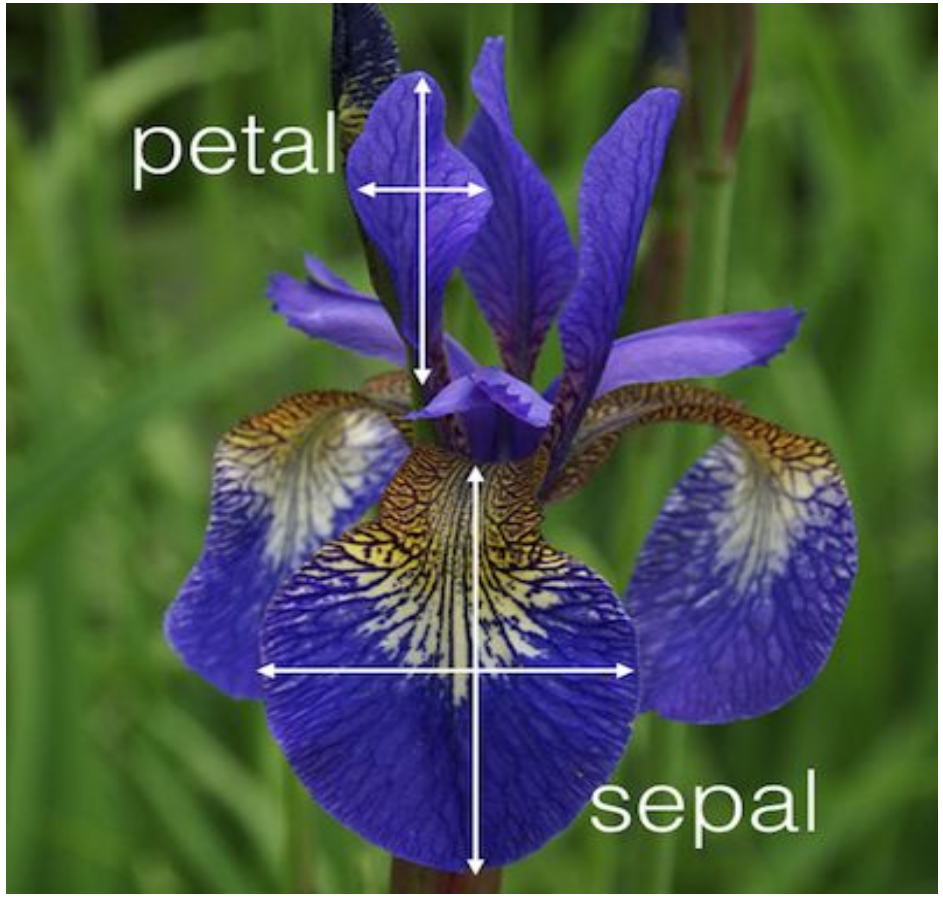
הטכניון
מכון טכנולוגי
לישראל

K-Nearest Neighbors (KNN)

