נושאי ליבה בבינה מלאכותית

גרסה 1.0 - תשפ"ו

פיתוח היחידה
- אריאל בר-יצחק
- ד"ר קובי מייק

ייעוץ והכוונה פדגוגית
- פרופ' אורן קורלנד, הפקולטה למדעי הנתונים וההחלטות, הטכניון

תוכן עניינים

מבוא

יחידה 5 עוסקת בנושאים תיאורטיים וסוגיות ליבה בלמידת מכונה ובינה מלאכותית. יחידה זו מהווה חטיבה בבינה מלאכותית - למידת מכונה יחד עם יחידת המעבדה (3) בלמידת מכונה. יחידה זו משלימה את הידע המעשי הנרכש ביחידה 3. הדגש ביחידה 5 הוא data centric, וההעמקה התיאורטית היא בנושאים הבאים: רקע מתמטי, רקע סטטיסטי, חקר נתונים (EDA), מודלי סיווג מבוססי מרחק ותהליך הפיתוח של מודל בינה מלאכותית.

מטרת היחידה

העמקת והרחבת הידע התיאורטי בנושאי ליבה בבינה מלאכותית: תהליך חקר נתונים ותהליך פיתוח מודל של למידת מכונה.

רשימת נושאים

יחידה 5 מעמיקה ומרחיבה את הידע המוקנה ביחידה 3 (פרויקט בלמידת מכונה). רשימת הנושאים הנלמדים ביחידה 5 מוצגת לכן ביחד עם הפרקים של יחידה 3. ניתן ללמד את יחידה 5 גם לתלמידיות שלא למדו את יחידה 3 בלמידת מכונה, במקרה זה באחריות המורה להשלים לתלמידיות את הידע החסר מחוץ למסגרת השעות של היחידה.

נושא	יחידה 3 - פרויקט בבינה מלאכותית 90 שעות	יחידה 5 - נושאי ליבה בבינה מלאכותית 90 שעות
תהליך העבודה במדעי הנתונים	שאלות מחקר איתור מאגרי נתונים ביצוע פרויקט	
בסיס מתמטי	L2 distance	Vectors in R^n L1 & L2 norm L1 & L2 distances inner product similarity cosine similarity cosine distance
קלט והכנת נתונים	ספריית pandas, ספריית numpy. זיהוי חסרים, כפולים וחריגים וטיפול בהם. המרת משתנים קטגוריאליים תקנון ונירמול הנדסת מאפיינים איזון תגיות	
EDA	סוגי משתנים התפלגות נורמלית סטטיסטיקה תיאורית ויזואליזציה	מדדי מרכוז ופיזור התפלגות נורמלית תקנון ונירמול יחסים בין משתנים קריאת גרפים

EDA - clustering		Intro - motivation K-means Evaluation: WCSS, Silhouette
מודלים	KNN רגרסיה לינארית ומדד R2 Perceptron & SVM Overfitting and underfitting Decision boundaries	Nearest Centroid KNN Overfitting and underfitting Decision boundaries
תהליך אימון של אלגוריתם לומד	Train, Validation and Test splits Hyperparameter Tuning Cross Validation Confusion Matrix Accuracy Precision, Recall, F1	Train, Validation and Test splits Hyperparameter Tuning Cross Validation Confusion Matrix Accuracy Precision, Recall, F1 Binary Decision Threshold

הקצאת שעות לימוד לפי פרק

פרק	תת פרק	עיוני	מעשי	סך שעות
מבוא		1		1
פרק 1 - בסיס מתמטי (9 שעות)	1.1 וקטורים	1	4	5
	1.2 מרחקים בין וקטורים	1	3	4
פרק 2 - בסיס סטטיסטי (26 שעות)	2.1 מדדי מרכוז ופיזור	2	3	5
	2.2 התפלגות נורמלית	2	3	5
	2.3 תקנון ונרמול	2	3	5
	2.4 יחסים בין משתנים	1	2	3
	2.5 קריאת גרפים	4	4	8
פרק 3 - למידה לא מונחית (21 שעות)	3.1 מבוא ללמידה לא מונחית	3	2	5
	3.2 אלגוריתם k means	2	6	8
	3.3 מדדי ביצוע k means	2	6	8
פרק 4 - למידה מונחית (25 שעות)	4.1 מבוא ללמידה מונחית	3	2	5
	4.2 אלגוריתם Nearest Centroid	3	4	7
	4.3 אלגוריתם KNN	3	4	7
	4.4 אלגוריתמי סיווג מבוסס מרחק - מבט על	3	3	6
פרק 5 - מדדי ביצוע (8 שעות)	5.1 הערכת ביצועים	2	2	4
	5.2 ולידציה וחקר פרמטרים	2	2	4
סה"כ		**37**	**53**	**90**

למידה דיפרנציאלית

להלן רשימת נושאי רשות שנועדו לטובת למידה דיפרנציאלית בכיתות - נושאים אלו לא יופיעו בבחינת הבגרות ונועדו להרחבה והעמקה במידה ותכנון הזמן מאפשר זאת.

פרק	נושא	הערות
1.1	ייצוג פולרי של וקטורים	
2.2	פונקציות התפלגות pdf,cdf	
3.1	זיהוי חריגים	
3.1	הקטנת ממדיות t-SNE, UMAP, PCA	
3.2	אלגוריתם האתחול הסטטיסטי K-Means++	נדרשת הבנה של k-means עם אתחול אקראי
3.3	מדד Silhouette	חישוב המדד וכתיבת קוד - רשות. ניתוח ביצועים על פי המדד - חובה.
4.3	סיבוכיות של אלגוריתם KNN ו- nearest centroid	
5.2	GridSearchCV	כתיבת קוד - רשות ניתוח פלט - חובה

פרק 1 - בסיס מתמטי - אלגברה לינארית

מבוא

עיסוק בבינה מלאכותית דורש רקע מתמטי בהיקף שאינו נלמד בתוכנית הלימודים בתיכון. היקף הרקע המתמטי הנדרש רב וכולל נושאים באלגברה לינארית, חשבון דיפרנציאלי, סטטיסטיקה, הסתברות ועוד. בפרק זה ילמדו נושאים נבחרים מתוך אלגברה לינארית הנדרשים להבנת התוכן בבינה מלאכותית. מיקוד פרק זה בהצגת המושג וקטור, הכללה של וקטורים בR^n ומדדי מרחק בין וקטורים.

קווים מנחים להוראת הפרק

- יש לשלב גם תרגיל דף ועט וגם תרגיל קוד. חלק מהתרגילים ידרשו חישובים באמצעות מחשבון ופתרון על דף, בדומה להוראת מתמטיקה, וחלק מהתרגילים ידרשו כתיבת קוד.
- מומלץ להשתמש בשפת python אולם אין מניעה להשתמש בשפה אחרת. בחינת הבגרות תתקיים בשפת python.
- מומלץ לעבוד בסביבת Google Colab או סביבת Jupyter Notebooks אחרת.

רשימת הנושאים

- וקטורים בR^n
- מרחק ודמיון בין וקטורים
  - מרחק L1, L2
  - inner product similarity
  - cosine similarity
  - מרחק cosine

נושא 1.1 וקטורים בRn

תלמידים/ות רבים לומדים את מושג הוקטור במסגרת לימודי פיזיקה. על מנת לקשר מושג הוקטור הכללי הנלמד ביחידה זו ללימודים קודמים ומקבילים, נתחיל בהוראה של וקטור כגודל פיזיקלי המייצג גודל עם כיוון בשני ממדים, ונמשיך לוקטור כללי בRn.

מושגים

- ★ סקלר
- ★ וקטור
- ★ ייצוג קרטזי
- ★ **(רשות)** ייצוג פולארי. תלמידי/ות פיזיקה מכירים את המונח וקטור מלימודי הפיזיקה כגודל עם כיוון. עבור תלמידים אלו יש להציג את הקשר בין הייצור הפולארי המוכר מפיזיקה לבין הייצוג הקרטזי החדש, על מנת להבנות להם המחשה אחודה של מושג הוקטור.
- ★ חיבור, חיסור, כפל dot, כפל איבר איבר
- ★ נורמה L1, נורמה L2

מטרות ביצועיות

- הגדרת וקטור ב Rn

- חישוב חיבור וחיסור וקטורים ב Rn
- חישוב כפל של וקטור בסקלר
- חישוב כפל איבר-איבר של שני וקטורים (Element-wise multiplication)
- חישוב מכפלה פנימית של שני וקטורים (Dot product)

$$dot\_product(u,v) = u \cdot v = \sum_{i=1}^{n} u_i v_i$$

- חישוב אורך (L1 norm) של וקטור

$$||V||_1 = \sum_{i=1}^{n} |v_i|$$

- חישוב אורך (L2 norm) של וקטור

$$||V||_2 = \sqrt{v \cdot v} = \sqrt{\sum_{i=1}^{n} v_i^2}$$

- **(רשות)** הבנת ייצור וקטור בייצוג פולרי (גודל וזווית)

$$v = (r, \phi)$$

- **(רשות)** הבנת ייצוג וקטור בייצוג קרטזי (גודל בכל מימד)

$$v = (x, y)$$

- **(רשות)** המרה מייצוג קרטזי לפולארי

$$r = ||v||_2 = \sqrt{x^2 + y^2}$$
$$\phi = atan2(y, x)$$
$$v = (x, y) \equiv (r, \phi)$$

- **(רשות)** המרה מייצוג פולארי לקרטזי

$$x = r \cos\phi$$
$$y = r \sin\phi$$

- חישוב נרמול וקטורים

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$$

$$\|\hat{\mathbf{v}}\| = \left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\| = \frac{\|\mathbf{v}\|}{\|\mathbf{v}\|} = 1$$

- זיהוי וקטור מנורמל כוקטור שהנורמה שלו היא 1.
- הסבר/תיקון/השלמת קוד של פעולות וקטורית עם ספריית numpy

נושא 1.2 דמיון ומרחק בין וקטורים

הרקע המתמטי של הלומדים כולל חישוב מרחק בין וקטורים על ידי משפט פיתגורס. ביחידה 5 נרחיב את ההבנה של מרחק אוקלידי לRn, וכן נלמד דרכים נוספות למדידת דמיון ומרחק בין וקטורים.

מושגים

- ★ L1 & L2 distances
- ★ inner product similarity
- ★ cosine similarity
- ★ cosine distance

מטרות ביצועיות

- חישוב מרחק בין שני וקטורים

$$d(u,v) = ||u - v||$$

- חישוב מרחק מנהטן L1 Norm בין שני וקטורים. הבנת המשמעות הגיאומטרית.

$$d_1(u,v) = ||u - v||_1 = \sum_{i=1}^{n} \left|(u - v)_i\right|$$

- חישוב מרחק אוקלידי L2 Norm בין שני וקטורים. הבנת המשמעות הגיאומטרית.

$$d_2(u,v) = ||u - v||_2 = \sqrt{\sum_{i=1}^{n} ((u - v)_i)^2}$$

- חישוב inner product similarity. הבנת המשמעות הגיאומטרית.

$$inner\_sim(u,v) = dot\_product(u,v) = u \cdot v$$

$$inner\_sim(u,v) = ||u||_2 \, ||v||_2 cos\,\theta$$

- חישוב cosine similarity ומרחק. הבנת המשמעות הגיאומטרית.

$$cos\_sim(u,v) = \frac{u \cdot v}{||u||_2 \, ||v||_2} = cos\,\theta$$

$$d_{cos}(u,v) = 1 - cos\_sim(u,v) = 1 - cos\,\theta$$

- הבנת ההבדלים בין השיטות השונות למדידת דמיון ומרחק בין וקטורים והמקרים בהם כדאי להשתמש בכל אחד מהם.
- הסבר/תיקון/השלמת קוד של חישוב מרחקים

פרק 2 - בסיס סטטיסטי לחקר נתונים

מבוא

עיסוק בבינה מלאכותית דורשת רקע מתמטי בהיקף שאינו נלמד בתוכנית הלימודים בתיכון. היקף הרקע המתמטי הנדרש רב וכולל נושאים באלגברה לינארית, חשבון דיפרנציאלי, סטטיסטיקה, הסתברות ועוד. בפרק זה ילמדו נושאים נבחרים מתוך סטטיסטיקה הנדרשים להבנת התוכן בבינה מלאכותית.

מושגי היסוד בסטטיסטיקה בפרק זה נלמדים בקונטקסט של חקר נתונים (Exploratory Data Analysis) על מנת לספק את הבסיס התיאורטי הנדרש להבנת מידול סטטיסטי של הנתונים, תקנון, זיהוי חריגים, ייצוג נתונים ועוד.

