למידה חישובית – תרגיל מספר 3

1. (35 נקודות) במחקר שנערך באנטארקטיקה החוקרים מצאו כי קיימים 3 מיני פינגווינים Adelie (152), Chinstrap(68), Gentoo (124), וכי קיימים הבדלים בין המינים במספר מדדים כמו אורך ועומק החלק העליון של המקור המכונה Culmen, אורך הכנפיים (המכונות flippers) וכן מסת הגוף (ראו דיאגרמות פיזור של המדדים באיור).

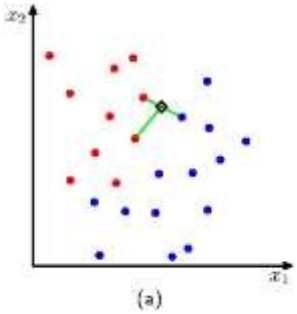
ה- dataframe מכיל 7 עמודות:

- species: penguin species (Chinstrap, Adélie, or Gentoo)
- culmen_length_mm: culmen length (mm)
- culmen_depth_mm: culmen depth (mm)
- flipper_length_mm: flipper length (mm)
- body_mass_g: body mass (g)
- island: island name

א. באמצעות הפקודות הבאות (ראו penguin_data במודל בתיקיה של exercise materials) קראו את הנתונים לתוך DataFrame:

ב. משני המינים Adelie ו- Gentoo בחרו ב- 100 ו- 80 הפרטים הראשונים בהתאמה, ומהמין Chinstrap 50 הפרטים הראשונים, ויצרו מטריצת אימון (np.array) X_train עם שתי תכונות flipper_length_mm, culmen_depth_mm, וכן וקטור y_train עם 230 תגיות מתאימות אותן יש לסמן ב- 0 (Adelie), 1 (Gentoo) ו- 2 (Chinstrap) כך שכל תגית (label) תתאים לפרט הנכון במטריצה X_train. באותו אופן יש ליצור מטריצת מבחן X_test וקטור תגיות y_test עבור הפרטים שלא נבחרו לנתוני האימון.

ג. כתבו פונקציית python למימוש אלגוריתם knn.
נתאר את שלבי האלגוריתם. בהינתן וקטור תכונות x ומדד מרחק כלשהו:



- מתוך N וקטורי אימון, זהו את k השכנים הקרובים ביותר לוקטור $\underline{x}$ ללא קשר לתיוג המחלקה שלהם.
- עבור k השכנים הקרובים ביותר, מנו את מספר הוקטורים k_i השייכים למחלקה $w_i, \quad i = 1,2,...,M$.
- שייכו את $\underline{x}$ למחלקת הרוב, כלומר למחלקה עבורה k_i הוא המקסימלי.

הפונקציה תקבל כקלט את מטריצות האימון והמבחן וכן את וקטורי התגיות המתאימים, את התכונות הנבחרות מתוך המטריצות (לדוגמה- אם רוצים להשתמש רק בשתיים מתוך התכונות), את k מספר השכנים הדרוש.
הפונקציה knn_classifier תשתמש בפונקציה dist_neigh לחישוב המרחק האאוקלידי בין כל דוגמת מבחן לכל דוגמאות האימון. הפונקציה dist_neigh

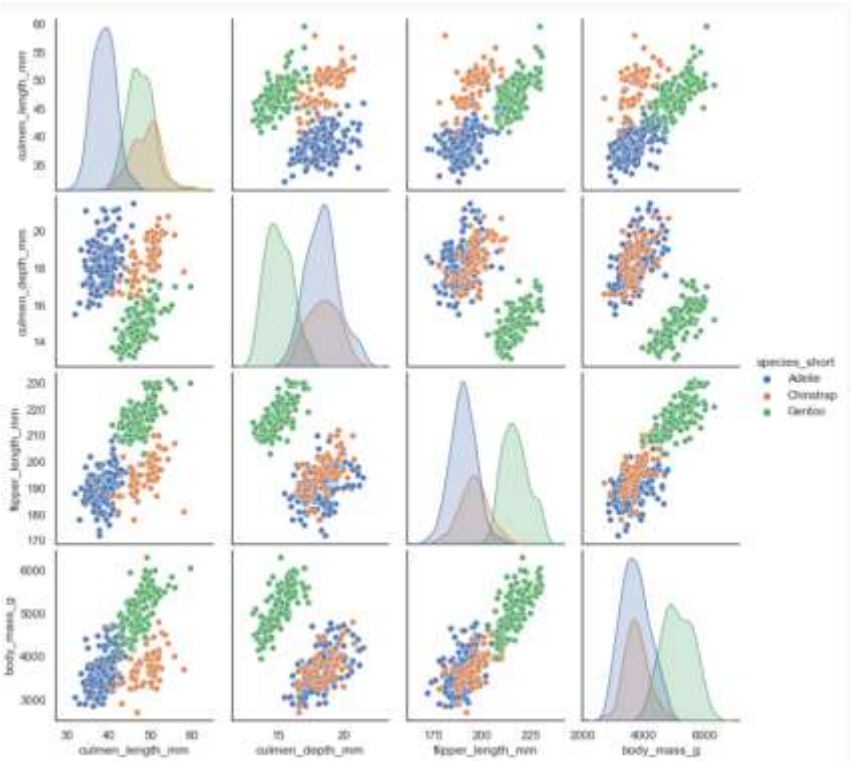
תשתמש בפקודה np.sort כדי לבצע sorting של המרחקים, תבחר את k השכנים הקרובים ביותר ותחזיר את התגיות של k השכנים הקרובים (list), ואת המרחק של כל אחד מהשכנים לדוגמת המבחן (numpy array).

