

תכנות מתקדם - מטלה 4

נושא: עיבוד נתונים

הנחיות:

1. נא לכתוב קוד מסודר, תוך שימוש בשמות משתנים בעלי משמעות והערות היכן שנדרש.
2. כל שאלה צריכה להיות מוגשת בנפרד
3. בתחילת כל מענה, נא להוסיף כהערה את שמות המגישים והאם התייעצתם נעזרתם בסטודנטים נוספים.

הגשה:

1. יש להגיש את העבודות בזוגות (עפ"י הקבוצות שנרשמתם).
2. יש להגיש מחברת בודדת (Jupyter notebook) עם המטלה.
3. שם הקובץ המוגש צריך להיות: student1_id_student2_id.ipynb

בהצלחה!

הקדמה

במטלה זו, המשימות ניתנות למימוש במספר דרכים שונות. בסעיפים בהם ניתן, יש להימנע משימוש באיטרציות (לולאות) ובמקום יש להשתמש בפונקציונליות של Pandas / Numpy. אוסף הנתונים המצורף, "StudentsPerformance.csv", הוא קובץ הנתונים למטלה זו, אשר מטרתה היא עיבוד הנתונים לקראת ניתוח והסקת מסקנות. מצורפות 5 שורות הראשונות בטבלה:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score |
|---|--------|----------------|-----------------------------|--------------|-------------------------|------------|---------------|---------------|
| 0 | female | group C | bachelor's degree | standard | completed | 100 | 100 | 100 |
| 1 | female | group A | associate's degree | free/reduced | none | 45 | 65 | 58 |
| 2 | male | group D | high school | free/reduced | completed | 87 | 81 | 87 |
| 3 | female | group B | bachelor's degree | free/reduced | none | 80 | 87 | 85 |
| 4 | male | group D | some high school | standard | none | 81 | 82 | 70 |

gender: female, male

ethnicity: group A, B, C, D, E

education: some college, associate's degree, high school, some high school, bachelor's degree, master's degree

lunch: standard, free/reduced

pre-course: none, completed

math, reading, writing: grades 0 - 100

חלק א' (50 נקודות) - טיפול בערכים חסרים ומשתנים קטגוריאליים

1. הדפס תחקור ראשוני עבור הנתונים: גודל הנתונים, שמות העמודות ומדדים סטטיסטיים לעמודות נומריות (תזכורת: describe).
 2. הצג דיאגרמת bar המציגה כמות הנתונים אשר קיימים עבור כל קבוצה ב-ethnicity.
 3. מהו אחוז הסטודנטים אשר לומדים בתיכון ביחס לכלל הנתונים?
 4. במידה וקיימים ערכים חסרים, הדפס את שמות העמודות בהן חסר ערכים ואת כמות הערכים החסרים. אם לא, הסבר כיצד הבחנת כי אין ערכים חסרים בנתונים.
 5. המר את עמודת gender לקידוד one-hot-vector. כלומר, לאחר פעולה זו, יתווספו שתי עמודות חדשות לנתונים; (gender_male, gender_female) המכילים ערכים בינאריים בלבד (0/1).
 6. בהתבסס על התשובה של סעיף 5 בלבד, כמה גברים וכמה נשים קיימים בנתונים?
 7. השתמש בפקודות Pandas, ושנה את שם העמודה "race/ethnicity" ל - "ethnicity".
 8. כתוב את הפונקציה get_last_char אשר מקבלת כקלט מחרוזת ומחזירה את התו האחרון במחרוזת.
 9. הפעל את הפונקציה הנ"ל על עמודות ethnicity (שימו לב: לא ליצור עמודה חדשה, אלא לדרוס את הקיימת).
 10. עבור עמודת test preparation course, המר את הערך "completed" לערך 1, ואת השאר ל-0. (כלומר, לאחר סעיף זה, עמודת test preparation course היא עמודה בינארית).
 11. השתמש ב-one-hot-vector encoding והמר את עמודות "lunch" בהתאם.
 12. עבור עמודת education, המר את ערכיה כך ש:
 - a. אם הערך הוא some college, bachelor's degree, master's degree - המר את ערך העמודה להיות "higher education".
 - b. אם הערך הוא associate's degree - המר את ערך העמודה להיות "degree".
 - c. אחרת, המר את ערך העמודה להיות "high school".
 13. המר את עמודת "ethnicity" לקידוד מספרי ללא שימוש ב-Pandas.
- הדרכה:** הגדר מילון אשר ממפה את קבוצות ה-ethnicity למספרים רצופים (למשל, A=1, B=2, וכו'). השתמש ב-apply, והפעל את מיפוי הערכים על העמודה הרלוונטית.