קווים מנחים להוראת הפרק

- לתלמידיות אין את הרקע הנדרש במתמטיקה, סטטיסטיקה והסתברות על מנת להבין לעומק את כל המושגים ביחידה זו, לדוגמא צפיפות הסתברות. לכן חלקים מהיחידה ילמדו דרך הסברים אינטואיטיביים ולא באופן מתמטי מלא.
- בהוראת היחידה מומלץ להשתמש בהדגמות בפיתון עם מאגרי נתונים אמיתיים, לדוגמא iris, penguin וכד'.

רשימת הנושאים

- מדדי מרכוז ופיזור
  - ממוצע
  - חציון
  - רבעון ראשון Q1
  - רבעון שלישי Q3
  - טווח בין רבעוני IQR
  - שכיח
  - שונות
  - סטיית תקן
- התפלגות נורמלית
  - התפלגות נורמלית תקנית
  - μ
  - σ^2
- תקנון ונרמול
- יחסים בין משתנים
- קריאת גרפים

נושא 2.1 - מדדי מרכוז ופיזור

הערה:

When ML transforms data → use population formulas (n).
When ML tries to discover patterns → use sample formulas (n−1).

מושגים

- ★ אוכלוסיה ומדגם
- ★ ממוצע של אוכלוסיה ומדגם

★ חציון
★ שכיח
★ התפלגות סימטרית ואסימטרית (skewed)
★ אסימטריה בעלת זנב ימני (right skewed)
★ אסימטריה בעלת זנב שמאלי (left skewed).

★ שונות של אוכלוסיה ומדגם
★ סטיית תקן של אוכלוסיה ומדגם

<u>מטרות ביצועיות</u>

- הגדרת אוכלוסיה ומדגם
- חישוב ממוצע (mean) של אוכלוסיה ומדגם:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- חישוב חציון (median). בנוסחאות נניח שהאינדקס מתחיל ב 0.

$$sort\ values \Rightarrow x_0 \le x_1 \le \ldots \le x_{n-1}$$

$$n\ odd \Rightarrow median = x\left[\frac{n}{2}\right]$$

$$n\ even \Rightarrow median = \frac{x\left[\frac{n}{2}-1\right] + x\left[\frac{n}{2}\right]}{2}$$

- חישוב רבעונים (Q1,Q3). נשתמש בשיטת החלוקה לחצאים (Tukey's hinges).
  - שלב 1: מיון הערכים
  - שלב 2: מציאת החציון
  - שלב 3: נפצל את הערכים לחצי תחתון + עליון, ללא החציון
  - שלב 4: נמצא את החציון של חצי תחתון (Q1) וחצי עליון (Q3).
- הגדרה וחישוב טווח בין רבעוני IQR = Q3 - Q1
- חישוב שכיח (mode)

$$mode = arg\,max_x\, frequency(x)$$

- הבנת הבדלים בין היסטוגרמה של התפלגות סימטרית ואסימטרית (skewed).
- זיהוי היסטוגרמה של התפלגות עם אסימטריה בעלת זנב ימני ואסימטריה בעלת זנב שמאלי.
- זיהוי המיקום של ממוצע, חציון ושכיח בגרף של התפלגות סימטרית וא-סימטרית.
- בחירת מדד מירכוז: ממוצע, חציון, או שכיח בהתאם לסוג הבעיה על בסיס הבנת ההבדלים ביניהם
- חישוב של שונות וסטיית תקן לאוכלוסיה:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} (x_i - \mu)^2 \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} (x_i - \mu)^2}$$

- חישוב של שונות וסטיית תקן למדגם:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- הסבר מדוע יש הבדל בין החישוב של שונות וסטיית תקן לאוכלוסייה ולמדגם.
- בחירת מדד פיזור: שונות, סטיית תקן או IQR בהתאם לסוג הבעיה על בסיס הבנת ההבדלים ביניהם
- הסבר/תיקון/השלמת קוד המחשב מדדי מרכוז ופיזור
- ניתוח תוצאות של df.describe

נושא 2.2 - התפלגות נורמלית

מושגים

★ התפלגות נורמלית כללית ותקנית
★ ציון תקן (Z-Score)
★ כלל 68% - 95% - 99.7%
★ **(רשות)** גרף PDF
★ **(רשות)** הסתברות מצטברת CDF

דגשים פדגוגיים

❖ לא מלמדים אינטגרלים. האינטגרל משמש כאן אך ורק לסימון שטח מתחת לגרף מ a ועד b.
❖ לא מלמדים את צפיפות ההסתברות. במקום זאת מסבירים שהגרף PDF מתאר את נוסחת ההתפלגות הנורמלית. ומדברים על השטח מתחת לגרף כהסתברות.

מטרות ביצועיות

- הסבר של התפלגות נורמלית כמתארת תכונות באוכלוסייה כגון גובה, משקל וכד'.
- ניתוח גרף של התפלגות נורמלית - זיהוי הממוצע והשונות מתוך הגרף.
- חישוב של ציון תקן (Z-score)

$$z = \frac{x-\mu}{\sigma}$$

- חישוב ערך מקורי מתוך ציון תקן.
- זיהוי ציון Z על גבי הגרף של התפלגות נורמלית.
- זיהוי רמת חריגות וזיהוי חריגים באמצעות ציון תקן (Z-score).
- ביצוע חישובים עם כלל 68% - 95% - 99.7% (סטיית תקן אחת, שתיים ושלוש בהתאמה).
- **(רשות)** זיהוי פונקציית ההתפלגות של התפלגות נורמלית כללית וזיהוי הממוצע והשונות בנוסחה

$$pdf(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- **(רשות)** זיהוי פונקציית ההתפלגות של התפלגות נורמלית תקנית וזיהוי כי הממוצע=0 והשונות=1

$$pdf(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})$$

- **(רשות)** הסבר התכונות של פונקציית התפלגות:
  a. אי שלילית
  b. השטח מתחת לגרף PDF מייצג את ההסתברות לקבל נתון בתחום הנבדק.

$$P(a \le X \le b) = \int_a^b pdf(x)dx$$

  c. השטח כולו שווה 1.

$$\int_{-\infty}^{\infty} pdf(x)dx = 1$$

- **(רשות)** הסבר התכונות של פונקציית הסתברות מצטברת CDF כמתארת את ההסתברות לקבל ערך קטן מ-a.

$$cdf(a) = P(X \leq a) = \int_{-\infty}^{a} pdf(x)dx$$

- **(רשות)** חישוב ההסתברות לערך קטן מ-a באמצעות טבלת CDF או ציון תקן או פונקציית ספריה norm.cdf.
- **(רשות)** חישוב ההסתברות בטווח (a,b) באמצעות טבלת CDF או ציון תקן או פונקציית ספריה norm.cdf.

$$P(a \leq X \leq b) = \int_{a}^{b} pdf(x)dx = cdf(b) - cdf(a)$$

- **(רשות)** חישוב ההסתברות לערך גדול מ-a באמצעות טבלת CDF או ציון תקן או פונקציית ספריה norm.cdf.

$$P(X \geq a) = \int_{a}^{\infty} pdf(x)dx = 1 - cdf(a)$$

- **(רשות)** הסבר/תיקון/השלמת קוד המחשב הסתברויות על ידי norm.cdf.
- זיהוי התפלגויות נורמליות בגרף היסטוגרמה

<u>נושא 2.3 - תקנון ונרמול</u>

הערה: בתהליך נרמול הנתונים והכנתם למודל נתייחס לנתונים כאוכלוסיה כולה

<u>מושגים</u>

★ standard scaler - תקנון
★ minmax scaler - נרמול
★ robust scaler - תקנון
★ norm scaler - נרמול וקטורי

<u>מטרות ביצועיות</u>

- חישוב הפרמטרים וביצוע תקנון עם Standard Scaler

$$x' = \frac{x - \mu}{\sigma}$$

- חישוב הפרמטרים וביצוע נרמול עם Minmax Scaler

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- חישוב הפרמטרים וביצוע תקנון עם Robust Scaler

$$x' = \frac{x - median}{IQR}$$

- חישוב נרמול וקטורי Normalizer Scaler. יש לציין ששיטת נרמול זאת פועלת על הדוגמאות ולא על

המאפיינים. ולכן היא שונה מהותית מיתר השיטות שצויינו.

$$x' = \frac{x}{||x||}$$

- הסבר המקרים בהם נדרש תקנון ונרמול כהכנה ליצירת מודל של מכונה לומדת
- הסבר על היתרונות וחסרונות של כל אחד מהאלגוריתמים
- בחירת אלגוריתם תקנון/נרמול בהתאם לסוג הבעיה
- ניתוח ההבדלים בין ביצועים של אלגוריתמי תקנון/נרמול שונים בבעיה אמיתית.
- הסבר/תיקון/השלמת קוד של תקנון/נרמול

נושא 2.4 - יחסים בין משתנים

הערה: בתהליך הבנת הקשרים בין המשתנים נתייחס לנתונים כמדגם.

מושגים

★ שונות משותפת (covariance)
★ מתאם או קורלציה (correlation)
★ סיבתיות (causation)

מטרות ביצועיות

- חישוב שונות משותפת (covariance)

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- הסבר הקשר בין משתנים על פי השונות משותפת (covariance) חיובית, שלילית ואפס
- חישוב של מתאם או קורלציה (correlation)

$$r_{xy} = \frac{cov(x,y)}{s_x s_y} \qquad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- הסבר הקשר בין משתנים על פי קורלציה חיובית, שלילית ואפס.

$$0 \le r_{xy} \le 1$$

- הסבר ההבדל בין מתאם (correlation) לבין סיבתיות (causation). זיהוי האם בבעיה נתונה אפשר להסיק סיבתיות או קורלציה.
- הסבר/תיקון/השלמת קוד של חישוב שונות משותפת וקורלציה

נושא 2.5 - קריאת גרפים

מושגים

★ היסטוגרמה (כולל הסבר אינטואיטיבי על KDE המופיע בגרף)
★ גרף פיזור - scatter plot
★ גרף עמודות - bar plot
★ גרף ספירה - count plot
★ box plot

מטרות ביצועיות

- הסקת מסקנות מהיסטוגרמה, זיהוי נורמליות, זיהוי אסימטריות, הערכה של ממוצע ושונות
- הסקת מסקנות מגרף פיזור, זיהוי מגמות לינאריות ולא לינאריות, הערכת הקורלציה
- הסקת מסקנות מגרף עמודות, זיהוי הממוצע והערכה של סטיית התקן
- הסקת מסקנות מגרף count
- הסקת מסקנות מגרף box, זיהוי חציון ורבעונים 1,3, זיהוי IQR, זיהוי חריגים

פרק 3 - קיבוץ נתונים - clustering

מבוא

למידה לא מונחית היא פרדיגמה חשובה בלמידת מכונה, הן תיאורטית והן מעשית. למידה לא מונחית כוללת אלגוריתמים המחפשים תבניות בנתונים מבלי שיש תגיות שסומנו מראש. לדוגמא ביישום של anomaly detection האלגוריתם לומד את ההתפלגות הרגילה של הנתונים מתוך הנתונים עצמם ולומד לזהות מצבים חריגים גם אם סט האימון לא כלל תיוג של נתונים חריגים או אף לא כלל דוגמאות של נתונים חריגים.

למידה לא מונחית שימושית מאוד הן למחקר של נתונים כחלק משלב exploratory data analysis, הן לזיהוי תבניות סמויות בנתונים לא מתוייגים, והן לאפליקציות שונות בעולם העסקי. לדוגמא בעולם השיווק קיבוץ נתונים משמש ליצירת פילוח לקוחות המאפשר התאמת קמפיינים מדויקים; ברפואה הוא מסייע בזיהוי תתי־סוגים של מחלות על סמך נתוני ביומרקרים; בתחום אבטחת המידע הוא תורם לזיהוי אנומליות והתנהגויות חשודות; ובעיבוד תמונה הוא משמש לקיבוץ פיקסלים לצורך חלוקה לאזורים בתמונות. במקרים רבים קיבוץ נתונים מהווה שלב ראשוני חשוב שמאפשר להבין את מבנה הנתונים ולהניח יסודות ליישומים מתקדמים יותר.

פרק זה מתמקד בנושא אחד מתוך למידה לא מונחית: קיבוץ נתונים (Clustering). קיבוץ נתונים מאפשר לזהות מבנים פנימיים ויחסים סמויים בתוך נתונים ללא צורך בתוויות מראש. באמצעות חלוקת הנתונים לקבוצות הומוגניות על בסיס דמיון, אפשר לחשוף תובנות שלא היו נגלות בעיבוד רגיל. קיבוץ נתונים מאפשר לסכם נתונים באופן שנותן יותר אינפורמציה מאשר סטטיסטיקה תיאורית. חשיבותו של קיבוץ הנתונים נעוצה בכך שהוא מאפשר לארגן מידע מורכב, לבצע הפחתת מורכבות, לגלות דפוסים חבויים ולסייע בקבלת החלטות מושכלת בתחומים מגוונים.