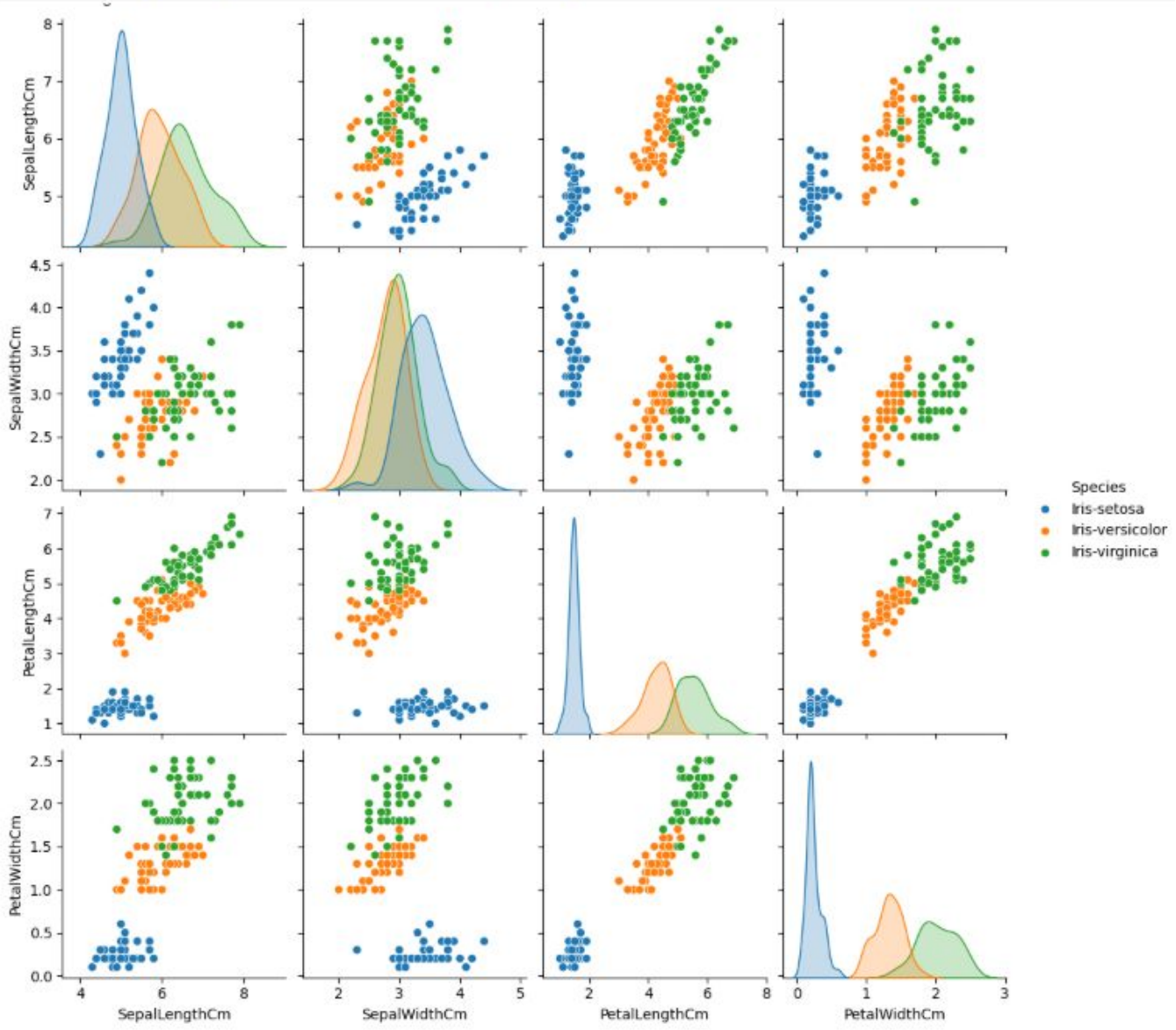
דוגמא: סיווג אירוסים

מאפיינים של פרח



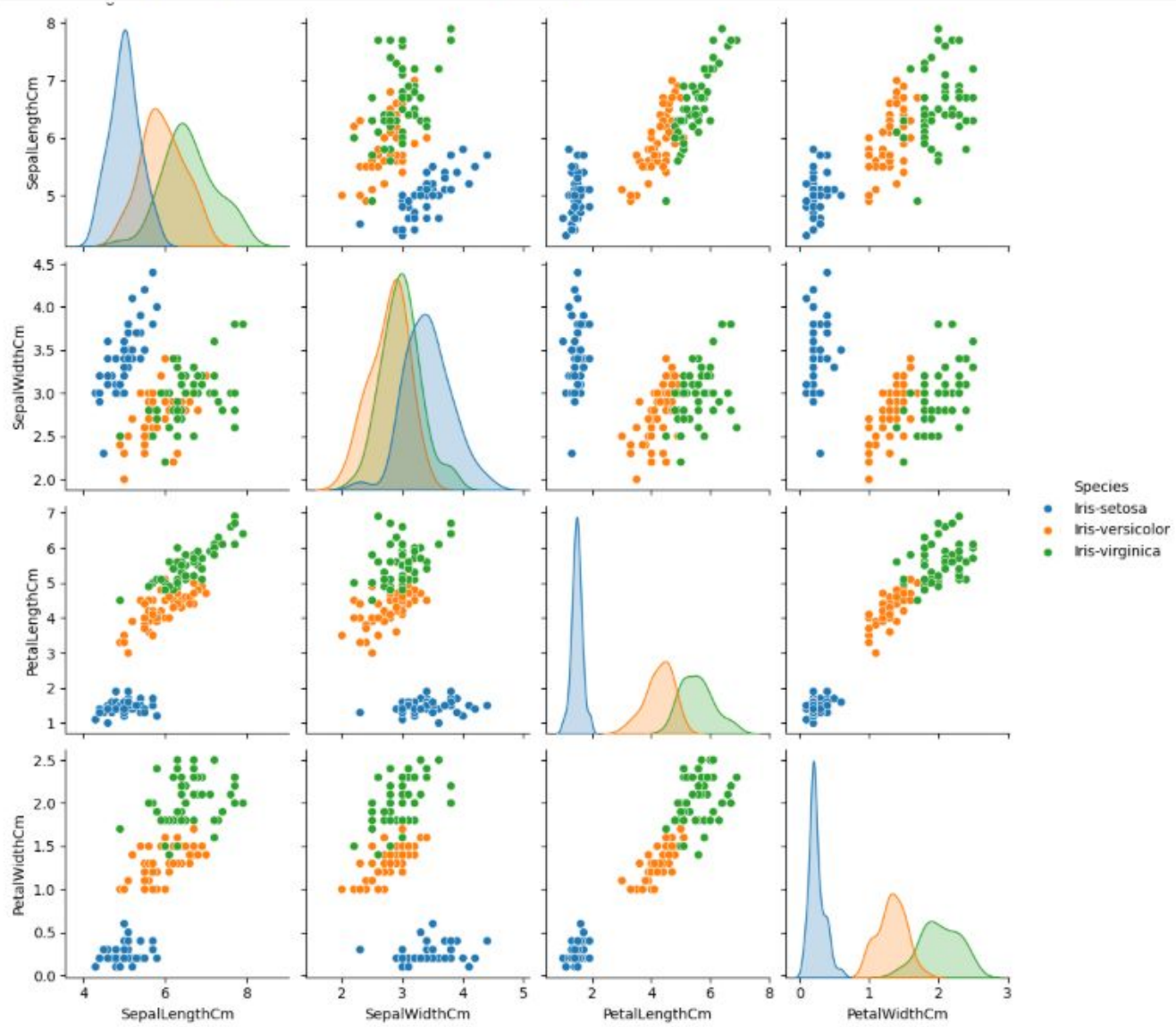
מאפיין	משמעות
Sepal length	אורך עלי גביע
Sepal width	רוחב עלי הגביע
Petal length	אורך כלי הכותרת
Petal width	רוחב עלי הכותרת

```
sns.pairplot(df.drop('Id',axis=1), hue = 'Species')
```