מדד מרחק אאוקלידי :אם x הוא וקטור המבחן, ו- y הוא וקטור מקבוצת האימון אזי המרחק האאוקלידי d מוגדר על-ידי : $d = \sqrt{\sum_{i=1}^{L} (x_i - y_i)^2}$, כאשר L הוא מימד כל וקטור.

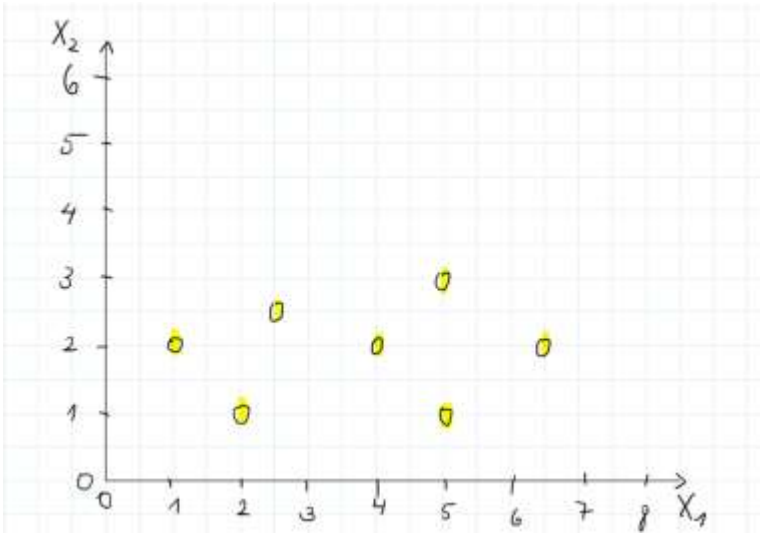
הפונקציה knn_classifier תשתמש ב- np.uniq כדי לקבוע מהם התיוגים של k השכנים, וכן את מספר השכנים בעלי אותו תיוג (השייכים לאותה מחלקה), ותשייך את דוגמת המבחן למחלקת הרוב (כלומר למחלקה עבורה מספר השכנים הוא מקסימלי). במידה ויש מספר זהה של שכנים משתי מחלקות או יותר, הפונקציה תחשב את סכום המרחקים של השכנים של כל מחלקה מדוגמת המבחן, ותשייך את דוגמת המבחן למחלקה עבורה סכום המרחקים הוא מינימלי.

ד. הפעילו את הפונקציה שכתבתם על ה- penguin dataset, ובדקו מהו אחוז הזיהוי הנכון עבור k=1, 3, 5. הדרכה : נתוני האימון הם שורות X_train, וקבוצת המבחן היא X_test, כאשר מתוך שתי המטריצות התכונות הרלבנטיות הן culmen_depth_mm ו- flipper_length_mm .

ה. חזרו על הסעיף הקודם אך עם כל התכונות של ה- penguins dataset.



2. (15 נקודות) וקטורי התכונות הבאים (נקודות ב- R^2), מייצגים מדדים של שנבוב (aardvark) שלא היה ידוע לחוקרים. הזואולוגים טוענים כי קיימים שני סוגים של שנבוב. כדי להבחין בין הסוגים מפעילים את אלגוריתם k-means עבור k=2. הצנטרואידים ההתחלתיים הוגרלו בנקודות C1=(2,3), C2=(4,3).