חלק ב' (20 נקודות) - נרמול נתונים + שאילתות

1. עבור עמודות math, reading, writing - המר אותם עפ"י הכלל הבא:
נסמן ב-m את ממוצע העמודה וב-s את סטיית התקן של העמודה. הערך החדש בכל אחת מהעמודות הוא הערך הקודם פחות m, לחלק ל-s. במילים אחרות, הפחיתו מכל ערך את ממוצע העמודה ולאחר מכן יש לחלק בסטיית התקן.
2. עבור כלל קבוצות ה-"ethnicity" הקיימות, מי היא הקבוצה אשר קיימים בה היא הרבה סטודנטים אשר השלימו קורס\מכינה לפני הלימודים?
3. הצג בדיאגרמה (לבחירתך) את היחס בין ממוצע ציוני הסטודנטים במתמטיקה עפ"י כל אחת מקבוצות ethnicity. כלומר, חשב לכל קבוצה את ממוצע הציונים במתמטיקה והצג תוצאה זו בצורה גרפית.
4. צור DataFrame חדש המכיל את העמודות המנורמלות של ציוני הסטודנטים במתמטיקה, קריאה וכתביה (מסעיף 1). הפעל את הפקודה הבאה:

```
new_df.hist()
```

מה הוא הפלט? הסבירו במשפט אחד.
5. הציגו שתי דיאגרמות Pie המציגות את התפלגות הציונים במתמטיקה, קריאה וכתביה עבור נשים \ גברים. במילים אחרות, גרף Pie ראשון עבור התפלגות הנשים והשני עבור התפלגות הגברים.

חלק ג' (30 נקודות)

חלק זה במטלה אינו קשור לחלקים הקודמים.

1. כתבו את הפונקציה distance, המקבלת שתי רשימות (זהות באורך) ומחזירה את המרחק בין רשימות אלו.

מרחק בין רשימות הוא כמו מרחק בין נקודות. למשל, אם איברי הרשימה הראשונה מסומנים ב-x ואיברי הרשימה השניה מסומנים ב-y, אז המרחק הוא:

$$d(lst1, lst2) = \sqrt{(x_0 - y_0)^2 + (x_1 - y_1)^2 + \dots + (x_{n-1} - y_{n-1})^2}$$

2. צור DataFrame בגודל 100 עם שתי עמודות: הראשונה id (מספר אקראי שלם כלשהו ללא כפילויות), והשניה עמודת item, אשר מכילה אות רנדומלית כלשהי מהקבוצה: A,B,C,D,E. הדפס את value_counts עבור עמודת item ובדוק כי אכן מופיעים רק האותיות הנ"ל.

3. המר את עמודת item ל-one-hot-vector encoding.

4. כעת, אוסף הנתונים אמור להיראות (בערך) כך:

| id | item_A | item_B | item_C | item_D | item_E |
|----|--------|--------|--------|--------|--------|
| 91 | 0 | 1 | 0 | 0 | 0 |
| 90 | 1 | 0 | 0 | 0 | 0 |
| 33 | 0 | 0 | 0 | 1 | 0 |

עבור כל שורה, הוסף עמודה חדשה המכילה את מרחק השורה מהרשימה [0,1,0,0,0].

למשל, עבור השורה השניה, נבדוק את המרחק בין [0,1,0,0,0] לבין [1,0,0,0,0].

5. הפעל groupby על העמודה החדשה, ובדוק מי הם השורות אשר מרחקם הוא בדיוק 0. מה ניתן להגיד על שורות אלה?

6. במידה ומרחק השורה הוא לא אפס, מה הוא המרחק המקסימלי אשר מתקבל מסעיף 4? הסבר במשפט אחד מדוע.