פיזור המאפיינים

```
sns.pairplot(df.drop('Id',axis=1), hue = 'Species')
```

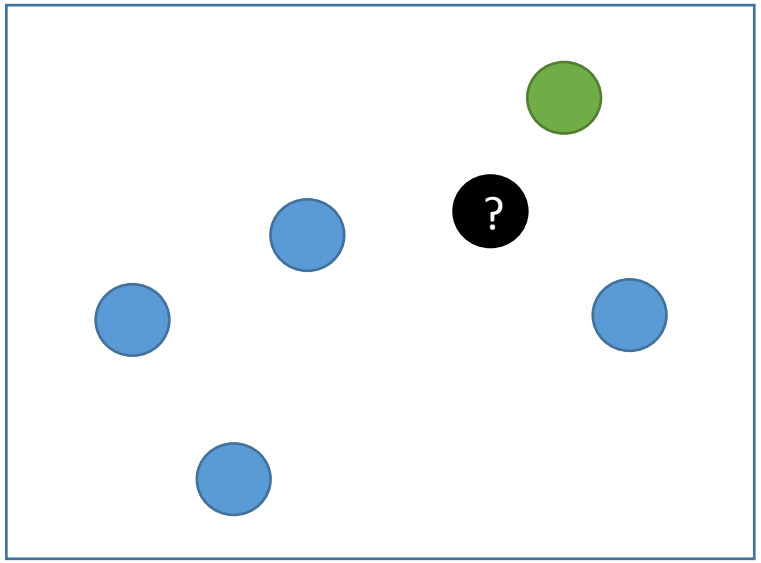


תרגיל

הציעו גישות לסיווג פרח על פי מאפייניו.

