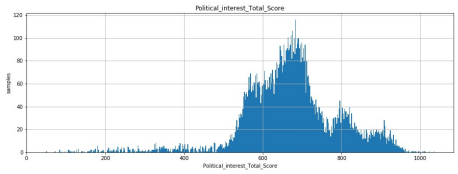
בבחינת המידע (בקובץ notebook) ראינו שיש שתי תכונות שלדעתנו צריכות להיות אי-שליליות (ממוצעי הוצאות), אז הנחנו שערכים אלו שגויים והתייחנו אליהם כאפסים.

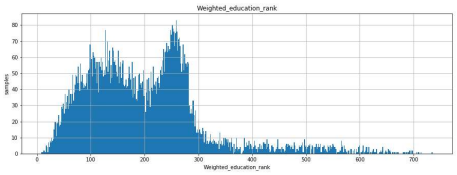
Data Transformation

מתכונות הבדידות שאין בהן סדר יצרנו תכונות חדשות בשיטת One-Hot.

את נרמול ביצענו בשיטות הבאות:

- Z-score עבור התכונות שראינו בבחינת המידע שלפילוג שלהן יש 'זנבות' ארוכים, כדי שמרכז המסה של ההתפלגות תהיה בקטע $[-1,1]$. למשל עבור התכונות הבאות:





- Min-Max עבור שאר התכונות, גם לקטע $[-1,1]$.

בנוסף, לאחר הנסיונות הראשוניים בבחירת התכונות, על כל תכונה עם ערכים חסרים הוספנו תכונה המהווה אינדיקטור שהמידע היה חסר במקור.

קבצים – בתקייה prepossessing

- understanding_the_data.ipynb, השתמשנו בקובץ לבחינת המידע (עבר שינויים, הבסיס שם)
- preprocessing.py, עיבוד המידע.

Feature Subset Selection

שיטה א' SFS

- השתמשנו ב-KNN וביער רנדומלי על מנת לבחור 15 תכונות. (ס נבחרים לכל מסווג)
- את האלגורית SFS מימשנו בעצמנו והחזרנו את התכונות שנבחרו לפי סדר בחירתם (כדי שנוכל לדמות SFS עבור מספר תכונות קטן יותר בהתסכלות על ראש הרשימה).
- כדי לקבל הערכה על כמות התכונות אשר מביאות לתוצאות טובות, עבור כל אחד מהסטים שהתקבלו, בדקנו את הדיוק המתקבל. (הפלט המלא נמצא בקבצים)
- נראה כי SFS עם KNN מצליח להביא לדיוק טוב יותר, גם עבור ולידציה עם KNN וגם עבור ולידציה עם היער.
- הדיוק עלה בהדרגה, כאשר המקסימום התקבל עבור סביב ה-15 תכונות (בפועל זה פחות בגלל הפיצול)

שיטה ב' Relief

- מימשנו את האלגוריתם בעצמנו.
- ביצענו כמו בשיטה קודמת רק שהמיון של התכונות התבצע בעזרת Relief.

שיטה ג' SBS

- מכיוון שהיו לנו המון תכונות, הפעם ביצענו את בדיקת הדיוק ובחירת התכונות בעזרת KNN בלבד.
- הרצנו את האלגוריתם עבור 15 תכונות.

שיטה ד' SFS with random double backward

- על מנת למצוא אולי סט יותר מעניין של תכונות, מימשנו אלגוריתם SFS אשר מתקדם צעד קדימה, ובהיסתברות נמוכה יחסית עושה שני צעדים אחורה. כלומר, מוריד זוג תכונות שהורדתם ממקסמת את הדיוק.

שיטה ה' Correlation Coefficient

- בין התכנות לתיוג.

בחירת התכונות ניתוח תוצאות

- SFS, למרות שלא היו הבדלים מאוד משמעותיים בדיוק, לקחנו בתור התחלה, משיקולי דיוק, את 15 התכונות הראשונות של KNN (נק' המקסימום), ואת 8 התכונות הראשונות שנבחרו ע"י היער. הייתה חפיפה בין התכונות שנבחרו לכן נראה כי בחירה זו טובה.
- היו תכונות ש-Relief בחר בהן עם ציון גבוה שלא נבחרו על ידי SFS (מאלו שלקחנו), התכונות הן Weighted_education_rank ,Avg_size_per_room. לקחנו את 10 הראשונים, רובם חפפו עם אלו שבחרנו מ-SFS. אנו מאמינים כי שילוב של שיטות יעזרו לנו בהמשך מכיוון שאלגוריתם SFS הוא אלג' חיפוש לוקאלי ולכן יתכן שנקלע למינימום מקומי, בעזרת Relief אנו מקטינים את הסיכוי לכך.
יתרון נוסף לאגוריתם זה הוא המהירות שבו הוא מוצא את התכונות הטובות (ובצורה לא רעה) לעומת SFS.
- כל התכונות שהתקבלו מ-SBS, למעט תכונה אחת, כבר נמצאו בסט התכונות שבחרנו כבר, אנו מאמינים כי זה מעיד על כך שהבחירות שעשינו למעלה כנראה טובות וייתכן כי סט התכונות האופטימליות נמצאות במרחב התכונות שבחרנו.
- SFS with random double backward, למרות היותו מיוחד באופיו, קיבלנו כי בשתי הרצות שהביאו את הדיוק למקסים, כל סט התכונות שנבחרו היו כבר מאלה שבחרנו למעט שני אינקטורים של תכונות שבחרנו, בפעם נוספת זה מחזק את התחושה שהבחירות לעיל טובות.
- בדיקת הקורלציה ולקיחת 10 התכונות בעלות הקורולציה הגבוהה ביותר לא הוסיפו תכונות חדשות לאלו שכבר בחרנו. שוב, חיזוק משמעותי להערכות שלו. אנו מאמינים שהקשר בין התכונות לתיוגים הוא מורכב, סיבה

מרכזית להערכה זו היא שהדיוק על פני מספר לא קטן של שיטות הביא לתוצאה כמעט זהה. מצד שני ראינו כי התכונות הנבחרות בין המודלים השונים זהים ברוב הפעמים ואלו הן התכונות בעלי הקורלציה הגבוהה ביותר.

- לסיכום, אנו חושבים שהשיטה ההיברידית הזו, של שימוש בסוגים שונים ובשיטות שונות של בחירת התכונות ואיחודם תגדיל את הסיכוי לבניית מסווג טוב, אנו מאמינים כי כיוונון הפרמטרים של אלגוריתמי הלמידה ויצירת מסווג מורכב יותר עם סט זה יתכן ויביא לתוצאות טובות יותר, נחכה ונראה 😊

*ניסיונות שנכשלו: נרמול בשיטה שונה, הסרת כל השורות עם הערכים החסרים, בלי להוסיף אינדיקטורים, איחוד כל התכונות ולקיחת התכונות הנפוצות ביותר, ועוד המון.... אבל יש לציין כי הפרש השגיאה לא היה מאוד משמעותי.

קבצים

- בתקיה הראשית נמצאים האלגוריתמים לבחירת התכונות שמימשנו בעצמנו. (SFS,RELIEF..)
- בתקיה feature_selection, נמצאים הסקיפטים והפלטים שבחרו את התכונות בכל אחת מהשיטות.
- בתקיה csv_dada_sets נמצאים קבצי המידע בפורמט csv.

מגישים

אביאל שמחי, 305376063

עמית סולומון, 305025785