רשימת הנושאים

- ❖ מבוא ליישומי למידה לא מונחית: קיבוץ נתונים, זיהוי חריגים, הקטנת ממדיות
- ❖ אלגוריתם k means
- ❖ evaluation: WCSS, Silhouette

נושא 3.1 - מבוא ליישומי למידה לא מונחית

מושגים

- ★ למידה לא מונחית Unsupervised Learning
- ★ קיבוץ נתונים Clustering
- ★ **(רשות)** זיהוי חריגים Anomaly Detection

★ **(רשות)** הקטנת ממדיות Dimensionality reduction.
★ **(רשות)** הכרות עם אלגוריתמים t-SNE, UMAP, PCA

<u>מטרות ביצועיות</u>

- הסבר מהי למידה לא-מונחית וכיצד היא שונה מלמידה מונחית
- הסבר המטרות של קיבוץ נתונים, הסבר הקלט והפלט של אלגוריתם לקיבוץ נתונים
- **(רשות)** הסבר המטרות של זיהוי חריגים באמצעות למידה לא מונחית. הסבר ההבדל בין זיהוי חריגים כלמידה מונחית ולא מונחית. הסבר הקלט והפלט של אלגוריתם זיהוי חריגים. הסבר עקרון הפעולה.
- **(רשות)** הסבר מהי הקטנת ממדיות, באילו מקרים היא נדרשת, מהו הקלט והפלט של אלגוריתם הורדת מימד.
- **(רשות)** ניתוח גרף ויזואליזציה של הקטנת מימד עם UMAP.

<u>נושא 3.2 - אלגוריתם k-means</u>

אלגוריתם K-Means מתבסס על מזעור סכום ריבועי מרחקי L2. סוגי מרחק אחרים (כגון: L1, Cosine) אינם נתמכים באלגוריתם הקלאסי, אך נתמכים באלגוריתמים אחרים כגון: K-medians, K-Medoids.

★ קיבוץ נתונים - cluster
★ מרכז האשכול - centroid
★ אתחול אקראי
★ **(רשות)** אתחול בשיטת K-Means++
★ מרחק בין דוגמא למרכז
★ ממוצע של נקודות באשכול
★ התכנסות

<u>מטרות ביצועיות</u>

- הסבר מהו קיבוץ נתונים - cluster
- הסבר מהו מרכז האשכול וכיצד הוא מייצג את האשכול
- חישוב ומעקב אחר כל השלבים של אלגוריתם k-means:
  - בחירת K מספר המרכזים
  - שלב 1: אתחול אקראי של מרכזי האשכול (בחירה מתוך מאגר קיים)
  - שלב 2: שיוך הנקודות לאשכול הקרוב ביותר

$$c_i = \ arg\,min_{k\in\{1,\dots,K)}\left|\left|x_i - \mu_k\right|\right|_2$$

  - שלב 3: עדכון מרכז האשכול ע"פ ממוצע של הנקודות באשכול

$$\mu_k = \frac{1}{|C_k|}\sum_{x_i\in C_k} x_i$$

  - שלב 4: ביצוע איטרציות של שלבים 2,3 עד להגעה להתכנסות או מקסימום צעדים.
- הסבר רגישות של K-means לאקראיות של האתחול וצורך בשכלול שיטת האתחול.
- **(רשות)** חישוב ומעקב אחר אלגוריתם אתחול K-Means++ (מחליף את שלב 1)
  - שלב 1.1: בחירת מרכז (centroid) ראשון אקראי
  - שלב 1.2: חישוב מרחקים ריבועיים ממרכז הכי קרוב

$$D(x_i)^2 = min_{\mu_k \in C} ||x_i - \mu_k||^2$$

- שלב 1.3: בחירת מרכז נוסף על פי נוסחת ההסתברות

$$P(x_i) = \frac{D(x_i)^2}{\sum_{j=1}^{m} D(x_j)^2}$$

- שלב 1.4: חזרה על שלבים 1.2,1.3 עד בחירת K מרכזים

- ניתוח תוצאות האשכול ותיאור מבנה הקבוצות שנוצרו
- הסבר על חשיבות נרמול הנתונים וניתוח תוצאות האלגוריתם עם ובלי נרמול נתונים
- הסבר/תיקון/השלמת קוד k-means

נושא 3.3 - clustering evaluation

מושגים

★ WCSS - Within Cluster Sum of Squares
★ Silhouette
★ Elbow Method
★ בחירת מספר אשכולות

מטרות ביצועיות

- חישוב מדד WCSS - Within Cluster Sum of Square. זהו מדד התלכדות בתוך האשכולות.

$$WCSS = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - u_k||_2^2 \qquad WCSS = \sum_{i=1}^{m} ||x_i - u_{c_i}||_2^2$$

- הסבר כי WCSS היא פונקציית המטרה של k-means. כלומר זוהי הפונקציה שהאלגוריתם מנסה למזער. ככל ש WCSS קטן יותר כך המדד טוב יותר.

$$0 \leq WCSS$$

- חישוב, מעקב וניתוח מדד Silhouette.
  - הסבר כי זהו מדד שמשלב מדד התלכדות בתוך האשכולות, עם מדד הפרדה בין האשכולות. ככל ש S גבוה יותר כך המדד טוב יותר.
  - הסבר של משמעות ערכי המדד בטווח האפשרויות

$$-1 \leq S \leq 1$$

  - **(רשות)** חישוב מרחק ממוצע בתוך ה cluster

$$a(i) = \frac{1}{|C_k|-1} \sum_{x_j \in c_k, j \neq i} d(x_i, x_j)$$

  - **(רשות)** חישוב מרחק ממוצע ל cluster הקרוב ביותר.
    פונקציית המרחק יכולה להיות: d_1, d_2, d_{cos}

$$d_{i,m} = \frac{1}{|C_m|} \sum_{x_j \in c_m} d(x_i, x_j)$$

$$b(i) = min_{m \neq k} d_{i,m}$$

- **(רשות)** חישוב המדד

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

- ניתוח איכות תוצאות האשכול על סמך המדדים
- בחירת מספר האשכולות לפי שיטת המרפק (Elbox Method)
- בחירת מספר אשכולות מתאים תוך שילוב שיטות Elbow ומדד Silhouette.
- הסבר/תיקון/השלמת קוד המחשב את מדד WCSS
- **(רשות)** הסבר/תיקון/השלמת קוד המחשב את מדד Silhouette

פרק 4 - מודלים מבוססי מרחק

מבוא

למידה מונחית כוללת אלגוריתמים לזיהוי דפוסים בתוך נתונים מתויגים על מנת לבצע משימות של חיזוי התגית מתוך הנתונים. אלגוריתמים אלו שימושיים מאוד לזיהוי מסמכים רלוונטיים בתוך מאגר מסמכים, זיהוי מחלות מתוך בדיקות מעבדה, לזיהוי מגמות של עליה וירידה במדדים כלכליים ועוד.

בפרק זה נלמד אלגוריתמים מבוססי מרחק. אלגוריתמים אלו אינם מורכבים מתמטית ומגלים ביצועי סיווג טובים. אלגוריתמים אלו מהווים לכן בסיס טוב לצורך הבנת הקונספטים הבסיסיים בלמידת מכונה, לפני שמתקדמים לאלגוריתמים מודרניים מבוססי רשתות נוירונים.

מודלים מבוססי מרחק הם שיטות פשוטות ואינטואיטיביות בלמידת מכונה, שבהן ההחלטה מתקבלת לפי **קרבה גאומטרית** בין דוגמאות במרחב התכונות. כל דוגמה מיוצגת כווקטור, והדמיון או המרחק בין וקטורים נמדד באמצעות מדדים כמו מרחק אוקלידי (L2), מרחק מנהטן (L1) או מרחק קוסינוס (cos).

Nearest Centroid

ב-Nearest Centroid כל מחלקה מיוצגת על ידי **המרכז (הממוצע)** של הדוגמאות שלה. דוגמה חדשה מסווגת למחלקה שהמרכז שלה הוא הקרוב ביותר. המודל פשוט ומהיר מאוד, אך מניח שכל מחלקה ניתנת לייצוג טוב על ידי נקודה אחת.

K-Nearest Neighbors

ב-KNN, כדי לסווג דוגמה חדשה, מחשבים את המרחק שלה מכל דוגמאות האימון, בוחרים את **K השכנים הקרובים ביותר**, והסיווג נקבע לפי **הצבעת רוב** (או ממוצע במקרה של רגרסיה). זהו מודל לא-פרמטרי, שאינו מבצע אימון מפורש אך דורש חישוב רב בזמן חיזוי.

מאפיינים כלליים

שני המודלים רגישים לסקייל של התכונות ולכן נרמול הנתונים חשוב. הם מתאימים במיוחד כמודלי בסיס (baseline) וככלי לימודי להבנת ייצוג גאומטרי של נתונים.

רשימת הנושאים

- ❖ מבוא ללמידה מונחית
- ❖ אלגוריתם Nearest Centroid
- ❖ אלגוריתם KNN

נושא 4.1 - מבוא ללמידה מונחית

בנושא זה בעיקר מבצעים רענון וחזרה על מושגים שנלמדו במסגרת יחידה 3.

מושגים

- ★ למידה מונחית
- ★ קלסיפיקציה ורגרסיה
- ★ טבלת נתונים
- ★ דוגמאות
- ★ מאפיינים
- ★ תגיות
- ★ שלב הכנת הנתונים
- ★ שלב אימון המודל model.fit()
- ★ שלב פרדיקציה/חיזוי model.predict()
- ★ נתוני אימון ומבחן
- ★ נרמול נתונים על אימון ומבחן
- ★ אימון הנרמול scaler.fit()
- ★ טרנספורמציה של הנרמול: scaler.transform()
- ★ מדד accuracy

מטרות ביצועיות

- הסבר מהי למידה מונחית ומהם היישומים המרכזיים של פרדיגמה זאת: קלסיפיקציה, רגרסיה
- הדגמת יישומים של קלסיפיקציה ושל רגרסיה
- הסבר מהי טבלת נתונים כולל: דוגמאות,מאפיינים ותגיות
- הסבר תהליך למידת מכונה ובכלל זה:
  - שלב הכנת הנתונים
  - שלב אימון המודל
  - שלב פרדיקציה/חיזוי
- חלוקה לסט אימון ומבחן והסבר תפקידם של נתוני אימון ומבחן.
- חישוב של תקנון ונרמול כחלק מתהליך האימון:
  - את פרמטרי הנרמול מחשבים רק על סט האימון, ומיישמים אותם גם על סט המבחן.
  - שלב האימון של הנרמול: scaler.fit()
  - שלב הטרנספורמציה של הנרמול: scaler.transform()
  - הסבר מימוש הנרמול על נתוני האימון והמבחן בכל אחת משיטות הנרמול והתקנון שנלמדו
- חישוב מדד accuracy
- הסבר/תיקון/השלמת קוד המבצע את תהליך הלמידה של מכונה לומדת.

נושא 4.2 - אלגוריתם Nearest Centroid

אלגוריתם **Nearest Centroid** תומך במרחקי **L2** ו־**L1**, שכן עבורם ניתן לחשב את הצנטרואיד בצורה סגורה (ממוצע וחציון, בהתאמה). מרחק קוסינוס אינו נתמך באלגוריתם הקלאסי, מאחר שאין עבורו צנטרואיד בעל צורה סגורה — כלומר, אין נוסחה המאפשרת לחשב את הפתרון ישירות, ללא חזרות או תהליך איטרטיבי.

מושגים

★ מרכזים (centroids)
★ מרכז הכי קרוב
★ היפר פרמטר metric

מטרות ביצועיות

- הסבר היפר-פרמטרים של האלגוריתם ובחירת היפר-פרמטרים בהתאם לבעיה:
  - metric - סוג המרחק L1/L2
- מעקב וחישוב שלב האימון (fit) של המודל
  - חישוב המרכזים (centroids) לכל המחלקות
  -

$$C_k = \{x_i \mid y_i = k\}$$

  - במדד L2 נרצה למזער את סכום ריבועי מרחקי L2 בין המרכז לדוגמאות באותה מחלקה. זה ייעשה באמצעות חישוב הממוצע של המאפיינים:

$$L2\ (Euclidean)\colon\ \mu_k = \frac{1}{|C_k|}\sum_{x_i \in C_k} x_i$$

  - במדד L1 נרצה למזער את סכום מרחקי L1 בין המרכז לדוגמאות באותה מחלקה. זה ייעשה באמצעות חישוב חציון של המאפיינים:

$$L1\ (Manhattan)\colon\ (\mu_k)_j = median\{x_{ij} \mid x_i \in C_k\}$$

- מעקב וחישוב שלב החיזוי (predict) של המודל
  - בעבור קלט v, נמצא את התגית של המרכז (centroid) הכי קרוב:

$$\hat{y} = arg\,min_{k\in\{1,\dots,C\}} \left\|v - \mu_k\right\|$$

- מעקב וחישוב החיזוי עבור ריבוי דוגמאות
  - נתון אוסף של t דוגמאות

$$V = [v_1, v_2, \dots, v_t]$$

  - עבור כל קלט v_i נבצע את שלב החיזוי ונקבל $\hat{y}_i$
  - נחזיר את וקטור החיזויים:

$$\hat{y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_t]$$

- הסבר על חשיבות נרמול הנתונים וניתוח תוצאות האלגוריתם עם ובלי נרמול נתונים
- הסבר/תיקון/השלמת קוד המבצע K-means.