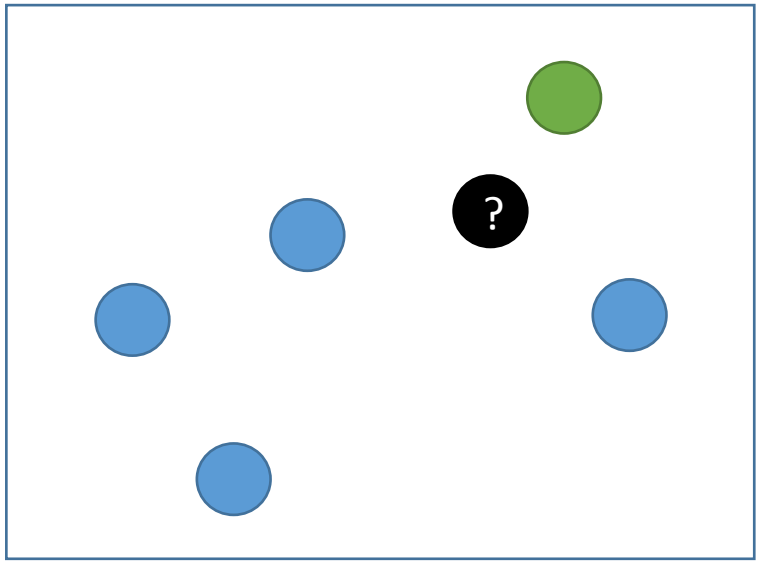
KNN

סיווג הדגימה על פי סיווג השכנים שלה

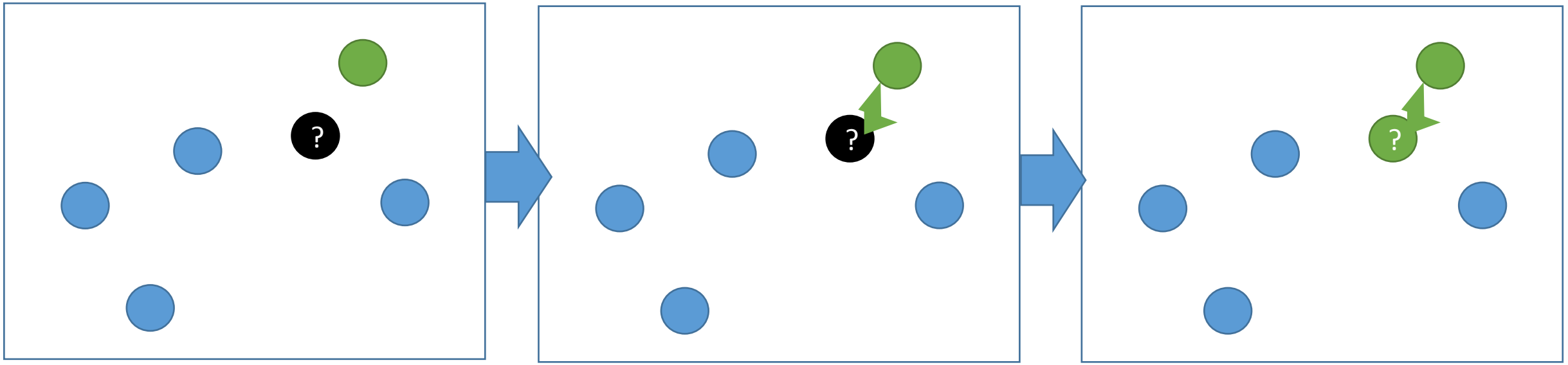


חוק הסיווג

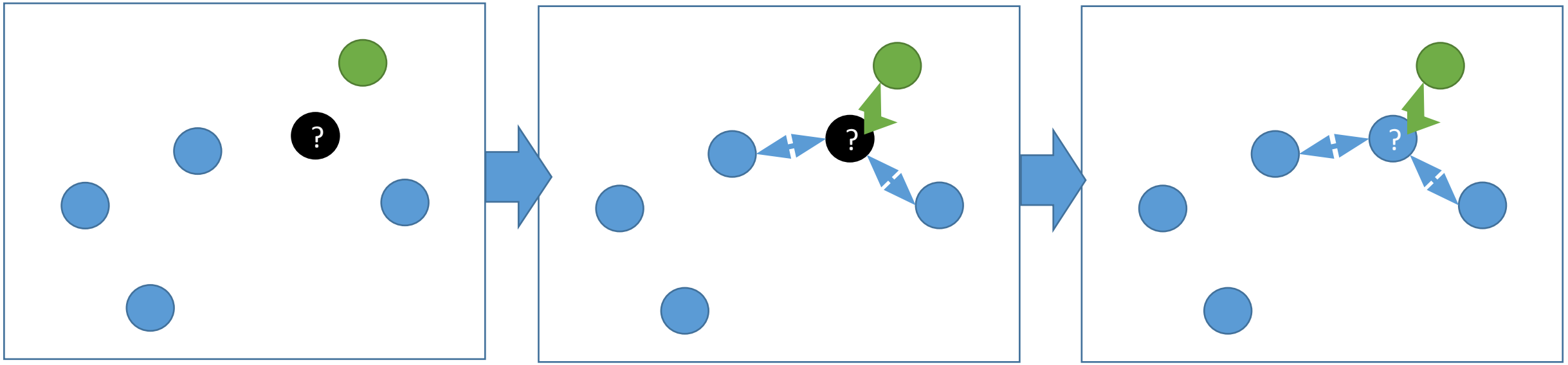
נבחר את ה-label של רוב הדוגמאות
מתוך K השכנים הקרובים



