Introduction to Data Science

תרגיל מספר 2

חלק א פייתון: יישום אלגוריתם K-means הכנת הנתונים וניתוחם בעזרת Jupyther.

חלק ב אקסל: יישום אלגוריתם K-means בעזרת solver.

מרצה:

ד"ר גייל גלבוע פרידמן

מתרגלים:

מר עידן לופו

מר אפי פקאני

עבודה ביחידים

תאריך הגשה - 21/05/2021

מטרת התרגיל

היכרות עם אלגוריתמי Clustering (K-Means), הכנת הדאטה + אימון וניתוח תוצאות האלגוריתם.

התרגיל מתייחס לקבצי דאטה עם נתונים שנאספו ממערכת מובילאיי בת"א ובברצלונה:

Mobileye_risk_Barcelona_NEW.csv.

הגשה

- **הגש תיקיית zip ששמה S1_YOURNAME_YOURID.zip המכילה 3 קבצים (יודגש באדום בהנחיות): Word, xlsx, ipynb.**
- ההגשה דרך הMoodle, בתיקייה ייעודית לתרגיל זה.
- יש להגיש Jupyter Notebook בשם: A2_YOURNAME_YOURID.ipynb.
- יש להגיש מחברת מסודרת המכילה:
  - תאי Markdown המחלקים את המחברת ל-sections בעזרת כותרות בהתאם לתרגיל (Part 1, Part 2 וכד')
  - תאי Code עבור שאלות תכנותיות
  - במידה ותרצו להוסיף הערות לקוד ניתן להוסיף עם סימון הערה – "#".
  - **<u>לא</u> לרשום תשובות מילוליות בתוך המחברת/בתוך האקסל אלא רק ב-Word.**
- עבור **<u>כל שאלה</u>** רשמו את התשובה בקובץ Word והציגו תצלום מסך של הפלט כסימוכין.
  - עבור שאלות בהן רשום הסבר, **שימו לב** להוסיף תשובה מילולית ב-Word.
  - עבור שאלות בהן רשום אקסל בלבד, ניתן לבצע את התשובה באקסל בלבד ללא תשובה מילולית.
  - עבור שאלות בהן רשום קוד בלבד, ניתן לצרף צילום מסך של הקוד והפלט ל-Word.
  - דוגמא לתשובה תקינה:

ניתן לראות שבקובץ שלנו ישנם 2823 שורות ו12 עמודות:

```
rows, cols = data.shape

print('There are', rows, 'rows.')
print('There are', cols, 'cols.')
```

```
There are 2823 rows.
There are 12 cols.
```

בהצלחה !

הקדמה

בהמשך לתרגיל הראשון, צירפנו לכם כחומר עזר/העשרה מאמר אינטראקטיבי המסביר כיצד מובילאיי משתמשים בנתונים שאספו על מנת להוציא תובנות, מאמר זה יכול לעזור בהבנת הנתונים של מובילאיי:

https://storymaps.arcgis.com/stories/43b9f0dbb18d4f8199033705b750ec49

- האם כעת קל יותר להבין את המשמעויות של המשתנים בדאטה שלנו? נסו להסביר במילים שלכם לגבי כל משתנה, מה הוא מודד וכיצד חושב (בקירוב) על מנת להבין טוב יותר את הנתונים (אין צורך להגיש תשובה).

חלק 1 - קריאת נתונים (7 נק')

לתרגיל זה מצורף קובץ csv המכיל נתונים ממערכת מובילאיי על מקטעי דרך ברחבי ברצלונה –

.Mobileye_risk_TelAviv_NEW.csv

להזכירכם, מדובר בנתונים מחודש ינואר, כל תצפית מייצגת מקטע דרך אחר בעיר ברצלונה.
ראו את פירוט השדות מטה:

start_lat – the start point latitude of drivable path.
start_long – the start point longitude of drivable path.
end_lat – the end point latitude of drivable path.
end_long – the end point longitude of drivable path.
detection_drives_count – number of drives during the month.
avg_speed – average speed during the month.
near_miss_pedestrian_ratio – ratio of pedestrian collision warning during the month.
near_miss_bicycle_ratio – ratio of bicycle collision warning during the month
near_miss_vehicle_ratio – ratio of Forward collision warning during the month
avg_pedestrian_on_road_volume – average number of pedestrians detected on the drivable path during the month.
avg_bicycle_on_road_volume – average number of cyclists detected on the drivable path during the month.
braking_count – total number of braking during the month.
cornering_count – total number of cornering during the month.
harsh_braking_ratio – normalized ratio of harsh braking out of the total braking during the month.
harsh_cornering_ratio – normalized ratio of harsh cornering out of the total cornering during the month.