נושא 4.3 - אלגוריתם KNN

אלגוריתם **KNN** תומך במגוון רחב של מטריקות (Cosine, L1, L2 ועוד), משום שהוא אינו מניח מבנה מיוחד למרחק.

מושגים

★ שכן קרוב / שכנים קרובים
★ מרחקים של וקטור מדוגמאות האימון

★ היפר פרמטר k - מספר השכנים
★ היפר פרמטר metric
★ היפר פרמטר weights - שקלול אחיד / מבוסס מרחק.
★ התגית השכיחה
★ לומד עצל (Lazy Learner)

<u>מטרות ביצועיות</u>

- הסבר היפר-פרמטרים של האלגוריתם ובחירת היפר-פרמטרים בהתאם לבעיה:
  - k - מספר השכנים
  - metric - סוג המרחק L1/L2/Cosine
  - weights - שקלול אחיד / מבוסס מרחק. פרמטר זה קובע האם לכל שכן יש משקל שווה בהצבעה, או שההצבעה תהיה משוקללת על פי מרחק, כאשר שכנים קרובים יקבלו משקל גבוה יותר.
- חישוב ומעקב אחר שלב אימון המודל
  - שמירת דוגמאות האימון X
- חישוב ומעקב אחר שלב חיזוי המודל
  - בעבור וקטור לחיזוי v
    - חישוב המרחקים של וקטור v מדוגמאות האימון X.

$$d_i(v) = \left\| v - x_i \right\| \qquad i = 1,\dots, m$$

- מציאת k השכנים הקרובים ביותר

$$NN_k(v) = \textit{the k training samples with the smallest distances to v}$$

- חישוב התגית השכיחה כאשר weights = uniform

$$S_c = \sum_{x_i \in NN_k(v),\, y_i = c} 1$$

$$\hat{y}(v) = arg\,max_c\, S_c$$

- חישוב התגית השכיחה כאשר weights = distance

$$w_i = \frac{1}{d_i(v)}$$

$$S_c = \sum_{x_i \in NN_k(v),\, y_i = c} w_i$$

$$\hat{y}(v) = arg\,max_c\, S_c$$

- חישוב ומעקב החיזוי עבור ריבוי דוגמאות
  - נתון אוסף של t דוגמאות

$$V = [v_1, v_2, \dots, v_t]$$

  - עבור כל קלט v_i נבצע את שלב החיזוי ונקבל $\hat{y}_i$
  - נחזיר את וקטור החיזויים:

$$\hat{y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_t]$$

- **(רשות)** הסבר שיטות למציאת k השכנים הקרובים ביותר
  - באמצעות מיון הדוגמאות על פי מרחק, ובחירת k הדוגמאות הקרובות ביותר. סיבוכיות: $O(m * log\, m)$
  - באמצעות מציאת דוגמה עם מינימום מרחק + הסרה מרשימה, k פעמים. סיבוכיות: $O(m * k)$
- **(רשות)** השוואת האלגוריתמים Nearest Centroid ו KNN מבחינת סיבוכיות זמן האימון וסיבוכיות זמן החיזוי.
- הסבר על חשיבות נרמול הנתונים וניתוח תוצאות האלגוריתם עם ובלי נרמול נתונים
- הסבר/תיקון/השלמת קוד המבצע KNN.

<u>נושא 4.4 - אלגוריתמי סיווג מבוסס מרחק - מבט על</u>

<u>מושגים</u>

- ★ למידה
- ★ מודל
- ★ מורכבות המודל
- ★ הכללה
- ★ משטחי הפרדה / גבולות הפרדה
- ★ הפרדה לינארית
- ★ התאמת יתר וחסר
- ★ פער ההכללה Generalization Gap
- ★ הפער האנושי Human Gap
- ★ כיוון היפר-פרמטרים

<u>מטרות ביצועיות</u>

- הסבר כי למידה היא תהליך שבו האלגוריתם מזהה חוקיות מתוך נתוני האימון.
- הסבר כי הכללה היא היכולת של המודל ליישם חוקיות זו על נתונים חדשים שלא נראו באימון.
  - אלגוריתם nearest centroid מבצע הכללה על ידי מציאת המרכז של כל מחלקה.
  - אלגוריתם KNN מבצע חיזוי על סמך דוגמאות דומות, והכללה מתקבלת מתוך מבנה הנתונים עצמו ולא מתוך בניית מודל פרמטרי.
- הסבר כי מודל מייצג את הנתונים. אלגוריתם הלמידה בונה את המודל ואלגוריתם הסיווג משתמש במודל.
- הסבר שלמידה היא הכללה של הדוגמאות לצורך בניית המודל.
- הסבר כי משטחי הפרדה מאפשרים לראות כיצד המודל מחלק את מרחב המאפיינים לתתי מרחב בהתאם לחיזוי מחלקות הנתונים.
- הסבר כי הפרדה לינארית היא חלוקה של מרחב המאפיינים לתתי מרחב על ידי קווים (2 ממדים), מישורים (3 ממדים) או היפר-מישורים (מעל 3 ממדים). הפרדה לינארית היא מצב שבו גבולות ההחלטה (Decision Boundaries) הם קווים, מישורים או היפר־מישורים.
  - אלגוריתם nearest centroid מאופיין בהפרדה לינארית.
- הסבר כי התאמת חסר היא מצב בו המודל לא מספיק משוכלל על מנת לייצג את המורכבות של הנתונים.
  - אלגוריתם nearest centroid נוטה להתאמת חסר כי המודל מאוד פשוט. נגיע למצב זה אם לא ניתן להפריד את הנתונים עם הפרדה לינארית.

- אלגוריתם KNN נוטה להתאמת חסר כאשר K גדול מדי. במקרה זה שכנים רחוקים משפיעים על ההחלטה ולכן משטחי ההפרדה הופכים חלקים מדי ואינם מתאימים למבנה האמיתי.
- הסבר כי התאמת יתר היא מצב בו משטחי ההפרדה יכולים להיות מפותלים מאוד ואז אין הכללה טובה
  - אלגוריתם KNN נוטה להתאמת יתר כאשר K נמוך מכיוון שבוחנים את המחלקות של מספר מועט של שכנים ולכן יכולה להיות השפעה גדולה של רעש.
  - אלגוריתם KNN מאופיין במשטחי הפרדה מפותלים
- זיהוי מצבים של התאמת יתר או התאמת חסר על סמך ביצועי אימון ומבחן מטבלה או גרף.
- בחירת היפר פרמטריים לאלגוריתם למידה מונחית על פי אזורי התאמת יתר וחסר.
- חישוב פער הכללה Generalization Gap והסבר של משמעותו.
- חישוב הפער האנושי Human Gap והסבר של משמעותו.
- הסבר כי אחת המטרות של כיוון היפר-פרמטרים של המודל היא מציאת נקודת עבודה של הכללה טובה שאינה באזורים של התאמת חסר או התאמת יתר. מטרת בחירת K ופרמטרי המרחק היא לא למקסם דיוק על נתוני האימון, אלא למקסם הכללה על נתונים חדשים.

פרק 5 - מושגי יסוד בפיתוח אלגוריתמי למידת מכונה

מבוא

תהליך האימון של מכונה לומדת הוא מורכב וכולל בחירה של אלגוריתם, בחירה של היפר-פרמטרים, אימון וכיוון הפרמטרים והערכת ביצועים. ללא תהליך מתודולוגי של אימון מכונה לומדת לא ניתן להגיע לביצועים טובים. לכן פרק זה מעמיק למספר נושאים תיאורטיים הנדרשים להבנת תהליך הפיתוח של מכונה לומדת.
בפרק זה נעסוק בהערכת מודלים בלמידת מכונה ובתהליכים הנדרשים לבחירה של מודל חיזוי איכותי ואמין. נלמד כיצד להעריך את ביצועי המודל באמצעות מטריצת הבלבול ומדדי ביצוע מרכזיים כגון Accuracy, Precision, Recall ו־F1, וכיצד מדדים אלו מסייעים להבין את חוזקותיו וחולשותיו של המודל.
לבסוף, נעסוק בתהליכי ולידציה ונכיר שיטות לחלוקת הנתונים לסטי אימון, ולידציה ומבחן, כולל ולידציה מוצלבת. את העקרונות שנלמד ניישם בתהליך פיתוח מלא של מודל KNN, שבו נבחר ערך (K) מתאים על סמך ביצועי ולידציה ונבחן את המודל הסופי על סט המבחן.

רקע נדרש

רקע תיאורטי ומעשי שנלמד במסגרת היחידה השלישית מדעי הנתונים. ביחידה זו נבצע חזרה והעמקה בתחומים הרלוונטיים.

רשימת הנושאים

- ❖ הערכת ביצועים
  - ◇ מטריצת הבלבול
  - ◇ מדדי ביצוע accuracy, precision, recall, F1
- ❖ תהליכי ולידציה
  - ◇ חלוקה לאימון, ולידציה, מבחן
    - ■ נתוני אימון - משמשים לאימון הפרמטרים של המודל
    - ■ validation - בדיקת ביצועים של המודל וכיוון היפר פרמטרים
    - ■ נתוני מבחן - בחינה של המודל הסופי
    - ■ ולידציה מוצלבת (Cross validation)

- ◇ תיאור תהליך פיתוח אלגוריתם למידת מכונה
  - ■ אימון פרמטרים
  - ■ ולידציה וקביעת היפר פרמטרים
  - ■ מבחן

נושא 5.1 - הערכת ביצועים

מבוא

אפליקציות שונות דורשות אופטימיזציה על מדדי ביצוע שונים. לדוגמא במשימת חיפוש באינטרנט חשוב לנו דיוק התוצאות - כלומר שיעור התוצאות שעונות על שאלת החיפוש מתוך כלל התוצאות. במשימה של חיפוש חפצים מסוכנים בסריקות אבטחה של נמל תעופה חשוב לנו האיחזור, כלומר שיעור החפצים המסוכנים שהמכונה גלתה מכלל החפצים. פרק זה מעמיק את ההבנה של מדדי ביצוע הנלמדים ביחידה 3 ומציג את מדדי הביצוע השונים ואת האפליקציות המתאימות למדדים השונים.

מושגים

- ★ מטריצת הבלבול
- ★ ספירות מטריצת הבלבול לתגית: TP, FN, FP, TN
- ★ מדד גלובלי accuracy
- ★ מדדי ביצוע לתגית precision, recall, f1
- ★ מדדי ביצוע גלובליים: ממוצע חשבוני ומשוקלל precision, recall, f1
- ★ סף החלטה בינארי Binary Decision Threshold
- ★ Precision-Recall Tradeoff

מטרות ביצועיות

- חישוב וניתוח של מטריצת הבלבול.
  - בשורות יופיע תגיות האמת. בעמודות יופיעו התחזיות של המודל.

$$y = (y_1,\dots, y_m) \quad t \in \{1,\dots, m\} \quad y_t \in \{0,\dots, K-1\}$$

$$\hat{y} = (\hat{y}_1,\dots, \hat{y}_m) \quad t \in \{1,\dots, m\} \quad \hat{y}_t \in \{0,\dots, K-1\}$$

$$C \in N^{KxK} \qquad C_{ij} = \sum_{t=1}^{m} (y_t = i \; and \; \hat{y}_t = j)$$

- חישוב ספירות מטריצת הבלבול לתגית

$$TP_k = \sum_{t=1}^{m} (y_t = k \; and \; \hat{y}_t = k)$$

$$FN_k = \sum_{t=1}^{m} (y_t = k \; and \; \hat{y}_t \neq k)$$

$$FP_k = \sum_{t=1}^{m} (y_t \neq k \; and \; \hat{y}_t = k)$$

$$TN_k = \sum_{t=1}^{m} (y_t \neq k \; and \; \hat{y}_t \neq k)$$

- חישוב מדד גלובלי Accuracy

$$Accuracy = \frac{1}{m}\sum_{t=1}^{m}(y_t = \hat{y}_t)$$

- חישוב מדדי Precision, Recall, F1 לכל תגית

$$Precision_k = \frac{TP_k}{TP_k+FP_k} \qquad Recall_k = \frac{TP_k}{TP_k+FN_k} \qquad F1_k = \frac{2\cdot Precision_k \cdot Recall_k}{Precision_k+Recall_k}$$

- הסבר כי F1 הוא ממוצע הרמוני של מדדי Precision, Recall מכיוון שהוא מאזן בין שני המדדים ומחמיר יותר מאשר ממוצע חשבוני רגיל.