דוגמא – K=1

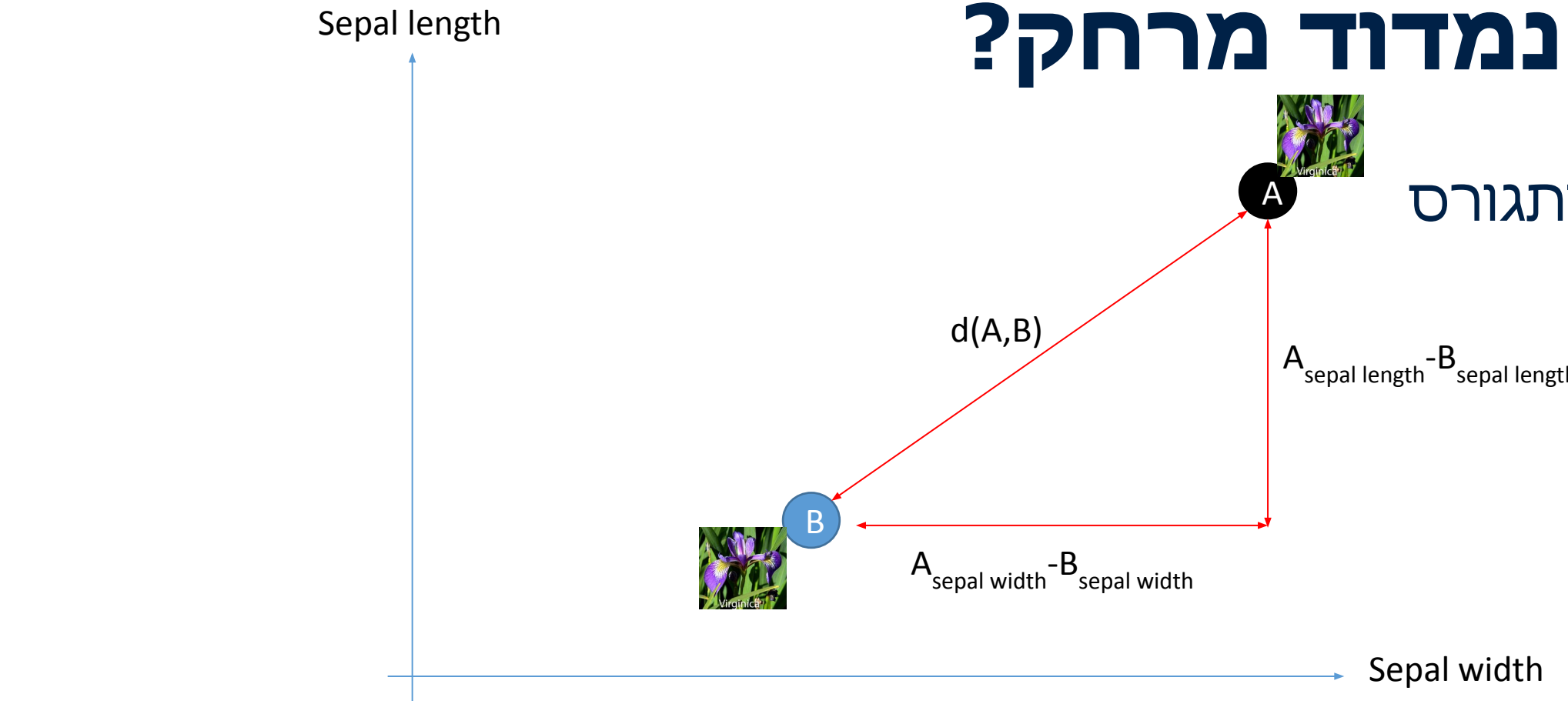


דוגמא – K=3



כיצד נמדוד מרחק?