1. הורד את קובץ הנתונים הנ"ל.

2. קוד + הסבר (3 נק') - ייבאו את ספריית **sklearn** וסכמו מהאינטרנט על מטרת הספריה + השימוש בספריה זו.

3. קוד (2 נק') - ייבאו את הספריות **pandas, matplotlib.pyplot,geopy**. (במידה ולא הצלחתם לייבא את geopy התקינו אותה ע"י פתיחת cmd והזנת הפקודה: pip install geopy)

4. קוד (2 נק') - קראו את קובץ הנתונים Mobileye_risk_TelAviv_NEW.csv לתוך DataFrame והריצו תיאור סטטיסטי על הנתונים.

(22 נק') חלק 2 – Preprocessing

5. קוד (4 נק') צרו עמודה חדשה ב-DF שיצרתם וקראו לה section_length בה תחשבו באמצעות הפונקציה distance את המרחקים בין נקודת ההתחלה לנקודת הסוף (**במטרים**) עבור כל מקטע.
קטע הקוד הבא מחשב מרחק במטרים עבור כל מקטע, צרו בעזרתו את העמודה:

```
from geopy.distance import geodesic

def distance(row):
    address1 = (row['start_lat'], row['start_long'])
    address2 = (row['end_lat'], row['end_long'])
    return (geodesic(address1, address2).m)   #in meters
```

(רמז: השתמשו בפונקציות: apply, lambda)

6. קוד + הסבר (4 נק') הריצו תיאור סטטיסטי על עמודת **section_length**, מהו המרחק הממוצע של מקטע דרך, מה המרחק של המקטע הקצר ביותר ומהו המרחק של המקטע הארוך ביותר?

7. קוד (4 נק') כעת כשיש בידינו את המרחק ואת המהירות הממוצעת נוכל לחשב את הזמן הממוצע שלוקח לעבור כל מקטע, צרו עמודה חדשה ב-DF וקראו לה "avg_time" אשר תציג את הזמן הממוצע למקטע ע"י החישוב הבא:

$$avg\_time(hour) = \frac{section\_length}{avg\_speed * 1000}$$

* שימו לב שהזמן הממוצע שקיבלנו הינו במונחי שעה, על מנת לקבל את היחידות בשניות הכפילו את הערך ב-3,600 וקבלו ערכים במונחי שניה, הזינו את הערכים במונחי שניה לתוך עמודת avg_time.

$$avg\_time(second) = avg\_time(hour) * 3{,}600$$

8. קוד (2 נק׳) - הסירו באמצעות הפקודה drop את כלל משתני המיקום (start_lat, start_long, end_lat, end_long) ושמרו את התוצאה ב-DF חדש שתיצרו תחת שם חדש לבחירתכם.

9. הסבר (4 נק') - האם כל המשתנים בדאטה ששלנו צריכים לעבור נרמול? במידה ולא, אילו משתנים לא צריך לנרמל?

10. קוד (4 נק') - בצעו נרמול 0-1 למשתנים הרלוונטיים בטבלה. את התוצאה יש לשמור ב-DataFrame חדש בשם X_normalize.

(15 נק') חלק 3 - הרצת K-means יצירת Silhouette, inertia

בחלק זה, נעזר בספריית **sklearn** על מנת ליצור אלגוריתם K-means.

11. קוד (4 נק') - כתוב את הפונקציה: create_kmeans_classifier המקבלת מספר טבעי (שלם וגדול מ-1, יש לבדוק זאת) המייצג את ה-K ומחזירה מסווג K-means עבור ה-K הרלוונטי.

12. קוד (3 נק') - הפעילו את הפונקציה מסעיף (1) על טווח ערכי K (בין 1 ל-15) ועבור כל K, יש לאמן את המודל עבור 2 קבוצות: הראשונה אוסף נתונים ללא נרמול והשניה אוסף הנתונים לאחר נירמול.

13. קוד (3 נק') - עבור כל K, שמרו ברשימה את ערכי סכום המרחקים הממוצע (SSE) של אוסף הנתונים המנורמלים, וברשימה נפרדת חזרו על הפעולה עבור סכום המרחקים הממוצע (SSE) של אוסף הנתונים הלא מנורמלים.

14. קוד (3 נק') - עבור כל K, שמרו ברשימה את ערכי ה-Silhouette של אוסף הנתונים המנורמלים, וברשימה נפרדת חזרו על הפעולה עבור ערכי ה-Silhouette של אוסף הנתונים הלא מנורמלים.