- חישוב גלובלי של המדדים Precision, Recall, F1 לפי ממוצע רגיל על המחלקות (מאקרו)

$$Precision_{macro} = \frac{1}{K}\sum_{k=1}^{K} Precision_k \qquad Recall_{macro} = \frac{1}{K}\sum_{k=1}^{K} Recall_k \qquad F1_{macro} = \frac{1}{K}\sum_{k=1}^{K} F1_k$$

- חישוב גלובלי של המדדים Precision, Recall, F1 לפי ממוצע משוקלל של כמות הדוגמאות בכל מחלקה (weighted)

$$n_k = support\ of\ class\ k \qquad N = \sum_{k=1}^{K} n_k$$

$$Precision_{weighted} = \frac{1}{N}\sum_{k=1}^{K} n_k \cdot Precision_k \qquad Recall_{weighted} = \frac{1}{N}\sum_{k=1}^{K} n_k \cdot Recall_k$$

$$F1_{weighted} = \frac{1}{N}\sum_{k=1}^{K} n_k \cdot F1_k$$

- הבנת ההבדלים בין חישוב גלובלי של המדדים Precision, Recall, F1 לפי ממוצע רגיל על המחלקות (מאקרו) ולפי לפי ממוצע משוקלל של כמות הדוגמאות בכל מחלקה (weighted).
- בחירת מדד ביצוע מועדף בהתאם למאפייני הבעיה (לדוגמא - מהו המדד המתאים לזיהוי מחלות)
- הסבר מהו סף החלטה בינארי Binary Decision Threshold
- חישוב סף החלטה בינארי באלגוריתם KNN עם משקולות אחודה ומשקולת עפ"י מרחק.
- הסבר ההשפעה של סף החלטה בינארי על המדדים השונים.
- ניתוח גרף Precision Recall Tradeoff ובחירת סף לבעיה נתונה לפי דרישות precision או recall.
- הסבר/תיקון/השלמת קוד המחשב מדדי ביצוע של מסווג.

נושא 5.2 - ולידציה וחקר פרמטרים

מבוא

לצורך אימון של מודל למידת מכונה טוב נדרשת מתודולוגיה נכונה לתהליך האימון של המכונה. ללא מתודולוגיה נכונה אפשר להיקלע למצבים של למידה לא מוצלחת ואף להגיע למצב בו לא ניתן לזהות שהביצועים של המודל לא טובים. בפרק זה נלמדת ומיושמת המתודולוגיה הנדרשת לאימון מכונה לומדת.

מושגים

★ נתוני אימון, ולידציה ומבחן
★ ולידציה מוצלבת - Cross Validation

★ אובייקט Pipeline - מאגד תהליכים
★ מילון Param_grid - פרמטרים וערכים לסריקה
★ אובייקט GridSearchCV - תשתית לחקר פרמטרים.

<u>מטרות ביצועיות</u>

- הסבר התהליך נכון של חלוקה לאימון, ולידציה ומבחן
  - הסבר מטרת תהליכי הוולידציה ואת תפקידם בהערכת מודלים.
  - הסבר הצורך בחלוקת הנתונים לסט אימון, ולידציה ומבחן.
  - שימוש בנתוני אימון ולידציה לצורך אימון מודל וכיוון היפר־פרמטרים.
  - שימוש בנתוני מבחן לצורך בחינה סופית והוגנת של המודל.
  - הסבר עקרון הוולידציה המוצלבת (Cross Validation) ומתי נכון להשתמש בה.

להלן דוגמה עם מודל KNN:

  - חלוקת הנתונים לסט אימון, ולידציה ומבחן
  - העברת סט האימון לאלגוריתם - בניית רשימת השכנים.
  - היפר פרמטרים של האלגוריתם הם: מספר השכנים, סוג המרחק, שקלול ה weights, סוג התקנון/נרמול.
  - בחנו את הביצועים של סט הולידציה עבור כל סט היפר פרמטרים.
  - בחירת סט היפר פרמטרים בעל הביצועים הטובים ביותר.
  - בדיקת המודל הנבחר על סט המבחן
- **(רשות)** הסבר/תיקון/השלמת קוד של אובייקט Pipeline - מאגד תהליכים
- **(רשות)** הסבר/תיקון/השלמת קוד של מילון Param_grid - פרמטרים וערכים לסריקה
- **(רשות)** הסבר/תיקון/השלמת קוד של אובייקט GridSearchCV - תשתית לחקר פרמטרים
- בחירת היפר-פרמטרים מתוך הפלט של GridSearchCV
- ניתוח והסקת מסקנות מתוך פלט של GridSearchCV

שאלון בגרות לדוגמא (1) - מדעי הבינה המלאכותית

שאלון בגרות לדוגמא - מדעי הבינה המלאכותית

בשאלון זה 3 שאלות. יש לענות על 2 שאלות מתוך 3. 50 נקודות לכל שאלה.

שאלה 1

מאיה עובדת על פרויקט לזיהוי מחלות לב על בסיס בדיקות מעבדה. במאגר הנתונים קיימים הנתונים הבאים: רמת כולסטרול בדם, רמת סוכר בדם והאם קיימת מחלת לב. להלן 6 הרשומות הראשונות בטבלת הנתונים:

מטופל	רמת כולסטרול (mg/dL)	רמת סוכר בדם (mg/dL)	מחלת לב
1	180	90	לא
2	220	105	כן
3	195	100	לא
4	250	130	כן
5	210	115	כן
6	170	85	לא

1. חשבו את הממוצע ואת השונות של רמת הכולסטרול על בסיס הנתונים בטבלה. (הערה: יש לחשב את השונות כאילו מדובר באוכלוסייה, אין צורך בתיקון למדגם)

2. עידית חיפשה עבור מיה פונקציות ספרייה שיוכלו לבצע את החישובים הנדרשים ומצאה את הפונקציות הבאות. מה מבצעת כל אחת מהפונקציות sod_1,sod_2?

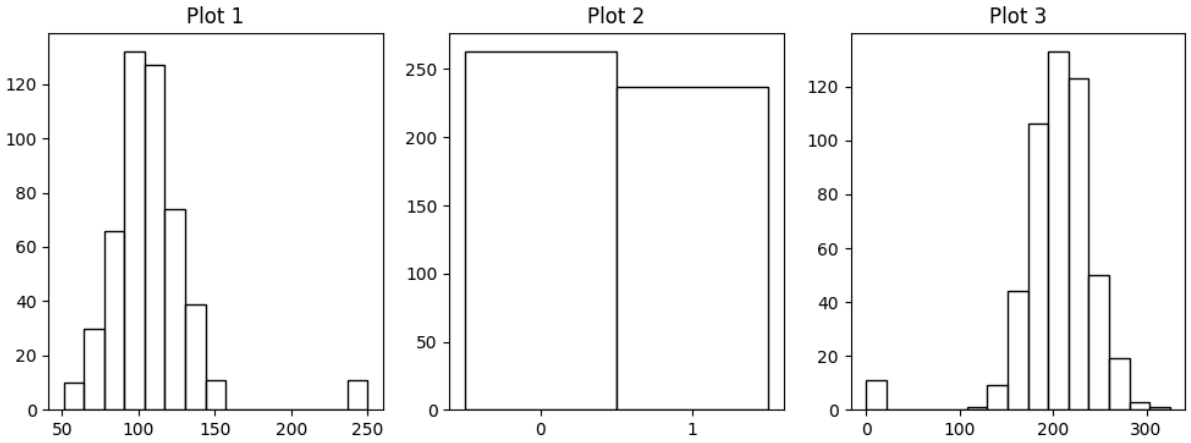
```
def sod_1(X):
  a = 0
  for val in X:
    a += val
  return a / len(X)

def sod_2(X):
  a = sod_1(X)
  b = 0
  for val in X:
    b += (val - a) ** 2
  return (b / len(X)) ** 0.5
```

3. כתבו פונקציה המקבלת רשימה של ערכים X וערך בודד y ומחזירה את ציון התקן Z של הערך y. ניתן להשתמש בפונקציות הספרייה sod_1,sod_2 לטובת החישוב.

```
def z_score(X,y):
```

```
return _____
```

4. ניר עוזר למאיה ושרטט היסטוגרמות של המשתנים במאגר (גרפים 1-3) אולם שכח להוסיף כותרות לגרפים. קבעו איזה גרף שייך לאיזה מאפיין. נמקו את קביעתכם.



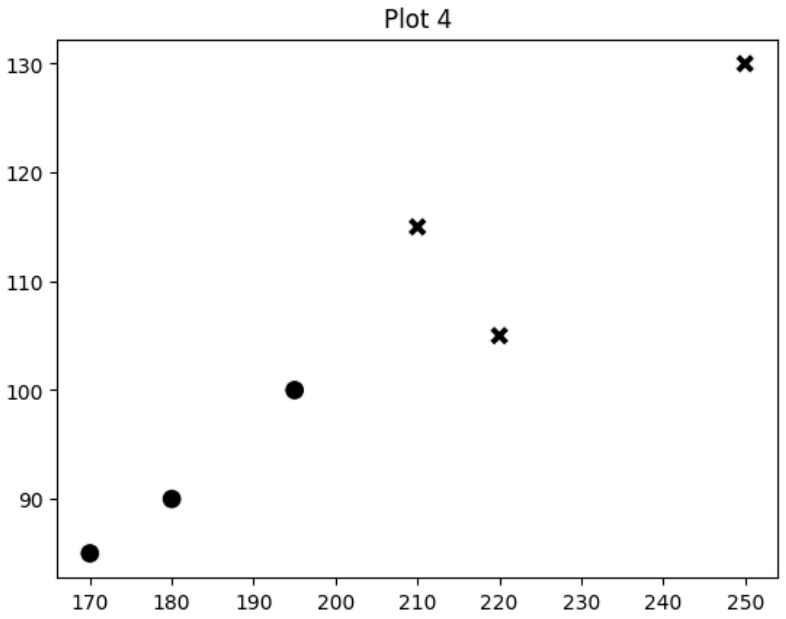
.a

5. על בסיס הגרפים בסעיף 4, העריכו כמה מטופלים כלולים במאגר.

6. על בסיס הגרפים בסעיף 4, העריכו האם השונות של רמת הסוכר נמוכה, שווה או גבוהה מהשונות של רמת הכולסטרול. נמקו את קביעתכם.

7. חשבו את ערך Z של נתון קיצוני עבור המשתנה רמת כולסטרול, קבעו על בסיס ערך Z זה האם הוא חריג.

8. מאיה שרטטה גרף פיזור של רמת הסוכר בדם לעומת רמת הכולסטרול (תרשים 4) אולם שכחה להוסיף כותרות לצירים.



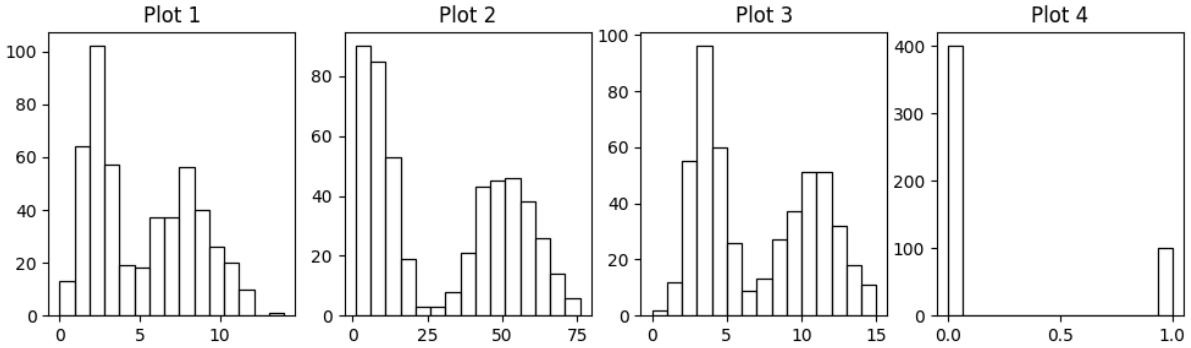
a. קבעו איזה מאפיין נמצא על ציר x. נמקו את קביעתכם.
b. הסבירו מדוע הגרף כולל שתי צורות סימון.
c. האם הגרף כולל את כלל הדוגמאות במאגר? נמקו את קביעתכם.
d. האם על פי הגרף יש קשר בין רמת כולסטרול לבין רמת סוכר בדם. נמקו את קביעתכם.