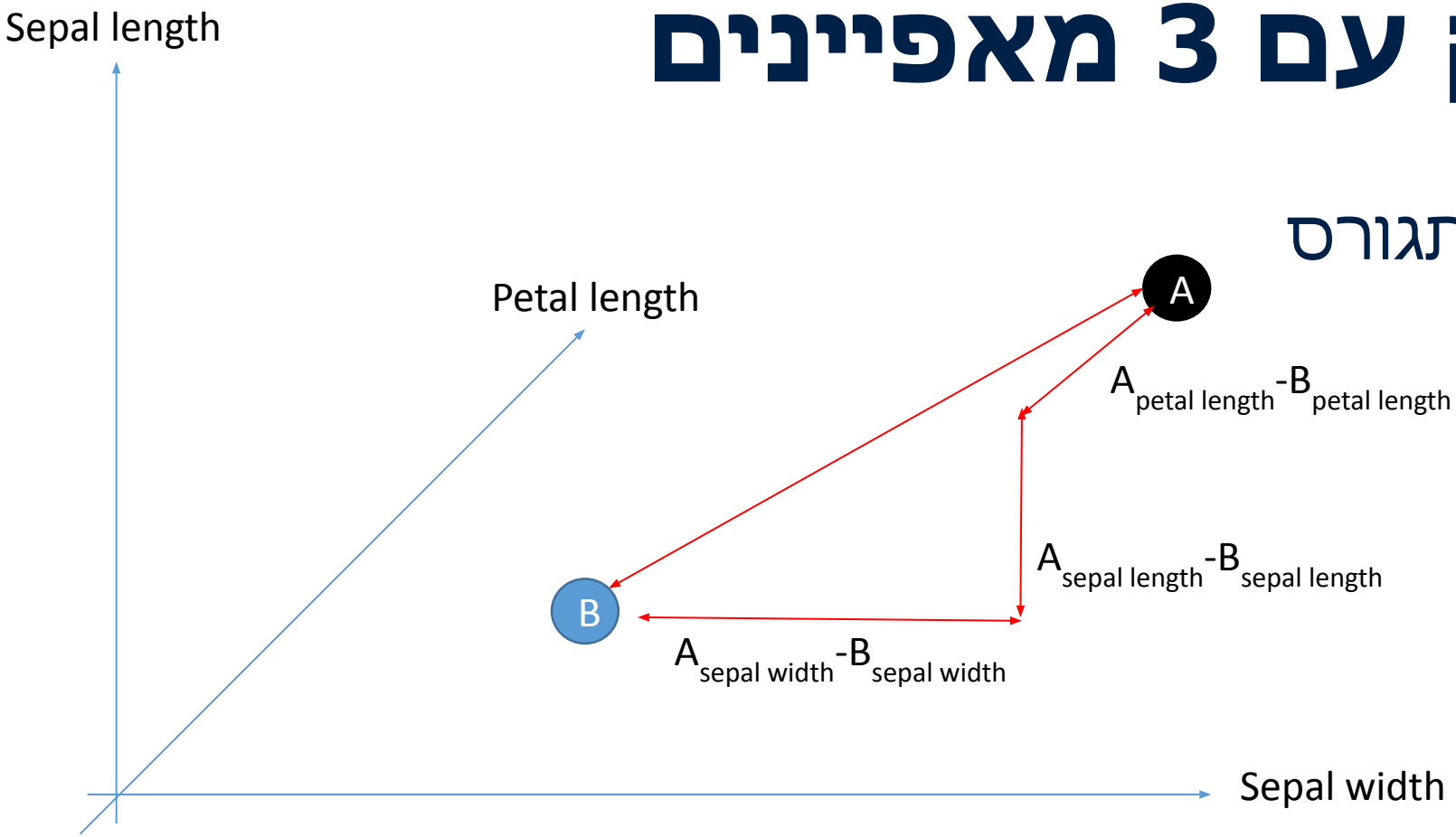
משפט פיתגורס



$$d(A,B) = sqrt((A_{sepal\ length}-B_{sepal\ length})^2 + (A_{sepal\ width}-B_{sepal\ width})^2)$$

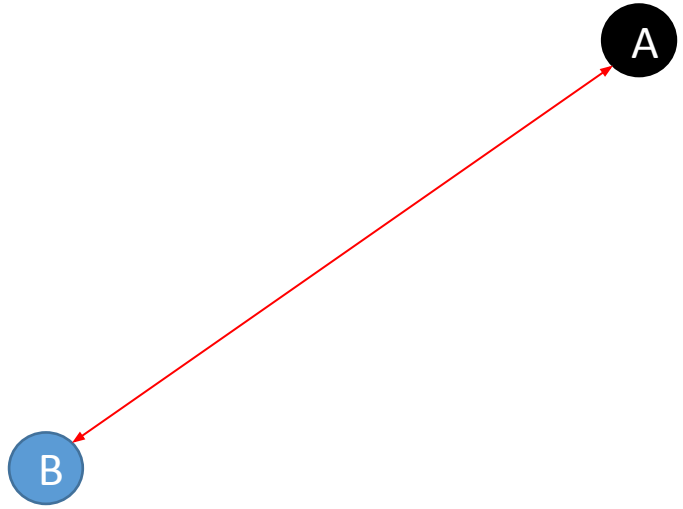
מרחק עם 3 מאפיינים

משפט פיתגורס



$$d(A,B) = \sqrt{(A_{sepal\ length}-B_{sepal\ length})^2 + (A_{sepal\ width}-B_{sepal\ width})^2 + (A_{petal\ length}-B_{petal\ length})^2}$$

מרחק עם 4 מאפיינים



$$d(A,B) = \sqrt{(A_{\text{sepal length}} - B_{\text{sepal length}})^2 + (A_{\text{sepal width}} - B_{\text{sepal width}})^2 + (A_{\text{petal length}} - B_{\text{petal length}})^2 + (A_{\text{petal width}} - B_{\text{petal width}})^2}$$

מרחק עם n מאפיינים

…,The distance between object A with features a1,a2

…,and object B with features b1,b2

:is

$$d(A,B) = \text{sqrt}\left((a_1-b_1)^2 + (a_2-b_2)^2 + \ldots + (a_n-b_n)^2 \right)$$

OR

דוגמה

בידינו מאגר נתונים של מתכונים המסווגים לשני סוגים: Muffin ו Cupcake:

לכל מתכון מוגדרים מספר מאפיינים (תכונות) כמו: כמות קמח, חלב, סוכר, חמאה, ביצים, אבקת אפייה, וניל ומלח

Type	Flour	Milk	Sugar	Butter	Egg	Baking Powder	Vanilla	Salt
Muffin	47	26	10	10	4	1	0	0
Muffin	50	17	17	8	6	1	0	0
Muffin	50	17	17	11	4	1	0	0
Cupcake	39	0	26	19	14	1	1	0
Cupcake	42	21	16	10	8	3	0	0
Cupcake	34	17	20	20	5	2	1	0