15. קוד (2 נק') - בדקו (ע"י הדפסת אורכי הרשימות) שרשימות ה-Silhouette מכילות 14 איברים (מספר ה-K השונים).

(14 נק') חלק 4 - הצגת הנתונים

בחלק זה, עליכם ליצור הצגות ויזואלזציות מלאות (הכוללת labels לצירים, כותרות, צבעים).

16. קוד (4 נק') - הצג את ערכי סכום המרחקים הממוצעים (SSE) עבור ערכי K בין 2 ל-15 עבור הנתונים <u>הלא-מנורמלים</u>.

17. קוד (4 נק') - הצג את ערכי ה-Silhouette עבור ערכי K בין 2 ל-15 עבור הנתונים הלא-מנורמלים.

18. קוד + הסבר (6 נק') - הצג את ערכי ה-Silhouette עבור ערכי K בין 2 ל-15 עבור הנתונים המנורמלים.

    הוסף לתשובתך, מה ההבדל בערכים בין שאלה 17 לשאלה 18? כיצד בא לידי ביטוי השינוי כאשר

    אוסף הנתונים הוא מנורמל?

מימוש (42 נק')

K-means באמצעות Excel

בחלק זה אנו ננתח חלק מהנתונים שנאספו במהלך הקורונה. המטלה מתבססת על הקובץ: corona_stats.xlsx
כל שורה בנתונים מייצגת מדינה ועבור כל מדינה מפורטים 22 מקורות מזון שונים שמהם נצרך שומן.
עבור כל מדינה מפורטת רמת השומן שנצרכת מכל מקור מזון. רמת צריכה נמוכה מוגדרת על ידי הערך '0' ורמת צריכה גבוהה מוגדרת על ידי הערך '1' (פירוט מקורות המזון מופיע בגיליון "מילון").

בנוסף, קיימים נתונים בגיליון **"סטטיסטיקת קורונה"** המפרטים עבור כ מדינה את אחוז האנשים הסובלים מהשמנת יתר, אחוז החולים המאומתים, אחוז המתים ואחוז המחלימים וכן גודל האוכלוסייה במדינה .

המטרה שלנו היא לבנות מודל שיסווג את המדינות לאשכולות בעלי שונות פנימית נמוכה ושונות חיצונית גבוהה.
מטלה זו יכולה להוסיף יעד נוסף במענה על השאלה האם אורח חיים בריא יכול לסייע בהתמודדות עם נגיף הקורונה .
לפני שתתחילו בפתרון המטלה, נא מלאו מספר תעודת זהות, שם פרטי ושם משפחה בגיליון "פרטים אישיים ."

חלק 1: הכנת הקלט עבור אלגוריתם K-MEANS

19. הסבר (2 נק') - הגדירו במילים שלכם: כיצד תבטאו על מדינה באמצעות נקודת נתונים במרחב (data-point) ?

20. אקסל (2 נק') - בגיליון "Q2" הפכו את הנתונים המופיעים בגיליון "נתונים" לסט של נקודות נתונים.

טיפ: תוכלו להשתמש באפשרות הדבקת ערכים והחלפתם VALUES PASTE ו-TRANSPOSE PASTE

חלק 2 :יישום אלגוריתם K-MEANS לקיבוץ נקודות הנתונים

משימה 1: בחירת מספר הסנטרואידים

21. אקסל (2 נק') - בגיליון "Q3" הגדירו אקראית את מיקומי 4 הסנטרואידים הראשוניים. השתמשו באקסל כדי לייצג את הסנטרואידים שבחרתם.

טיפ: בחירת המיקומים של הסנטרואידים אינה ייחודית. וודאו כי הינכם מבינים מדוע.

משימה 2: הקצאה ראשונית של נקודות הנתונים לסנטרואידים

22. אקסל (2 נק') - הציגו את ארבעת הסנטרואידים המופיעים בגיליון "Q3" ואת נקודות הנתונים של המדינות המופיעות בגיליון "Q2"כך שהנתונים יוצגו זה לצד זה.

23. אקסל (3 נק') - חשבו בשורות 26-29 את המרחק של כל אחת מהנקודות מכל אחד מהסנטרואידים.

24. אקסל (3 נק') - חשבו עבור כל אחת מנקודות הנתונים את המרחק הקצר ביותר מבין המרחקים שחישבתם בשאלה 22 ומלאו זאת בשורה 31.

25. אקסל (3 נק') - הקצו עבור כל אחת מנקודות הנתונים את מספר הסנטרואיד שהמרחק אליו הינו הקצר ביותר ומלאו זאת בשורה 32 .