סמנו את הצנטרואידים ההתחלתיים על התרשים.

א. בצעו את צעד השיוך ההתחלתי, רשמו את הנוסחה באמצעותה מתבצע השיוך כתבו את החישובים שאתם מבצעים. באיזה כלל השתמשתם?

ב. חשבו את הצנטרואידים של הצעד השני, ורשמו את הנוסחה בה השתמשתם.

ג. מהי שגיאת ה- clustering לאחר הצעד השני?

3. (20 נקודות) א. ממשו את אלגוריתם ה- k-means. הפונקציה הראשית תיקרא k_means ותעשה שימוש בשלוש פונקציות python:

- פונקציית איתחול הצנטרואידים – init_centroid, הפונקציה תבחר באופן אקראי k צנטרואידים. הצנטרואידים ייבחרו מתוך הנקודות של ה- dataset.
- פונקציית שיוך נקודות הדגימה הנתונות במטריצה X לפי המרחק מהצנטרואידים assign_samples (זוהי למעשה פונקציה המבצעת 1-NN).
- פונקציית חישוב הצנטרואידים לכל קבוצת שיוך - centroid_calc
- פונקציה המחשבת את שגיאת ה- clustering

הפונקציה K_means תקבל כקלט נקודות X data, מספר clusters רצוי k, ערך סף מתוך שגיאת ה- clustering לעצירת הריצה כאשר השינוי לא יותר גדול מערך הסף הנ"ל, וכן מספר איטרציות מקסימלי.

הפונקציה תחזיר וקטור y עם שיוך כל נקודת data ל- cluster, כאשר כל אחד מה- clusters יסומן במספר 0,1,2 וכו' לפי מספר ה- k clusters. כמו כן הפונקציה תחשב ותחזיר את שגיאת ה- clustering.

ב. הפעילו את האלגוריתם על ה- penguins dataset עבור כל נקודות הנתונים אך רק עבור התכונות culmen_depth_mm ו- flipper_length_mm , וצבעו כל אחת מהנקודות בציור על-פי הצביר (cluster) אליו היא שייכת. סמנו את הצנטרואידים של כל אחד מהצבירים על-ידי x בצבע המתאים ל- cluster.
חשבו וכתבו כמה פרטים מכל אחד מהמינים שויכו על-ידי אלגוריתם ה- k-means לאותו cluster. הפעילו מספר פעמים את הפונקציה שכתבתם על הנתונים, כאשר בכל פעם בדקו את שגיאת ה- clustering. בחרו את הפתרון המביא לשגיאה מינימלית.

ג. חזרו על הסעיף הקודם אך עם כל התכונות.

4. (15 נקודות) בתרגיל זה נבצע שוב Kmeas על ה- penguin dataset אך עתה באמצעות KMeans של scikit learn. צפו בהקלטה של הרצאה 16 וכתבו קוד המצייר את נקודות ה- data כאשר לכל מין של פינגווין סימון אחר מתוך הסימנים (x, +, o, triangle, diamond etc.).
הפעילו את KMeans והשוו את ציור ה- clusters וכן את שגיאת ה- clustering עבור n_init = 10 ,n_init= 1, עם 5 הרצות לכל אחד.
מהי מסקנתכם?
הריצו שוב עם n_init = 10 וציירו את ה- clusters אותם מקבלים, כאשר כל cluster מסומן בצבע אחר, והשוו לציור הקודם.

5. (15 נקודות) תהי X קבוצת הלימוד הבאה:

$$X = \{x_i,\ i = 1,2,3,4,5\}$$
$$x_1 = \begin{pmatrix}1\\2\end{pmatrix},\ x_2 = \begin{pmatrix}2\\3\end{pmatrix},\ x_3 = \begin{pmatrix}5\\5\end{pmatrix},\ x_4 = \begin{pmatrix}8\\5\end{pmatrix},\ x_5 = \begin{pmatrix}7\\9\end{pmatrix}$$

$$I(x) = \begin{pmatrix}1 & 2\\2 & 3\\5 & 5\\8 & 5\\7 & 9\end{pmatrix}$$

א. כתבו פונקציה המחשבת את מטריצת המרחק D(x) לפי מטריצת העוצמה I(x). השתמשו במרחק אאוקלידי כדי ליצור את מטריצת המרחק.
ב. הפעילו hierarchical clustering עם מרחק ממוצע (average linkage).
ג. ציירו את הדנדרוגרמה המתקבלת עם אורך ענפים לפי המרחק.