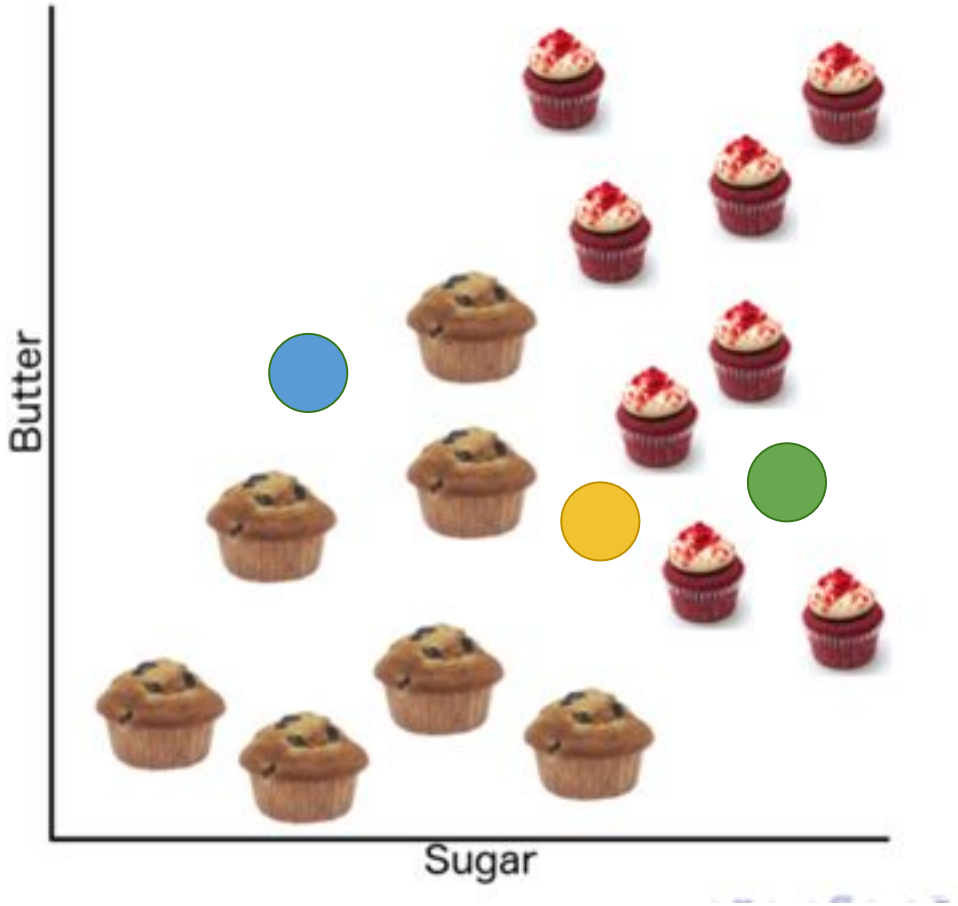
נרצה לקבל מתכון חדש ולקבוע האם הוא נחשב למאפין או קאפקייקס

אלגוריתם השכן הקרוב ביותר - NN

לצורך הסבר האלגוריתם נבחרו מתוך כל התכונות המאפיינות מתכון רק שתי תכונות.

מתוך הנתונים המופיעים בגרף:

- כמה דוגמאות אימון יש?
- אילו תכונות נבחרו?
- איזה סיווג הייתם בוחרים עבור
  - הנקודה הירוקה?
  - הנקודה הצהובה?
  - הנקודה הכחולה?



אלגוריתם השכן הקרוב ביותר - NN - חישוב מרחק

נסתכל על שתי הרשומות הבאות:

	Type	Flour	Milk	Sugar	Butter	Egg	Baking Powder	Vanilla	Salt
1	Muffin	47	26	10	10	4	1	0	0
2	Muffin	50	17	17	8	6	1	0	0

נחשב את המרחק בין מתכון 1 למתכון 2:

$$d(recipe1, receipe2) = \sqrt{(47-50)^2 + (26-17)^2 + (10-17)^2 + (10-8)^2 + (4-6)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2}$$

סיכום ביצוע אלגוריתם KNN

- שלב האימון (fit):
  - שמירת נתונים במאגר
- שלב החיזוי (predict):
  - קבלת דוגמאות לחיזוי
  - בעבור כל דוגמה
    - מציאת המרחקים של הדוגמה מכל הדוגמאות שבמאגר האימון
    - מציאת K השכנים הקרובים ביותר
      - באמצעות לולאה מקוננת
      - או באמצעות מיון ולולאה יחידה
    - בדיקת הסווג השכיח מבין K השכנים הקרובים ביותר
      - נמנה כמה יש מכל תגית
      - נמצא את המקסימום ונחזיר את התגית

