למידה סטטיסטית - תרגיל בית מספר 2

עיבוד מקדים של נתונים ו- *KNN*
ישום ב- *Python*

מרצה: מר' אבי זכאי

מתרגל: מר' סאלח אבו שאהין

מר' אנדריאס נסייר

הערות:

- יש לענות על כל השאלות כולל תיעוד הקוד
- תאריך אחרון להגשת התרגיל: **6.1.2025 בשעה 23:55**
- **חשוב מאוד לא להשתמש באף פונקציה מוכנה מראש של פייתון שעושה את הסיווג אלא אתם צריכים לכתוב את הקוד בעצמכם כמו שהצגנו בכיתה.**
- **ההגשה בזוגות**

בעבודת בית זו אנו הולכים לממש את אלגוריתם KNN באופן מלא על בסיס נתוני רכבים שונים המצורף למטלה. המטרה שלנו לסווג את הרכבים השונים לשני סוגי דלק: [Petrol, Diesel].

קבוצת האימון שלנו מכילה מידע על 2 סוגי דלק אלו של הרכבים כאשר לכל רכב נאספו ארבעה תכונות :

1. מחיר רכב (Price)
2. שנת מכירה (Year)
3. מספר קילומטרים שהרכב נסע (Kilometer length)
4. אורך (Length)

5. רוחב (Width)
6. אורך (Height)
7. מספר מקומות ישיבה (Seating Capacity)
8. קיבולת מיכל דלק (Fuel Tank Capacity)
9. רכב (Make)
10. מיקום מכירה (Location)

למטלה זו מצורף קובץ פייתון המכיל קטע קוד קצר. עם הרצת הקוד תקבלו את הנתונים הבאים:

1. trainingSet , מערך של numpy בממדים של 1508x8 המכיל את הנתונים של 1508 רכבים שמכרו (כל שורה זה רכב בודד) משתי סוגי דלק שונים (Petrol, Diesel), כל עמודה מייצגת תכונה אחת משמונת התכונות הראשונות **(הכמותיות בלבד)**.
2. trainingTargets , מערך של numpy בגודל 1508 המכיל את הסיווגים האמיתיים של כל אחד מהדגימות ב 1508 דגימות האימון. הסיווגים הם: ['Petrol', 'Diesel'].
3. testSet, מערך של numpy בממדים של 378x8 המכיל את הנתונים של 378 רכבים שנמכרו (כל שורה זה רכב בודד) הרכבים לא מסווגים, כל עמודה מייצגת תכונה אחת משמונת התכונות הראשונות שהזכרנו קודם **(הכמותיות בלבד)**.. זו הקבוצה שעליכם לסווג אותה לאחד משני סוגי הדלק.
4. testTargets, מערך של numpy בגודל 378 המכיל את הסיווגים האמיתיים של קבוצת הבדיקה. מערך זה ישמש אותנו אחרי סיווג קבוצת הבדיקה testSet בעזרת אלגוריתם KNN על מנת לבדוק את מידת ההצלחה של האלגוריתם שלנו.

<u>הוראות לעבודת בית:</u>

הורידו את קובץ הנתונים והפייתון למחשב האישי שלכם **ושמרו את שניהם באותה תיקייה** (בלי לשנות את שם קובץ הנתונים). הריצו את קובץ הפייתון בסביבת Spyder, בשלב זה אתם אמורים לקבל את ערכי המשתנים לעיל. תחקרו את הנתונים על מנת לוודא שהינכם מבינים את ארבעת המערכים שהזכרנו למעלה, כמו כן יש לחקור את הנתונים השמורים במשתנה dataSet ולהתייחס לסוגי הפיצ'ירים בנתונים. **שימו לב שסקריפט הקוד שיש לכם מוריד בהמשך את התכונות הקטיגוריאליות Make ו' Location. אתם לא צריכים לדאוג לטפל בזה.**

- כתבו פונקציה max_min_Scalling(trainingSet, testSet) המקבלת שני מערכי numpy מנרמלת אותם לפי נרמול min/max ומחזירה אותם. שימו לב, הנרמול של שני המערכים נעשה לפי ה min/max בקבוצת האימון (trainingSet).

- כתבו פונקציה euclidean_distance(NtrainingSet,vec) המקבלת דגימה בודדת מקבוצת הבדיקה (vec) ומקבלת קבוצת האימון המנורמלת (NtrainingSet). הפונקציה תחשב את המרחק האוקלידי של vec מכל דגימה מקבוצת האימון ותחזיר את ויקטור המרחקים.
- כתבו פונקציה manhattan_distance(NtrainingSet,vec) המקבלת דגימה בודדת מקבוצת הבדיקה (vec) ומקבלת קבוצת האימון המנורמלת (NtrainingSet). הפונקציה תחשב את מרחק מנהטן של vec מכל דגימה מקבוצת האימון ותחזיר את ויקטור המרחקים.
- כתבו פונקציה predict(k, distance, trainingTargets) המקבלת את וקטור המרחקים של הדגימה שאנו מסווגים, מקבלת את הסוגים של כל דגימה בקבוצת האימון ואת K המסמן את מספר השכנים. הפעולה תחזיר את הסיווג (Class), של הדגימה לפי K השכנים הקרובים ביותר KNN.
- תבנו פונקציה ראשית main_knn(k) הפונקציה מקבלת את מספר השכנים k , משתמשת בפונקציות שבנינו קודם ומיישמת את האלגוריתם KNN על מנת לסווג את כל אחד מהרכבים בקבוצת הנתונים testSet. **לצורך מציאת K השכנים אנחנו נשתמש במרחק אוקלידי בלבד.**

  את תוצאות הסיווג יש לשמור במערך של numpy בגודל 378 בשם result. על הפונקציה לחזור על תהליך הסיווג עבור $k = 1,3,5,7,9,11$. ולהדפיס את תוצאות הסיווג בכל פעם. כמו כן הפעולה תשווה בין תוצאות הסיווג result לבין הסוגים האמיתיים של קבוצת הבדיקה testTargets ותדפיס את אחוז ההצלחה שלנו בסיווג.
- **בסוף עליכם להגיש את קובץ הפייתון שמכיל את הקוד שלכם. שימו לב קוד לא רץ לא יקבל ניקוד כלל.**

כמה דגשים:

1. **המשתנים: testTargets, testSet, trainingSet, trainingTargets**
   הם משתנים גלובליים כלומר תוכלו להשתמש בהם ישירות בפונקציה הראשית.

2. **אחוז ההצלחה של המסווג (Accuracy) מוגדר כמספר הדגימות שהצלחנו לסווג אותם נכון בקבוצת הבדיקה חלקי סך כל הדגימות בקבוצת הבדיקה.**
   כלומר אם הצלחנו לסווג נכון 25 דגימות בקבוצת הבדיקה מתוך ה 30 שהיו לנו אז ההצלחה שלנו היא: $\frac{25}{30} = 0.8333$ או 83.33%.

בהצלחה!