9. אורן טוען שלטובת משימת סיווג באמצעות אלגוריתם KNN יש צורך לבצע המרת תחום לנתונים. האם אורן צודק? נמקו את קביעתכם.

שאלה 2

אריאל עובד על מערכת לזיהוי תקלות תוכנה על בסיס שלושה מאפיינים של קוד: מספר הלולאות, מספר המשתנים, ומספר התנאים.

1. אריאל שרטט היסטוגרמה של המאפיינים במאגר הנתונים (גרפים 4-1). אילו מבין הגרפים מייצג את המשתנה "תקלה"? נמקו את קביעתכם.



2. על בסיס הגרפים בסעיף 1, קבעו האם קובץ הנתונים מאוזן? נמקו את קביעתכם.

3. על סביב הגרפים בסעיף 1, קבעו כמה דוגמאות יש במאגר.

4. קובי מנסה לסווג את הפרויקט שלו (בעל 4 לולאות, 10 משתנים ו-6 תנאים) באמצעות אלגוריתם KNN שאומן על מאגר הנתונים שאריאל אסף. קובי התחיל לחשב את המרחקים בין הפרויקט שלו לבין הדוגמאות במאגר, אולם הלולאה נעצרה אחרי 5 חישובים בלבד.

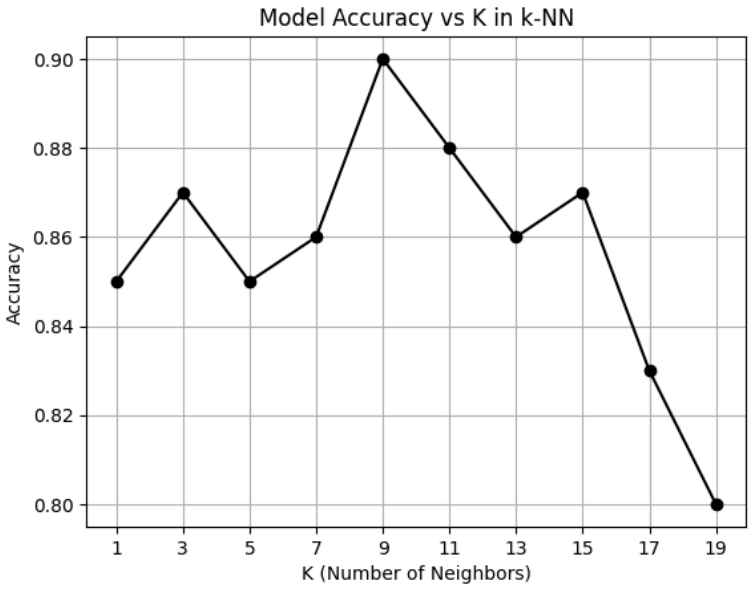
פרויקט	מספר לולאות	מספר משתנים	מספר תנאים	תקלה	מרחק (4 לולאות, 10 משתנים, 6 תנאים)
חדש	4	10	6	?	
1	2	8	3	לא	4.12
2	5	15	6	כן	5.10
3	3	10	4	לא	2.24
4	7	18	8	כן	8.77
5	4	12	5	כן	2.24
6	1	6	2	לא	

a. קבעו על פי חמשת הדוגמאות הראשונות באיזה מרחק משתמש קובי: L1, L2 או קוסינוס?
b. חשבו את המרחק של הפרויקט של קובי מדוגמא 6.
c. מהו הסיווג של הפרויקט החדש עבור k=3. נמקו את קביעתכם.

5. קובי מצא שתי פונקציות ספרייה לחישוב מרחקים: sod_x, sod_y. מה מחשבות פונקציות ספרייה אלו?

```
def sod_x(a,b):
  d = 0
  for i in range(len(a)):
      d += abs(a[i]-b[i])
  return d

def sod_y(a,b):
  for i in range(len(a)):
    d += (a[i]-b[i])**2
  return np.sqrt(d)
```

6. נתון גרף של דיוק המסווג על נתוני הולידציה (validation set) עבור K שונים.



a. מהו ה-K האופטימלי על פי גרף זה? נמקו את קביעתכם.
b. מדוע ביצועי המסווג נמוכים עבור K קטן מערכי K אופטימלי?
c. מדוע ביצועי המסווג נמוכים עבור K גדול מערכי K אופטימלי?

7. נתונה טבלת confusion matrix של ביצועי המסווג.

	נחזתה תקלה	**לא נחזתה תקלה**
יש תקלה (תגית 1)	7	5
אין תקלה (תגית 0)	3	5

   a. על כמה פרויקטי תוכנה נבדק המסווג?
   b. חשבו accuracy של המסווג.
   c. חשבו precision ו- recall לכל תגית.
   d. נציג מחלקת בקרת איכות טוען כי בפרויקטים רבים יש תקלה והיא לא מזוהה. לאיזה מדד ותגית מתייחס נציג מחלקת בקרת איכות: accuracy, precision או recall? נמקו את קביעתכן.
   e. נציג מחלקת פיתוח טוען כי בפרויקטים רבים מזוהה תקלה גם כשהם תקינים. לאיזה מדד ותגית מתייחס נציג מחלקת בקרת איכות: accuracy, precision או recall? נמקו את קביעתכן.

8. עמוס בנה ממסווג nearest centroid על בסיס הנתונים של אריאל וגילה כי זמן האימון וזמן החיזוי שונים לעומת זמן האימון והחיזוי של אלגוריתם KNN
   a. הסבירו את ההבדלים בזמני האימון והסיווג בין האלגוריתמים.
   b. נתונה טבלה עם זמני הריצה לאימון ולחיזוי של האלגוריתים שאמנו אריאל (KNN) ועמוס (nearest centroid). זהו איזו עמודה שייכת לאיזה אלגוריתם

	אלגוריתם 1	**אלגוריתם 2**
זמן אימון (ננו שניות)	5	445
זמן חיזוי (ננו שניות)	853	50

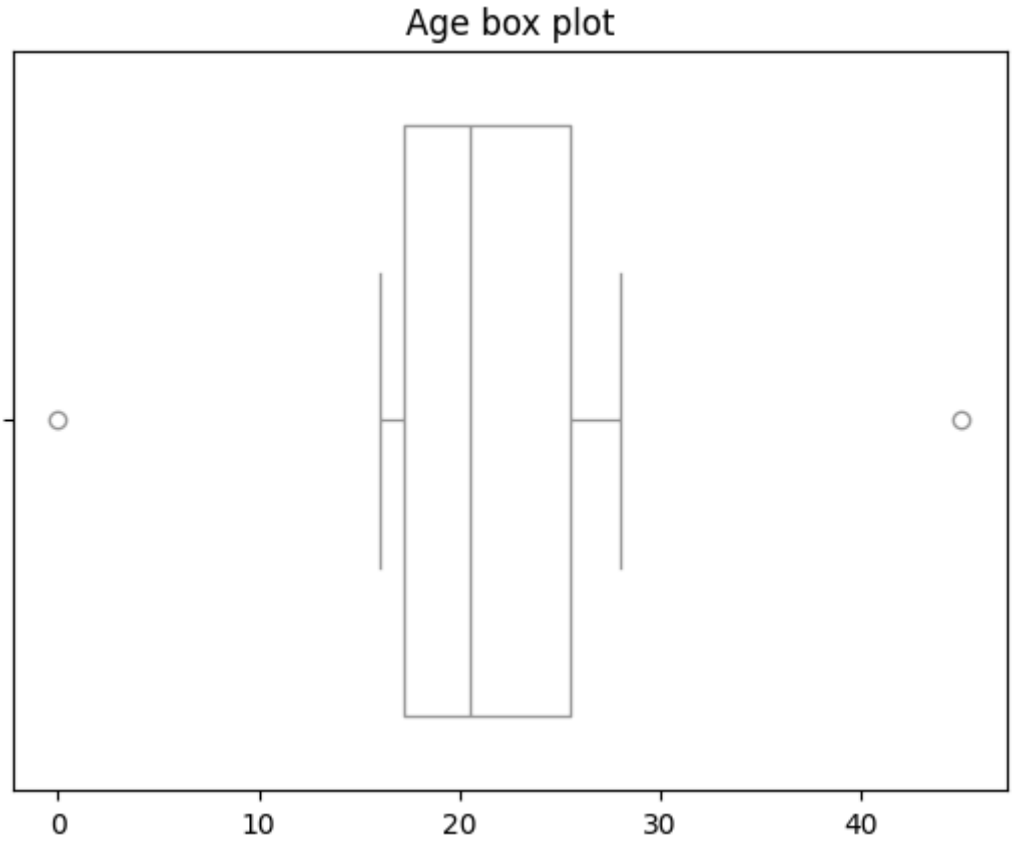
שאלה 3

רינת חוקרת סרטים בנושא הוראת בינה מלאכותית. לצורך כך היא אספה נתונים על צפיה בסרטוני YouTube.

1. שניים מהמאפיינים הם גיל והשכלה. להלן טבלת נתוני גיל והשכלה של 10 רשומות מתוך המאגר. חשבו מהי ההשכלה השכיחה בנתונים?

רשומה	גיל	השכלה
1	16	תיכון
2	18	תיכון
3	22	תואר ראשון
4	0	תואר ראשון
5	24	תואר ראשון
6	26	תואר שני
7	17	תיכון
8	19	תיכון
9	28	תואר שני
10	45	תואר שני

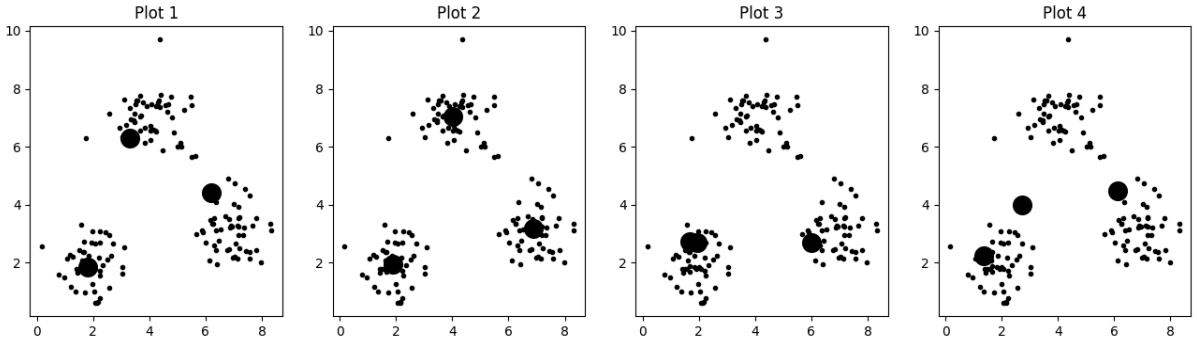
2. רינת הכינה box plot של המשתנה גיל, האם נראים בו נתונים חריגים? אם כן - פרטו אילו נתונים חריגים בגרף.



3. רינת טוענת כי רשומה 4 היא חריגה ללא קשר לסעיף הקודם. הסבירו את טענתה של רינת?

4. חנה מציעה להמיר את המשנה השכלה למשתנה כמותי עם הערכים תיכון = 1, תואר ראשון = 2, תואר שני = 3 האם הצעתה של חנה מתקבלת על הדעת? נמקו את קביעתכם

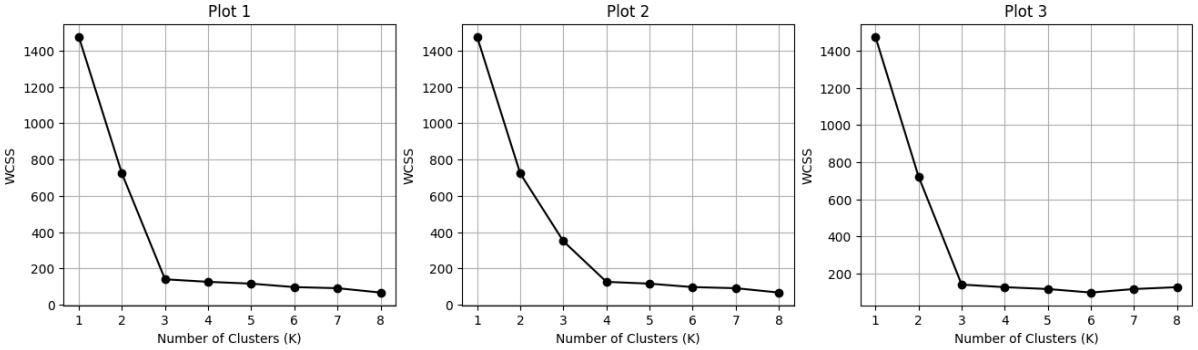
5. רינת מאמנת אלגוריתם k means על מנת לזהות אשכולות בנתונים. כדי לעקוב אחרי תהליך האימון רינת הדפיסה תרשים פיזור של הנתונים עם סימון ⭕ המייצג את מרכזי האשכולות, אולם סדר הגרפים השתבש. קבעו מהו הסדר הנכון של התרשימים בתהליך האימון של האשכולות. נמקו את קביעתכם.