טיפים :העתיקו את הסנטרואידים (גיליון Q3) ונקודות הנתונים (גיליון Q2) לגיליון חדש בשם "Q4". לאחר מכן, הקצו את נקודת הנתונים הראשונה לסנטרואיד שאליו היא הכי קרובה, לאחר מכן גררו את הנוסחאות עד שתגיעו למדינה האחרונה.
בדיוק כמו שביצענו בתרגול
(רמז: השתמשו בסימן ה-$ במסגרת החישובים שלכם כדי להבטיח חישובים מהירים ויעילים עבור כל נקודות הנתונים).

משימה 3: הגדרת בעיית האופטימיזציה (קביעת המיקום האופטימאלי של הסנטראידים)

26. הסבר (3 נק׳) - הגדירו באופן מילולי מהי פונקציית המטרה של בעיית האופטימיזציה, מהם משתני ההחלטה ומהם האילוצים?

27. אקסל (3 נק׳) - העתיקו את גיליון "Q4" לגיליון חדש ושנו את שמו של הגליון החדש ל-"Q11".
בגיליון החדש, חשבו בתא F35 את הערך הראשוני של פונקציית המטרה על ידי שימוש בחישוב סכום ריבועי השגיאות (שגיאה = המרחק האוקלידי של נקודת נתונים לסנטרואיד הקרוב ביותר אליה).

28. אקסל (4 נק׳) - העתיקו את גיליון "Q11" לגיליון חדש ושנו את שמו של הגיליון החדש ל-"Q12". פתרו את בעיית האופטימיזציה על ידי שימוש ב-SOLVER תוך שימוש בשיטת הפתרון "EVOLUTIONARY". שנו את צבע התא שבו מופיע הערך החדש של פונקציית המטרה .

משימה 4: הצגת התוצאות

29. אקסל (4 נק') - בגיליון חדש בשם "Q13" הציגו 4 רשימות של מדינות, המסודרות לפי הקלאסטרים שהתקבלו בשאלה 12.
כל רשימה תכלול בראשה כותרת ובה יהיה רשום מספר הקלאסטר (לדוגמה קלאסטר #1), תחת כל אחת מהכותרות הציגו את רשימת שמות המדינות ("מדינה") שהוקצו לאותו קלאסטר.

משימה 5: ויזואליזציה

30. הסבר (6 נק') - זהו את המאפיינים הבאים:
א. (3 נק') מהו הגודל של כל קלאסטר (מספר המדינות בכל קלאסטר)?
ב. (3 נק') מהם מקורות המזון המובהקים ביותר בכל קלאסטר מהם נצרך שומן?

הציגו בצורה ויזואלית מענה לשאלות אלו .

לצורך מענה על שאלות אלו היעזרו בתוצאות האלגוריתם שהתקבלו בשאלות 28-29 וכן בנתונים המופיעים בגיליון "סטטיסטיקת קורונה ."

חלק 3 :הערכת התוצאות (EVALUATION)

משימה 6: חישוב סילואט של נקודת נתונים

31. אקסל (5 נק') - בגיליון "Q12" העריכו את ההקצאה לקלאסטר של נקודת הנתונים עבור "ישראל ."לשם כך, עליכם לחשב את הסילואט של נקודת הנתונים של מקורות המזון שמהן נצטרך שומן ב" -ישראל". שימו לב, בגליון זה צריכים להופיע **כל החישובים הרלוונטים לחישוב הסילואט** .

טיפים:

I. חשבו את המרחקים של נקודת הנתונים של ישראל מנקודות הנתונים של כל שאר המדינות.
II. לאחר מכן, חשבו את ממוצע המרחקים של כל הנקודות השייכות לאותו הקלאסטר אליו שייכת ישראל (תוכלו להיעזר בפונקציית AVERAGEIF שכוללת 3 פרמטרים: 1) הטווח שעליו בודקים את הקריטריון; 2) הקריטריון הנבחן (במקרה שלנו מספר הקלאסטר שאליו משויכת "ישראל"; 3) הטווח שעליו מחשבים את הממוצע).
III. עבור שלושת הקלאסטרים הנותרים השונים מהקלאסטר שאליו משויכת "ישראל", חשבו את המרחקים של נקודת הנתונים של "ישראל" מנקודות הנתונים של כל שאר המדינות המשתייכות לכל אחד מקלאסטרים אלו.
IV. מצאו מאיזה קלאסטר (השונה מהקלאסטר אליו משוייכת "ישראל") המרחק הממוצע הינו הנמוך ביותר זהו בעצם, הקלאסטר השכן.
V. השתמשו בנוסחת חישוב הסילואט שנלמדה בכיתה ובתרגול.

בהצלחה!