6. רינת כתבה את הקוד הבא כחלק מהמחקר. הפונקציה מקבלת את בסיס הנתונים במשתנה X ואת רשימת המרכזים של k-means במשתנה C. מה מבצע הקוד ולמה רינת יכולה להשתמש בו?

```
def sod(X, C):
    t = 0
    for point in X:
        a = []
        for center in C:
            a.append(np.linalg.norm(point - center) ** 2)
        t += np.min(a)
    return t
```

7. רינת הריצה את אלגוריתם k-means עבור 8 ערכי k שונים ויצרה גרפים של WCSS ביחס ל-k. היא הכינה 3 גרפים אולם 2 נראים לה שגויים. איזה גרף מהתרשימים הבאים הוא הגרף הנכון? נמקו את קביעתכם.



8. חנה מאמנת אלגוריתם k-means על נתוני אורך הסרטים (בשניות) ומספר הלייקים שקיבל כל סרטון. בטבלה להלן מופיעות 4 הדוגמאות ראשונות מתוך הנתונים. חנה מחליטה לאתחל את דוגמאות 1,2 למרכז A ואת דוגמאות 3,4 למרכז B. מה יהיה המיקום של מרכז A לאחר האיתחול?

דוגמה	אורך סרטון (T)	מספר לייקים (L)
1	60	105
2	65	120
3	80	95
4	40	125

9. בניסוי אחר חנה השתמשה באתחול אקראי ומיקום המרכזים הוגרל בהתאם לטבלה להלן. לאיזה מרכז תשתייך הדוגמה 1 לאחר האתחול?

	T	L
מרכז A	80	95
מרכז B	40	125

שאלון בגרות לדוגמא (2) - מדעי הבינה המלאכותית

שאלון בגרות לדוגמא - מדעי הבינה המלאכותית

בשאלון זה 3 שאלות. יש לענות על 2 שאלות מתוך 3. 50 נקודות לכל שאלה.

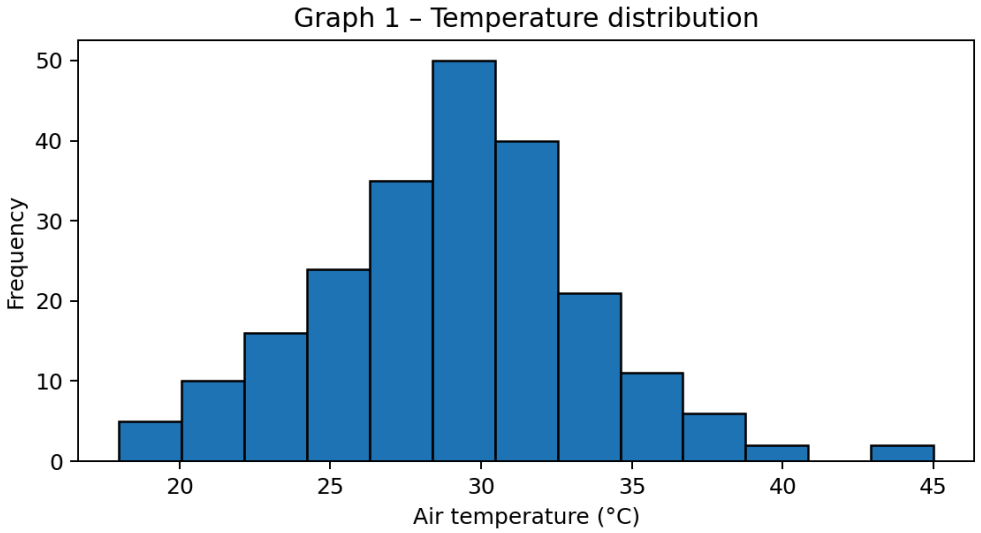
שאלה 1

נועה מפתחת מערכת לזיהוי מצוקת חום בצמחים בחממה. לכל דוגמה שלושה מאפיינים: טמפרטורת האוויר, לחות הקרקע ועוצמת האור. התגית היא האם הצמח נמצא במצוקה. בטבלה הבאה מופיעות שמונה הדוגמאות ראשונות מתוך המאגר.

דוגמה	טמפרטורה (C°)	לחות קרקע (%)	עוצמת אור (lux)	מצוקה
1	22	70	300	לא
2	24	66	400	לא
3	26	60	500	לא
4	28	55	600	לא
5	30	45	700	כן
6	32	38	800	כן
7	34	30	900	כן
8	36	20	1000	כן

1. חשבו את הממוצע ואת השונות של הטמפרטורה על בסיס הנתונים בטבלה. (הערה: יש לחשב את השונות כאילו מדובר באוכלוסייה, אין צורך בתיקון למדגם)

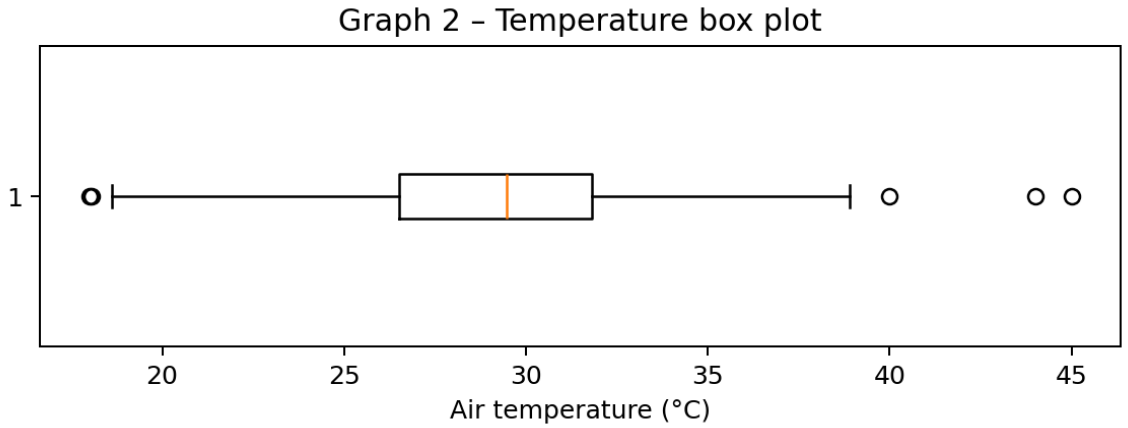
2. נתונה היסטוגרמה של ערכי הטמפרטורה (גרף 1). מה ניתן לומר על צורת ההתפלגות של המשתנה טמפרטורה?



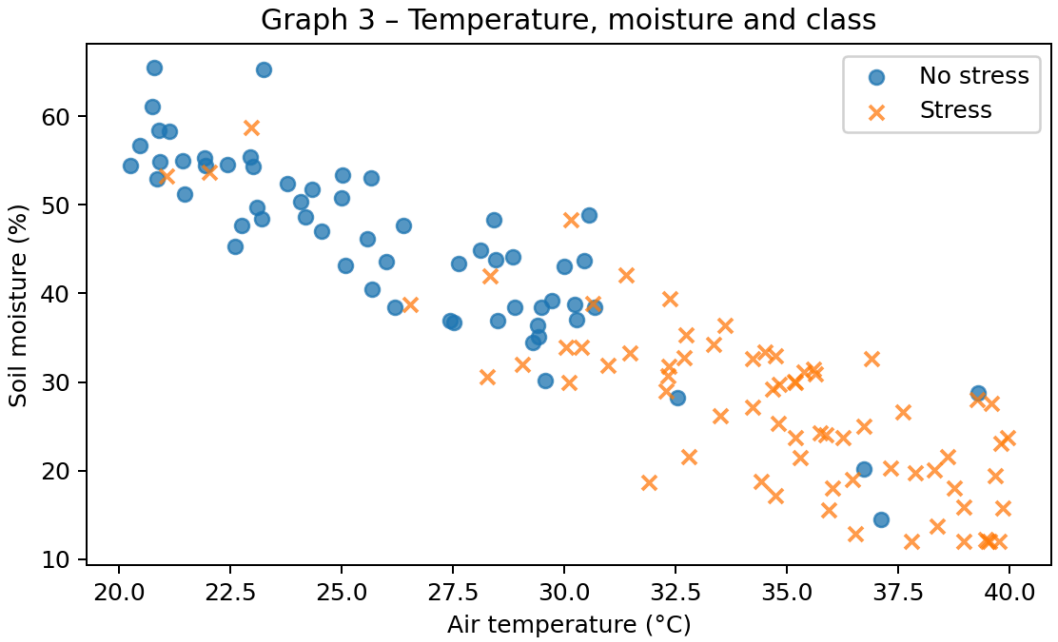
3. העריכו את ממוצע המשתנה טמפרטורה על בסיס גרף 1.

4. אורנה שמה לב שהתפלגות הטמפרטורה בטבלה הכוללת 6 דוגמאות היא אחידה ואינה תואמת את התפלגות הטמפרטורה בגרף. כיצד ניתן להסביר תופעה זו?

5. נועה הכינה.גרף קופסא של המשתנה טמפרטורה. האם נראות בגרף טמפרטורות חריגות? אם כן - אילו דוגמאות הן חריגות?



6. נועה הכינה גרף פיזור של הלחות ביחס לטמפרטורה.



a. מה ההבדל בין דוגמאות המסומנות בעיגול לבין דוגמאות המסומנות בX?
b. תארו את הקשר בין הטמפרטורה ללחות הקרקע בגרף 3.
c. האם ניתן להסיק מן הגרף שטמפרטורה גבוהה גורמת לירידה בלחות? נמקו.
d. על פי גרף 3, האם ניתן לבנות גלאי מושלם למצוקת צמחים על בסיס הטמפרטורה והלחות?

7. נועה מעוניינת לשנות את הסקלה של המשתנה עוצמת האור באמצעות אלגוריתם min max scaler. מה יהיה הערך של דוגמאות 1,5,8 אחרי שינוי הסקלה?

8. עידית חיפשה עבור נועה פונקצית ספרייה שתוכל לבצע את המרת התחום ומוצאת את פונקציית הספרייה scaler. נועה טוענת כי הפונקציה אינה מבצעת min max scaler. האם נועה צודקת? מה מבצעת הפונקציה scaler?
(הערה - הפונקציה np.percentile מחשבת אחוזונים. לדוגמא עבור הפרמטרים (x,50) מחושב החציון)

```
def scaler(X):
  a = np.percentile(X, 25)
  b = np.percentile(X, 50)
  c = np.percentile(X, 75)

  i = c - a

  X_scaled = (X - b) / i
  return X_scaled
```

9. עידית טוענת שניתן לתקן בקלות את הפונקציה כך שתבצע min max scaling. תקנו את הפונקציה כך שתבצע את המרת התחום הנכונה.

```
def min_max_scaler(X):
  a = np._______(X)
  b = np._______(X)
  c = np._______(X)

  i = c - a

  X_scaled = (X - b) / i
  return X_scaled
```

שאלה 2

אמיר מפתח מסנן לזיהוי הודעות דיוג (פישינג). כל הודעה מיוצגת בווקטור של שכיחויות מילים. בשלב הראשון נבחנות השכיחויות של שלוש מילים: urgent, link, money. אמיר מאמן אלגוריתם nearest centroid ומקבל את המרכזים הבאים:

וקטור	urgent	link	money	משמעות
c_0	1	1	0	הודעה תקינה
c_1	4	2	3	הודעת דיוג

1. בטבלה הבאה נתונה הודעה חדשה שהתקבלה. חשבו את מרחק L1 בין v לבין כל אחד מהמרכזים.

וקטור	urgent	link	money	סיווג
v	2	1	2	

2. מה יהיה הסיווג של ההודעה על פי מרחק L1?

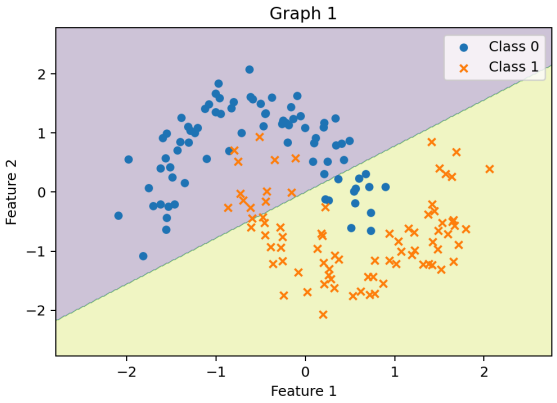
3. חשבו את מרחק L2 בין v לבין כל אחד מהמרכזים.

4. מה יהיה הסיווג של ההודעה על פי מרחק L2?

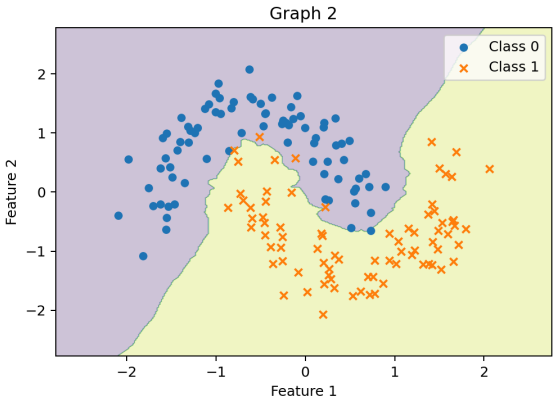
5. אמיר מצא שתי פונקציות ספריה לחישוב מרחקים, אולם הפונקציות הגיעו משובשות. השלימו את הקוד:

```
def L1_distance(a,b):
  d = 0
  for i in range(len(a)):
      d += _________________
  return d

def L2_distance(a,b):
  for i in range(len(a)):
    d += _________________
  return np.sqrt(d)
```

6. שגית הכינה שני גרפים המציגים את גבולות החלטה של שני מסווגים שאומנו על אותם נתונים: KNN ו-Nearest Centroid. קבעו איזה גרף מתאים ל-Nearest Centroid ואיזה ל-KNN. נמקו לפי צורת גבול ההחלטה.



גרף 1



גרף 2

7. איזה מן המודלים עלול לסבול מהתאמת חסר בנתונים אלו? נמקו.

8. כיצד צפוי גבול ההחלטה של KNN להשתנות כאשר עוברים מ-k=1 ל-k גדול מאוד?

9. אמיר ביצע חיפוש היפר-פרמטרים באמצעות Cross Validation. בחרו את המודל המועדף וחשבו את פער ההכללה שלו.

מודל	Scaler	Metric	k	Weights	דיוק אימון	דיוק CV
A	Standard	Manhattan	3	Uniform	0.99	0.91
B	Robust	Manhattan	7	Distance	0.95	0.94
C	MinMax	Cosine	11	Distance	0.93	0.935
D	ללא	Euclidean	7	Uniform	0.89	0.84

10. איזה מודל מציג סימנים להתאמת יתר ואיזה להתאמת חסר?

11. מה ניתן ללמוד מן הקרבה בין תוצאות B ו-C?

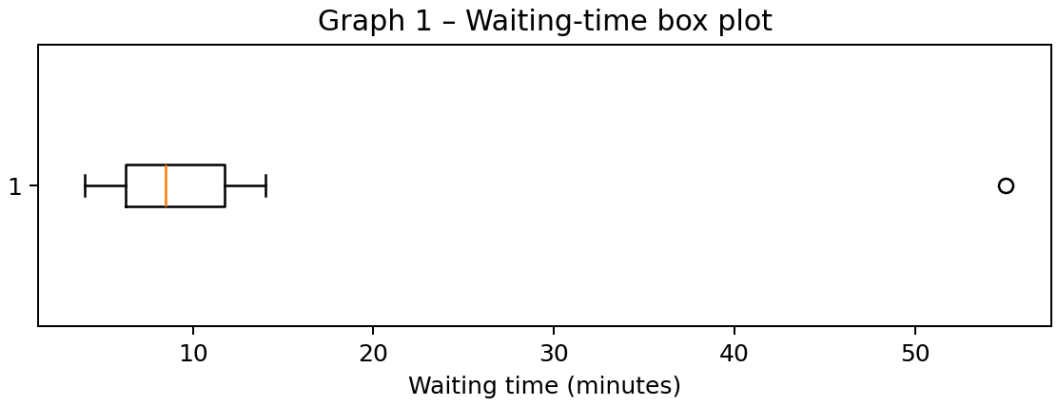
12. המודל הסופי נבדק על 200 הודעות. חשבו Accuracy, Precision, Recall ו-F1 עבור התגית "דיוג". (7 נקודות)

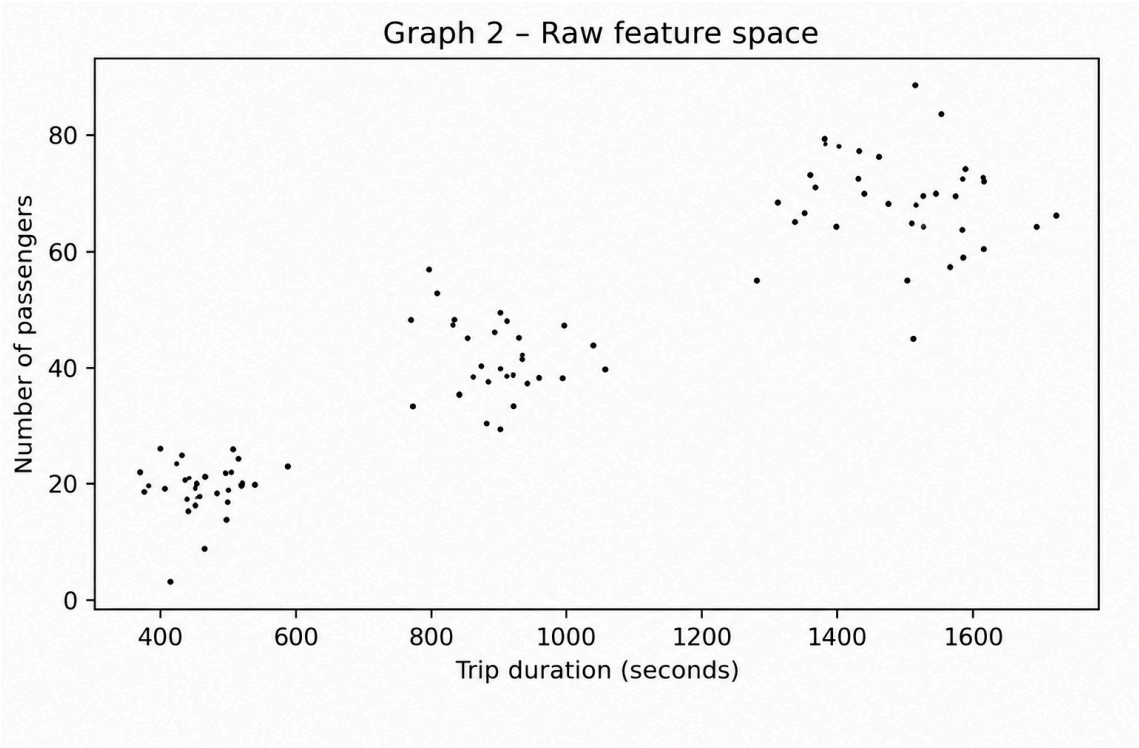
תגית אמיתית	נחזתה הודעת דיוג	נחזתה הודעה תקינה
דיוג	34	6
תקינה	12	148

13. הסבירו מדוע Accuracy לבדו עלול ליצור תמונה מטעה במקרה זה.

שאלה 3

מיכל מנתחת נסיעות בתחבורה ציבורית בעזרת K-means. בין המאפיינים: משך נסיעה, מספר נוסעים וזמן המתנה. מיכל הפיקה שני גרפים בשלב חקר הנתונים: גרף קופסא (גרף 1) גרף פיזור (גרף 2) המופיעים להלן:



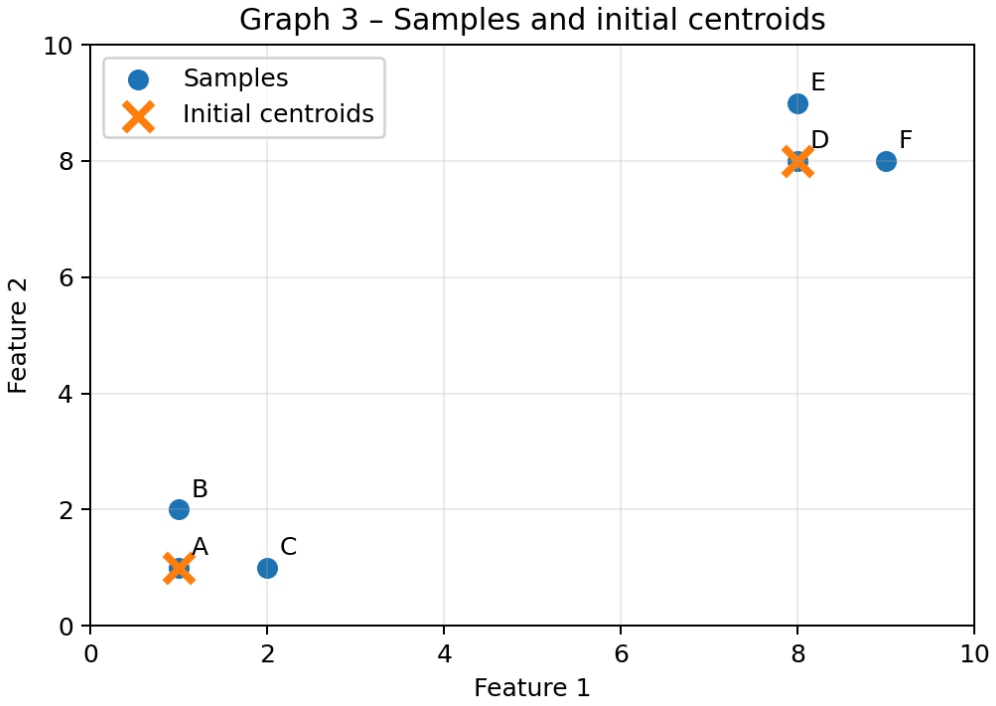


1. האם מופיעים בגרף 1 ערכים חריגים? אם כן - אלו ערכים חריגים מופיעים?
2. תארו את הקשר הנראה בגרף 2?
3. האם ניתן להסיק מגרף 2 שמספר נוסעים גדול גורם לנסיעה ארוכה יותר?
4. האם לדעתכן נדרשת המרת תחום לפני הפעלת אלגוריתם K-means על משך הנסיעה מספר הנוסעים?

5. בטבלה הבאה מופיעות 6 דוגמאות מתוך המאגר. באתחול אקראי של אלגוריתם k-means נבחרו המרכזים ההתחלתיים על דוגמאות A,D. השתמשו במרחק L2 על מנת לשייך את דוגמא C למרכז ההתחלתי הקרוב אליה. יש להציג את החישובים.

נקודה	Feature 1	Feature 2
A	**1**	**1**
B	1	2
C	2	1
D	**8**	**8**
E	8	9
F	9	8

6. בגרף 3 מופיעות הדוגמאות המפורטות בסעיף הקודם. השתמשו במרחק L2 על מנת לשייך את דוגמאות A-F למרכז ההתחלתי הקרוב אליה. בסעיף זה אין צורך לבצע חישובים.



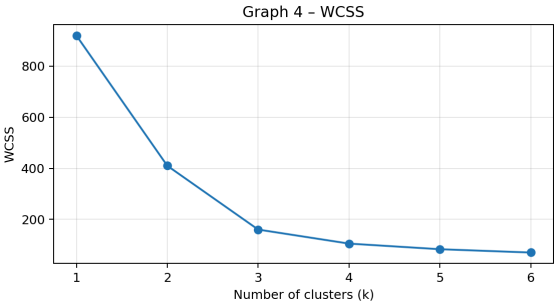
7. חשבו את מיקום שני המרכזים לאחר שלב העדכון הראשון.

8. הסבירו מהו מדד WCSS.

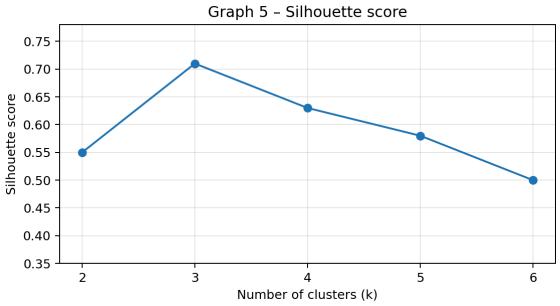
9. חשבו את WCSS מיד לאחר העדכון. הציגו את החישוב.

10. הסבירו מהו מדד silhouette.

11. מיכל בדקה מספרים שונים של אשכולות ויצרה גרפים של מדד WCSS ומדד silhoutte. בחרו מספר אשכולות מתאים על סמך שני הגרפים ונמקו.



גרף 4 – WCSS



גרף 5 – Silhouette

12. אלון טוען כי יש לבחור k=6 משום שה-WCSS שלו הוא הנמוך ביותר. האם אלון צודק? נמקו את קביעתכם.