תרגיל 4.1 – דף עבודה KNN

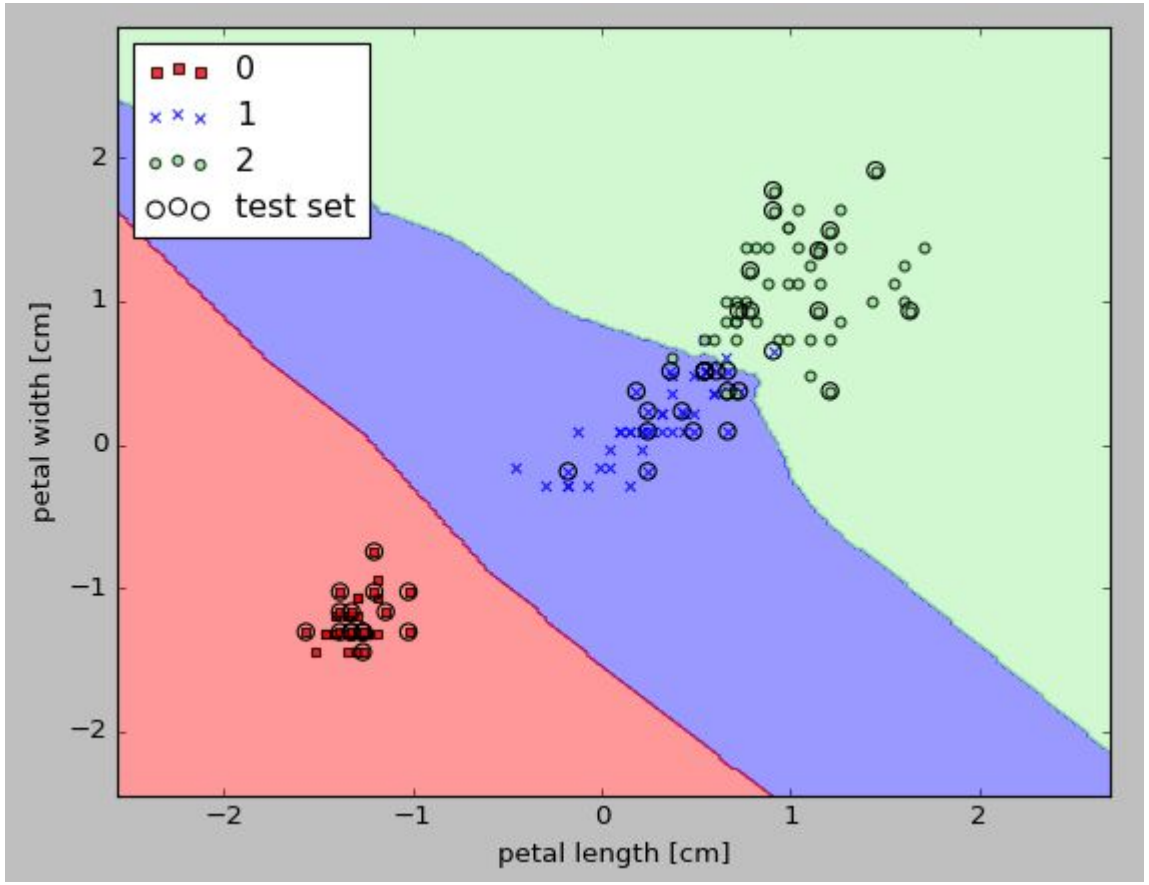
עוד קצת על חישוב מרחקים

- מה עושים עם מאפיינים קטגוריאליים?
  - אם אין סדר: One hot encoding
  - אם יש סדר: המרה למספרים
- מה עושים עם מאפיינים בסקאלות שונות?
  - scaling
- מה עושים עם מאפיינים בעלי חשיבות שונה?
  - Weighted scaling

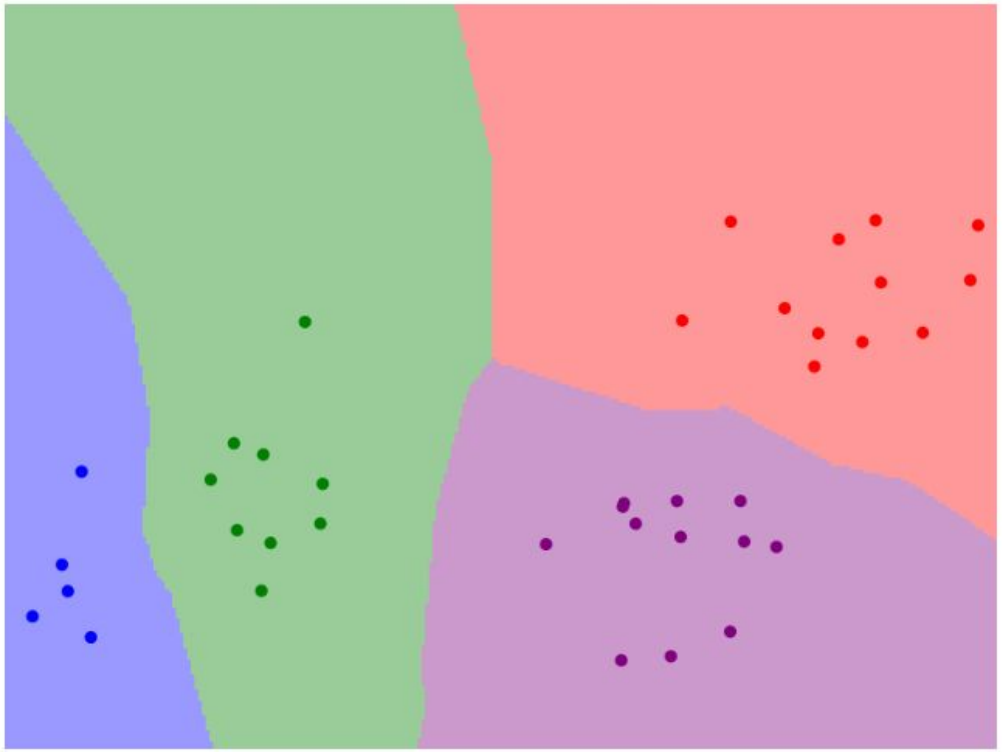
תרגיל 4.2 – תרגיל חזרה אישי

תרגיל חזרה אישי

גבולות החלטה



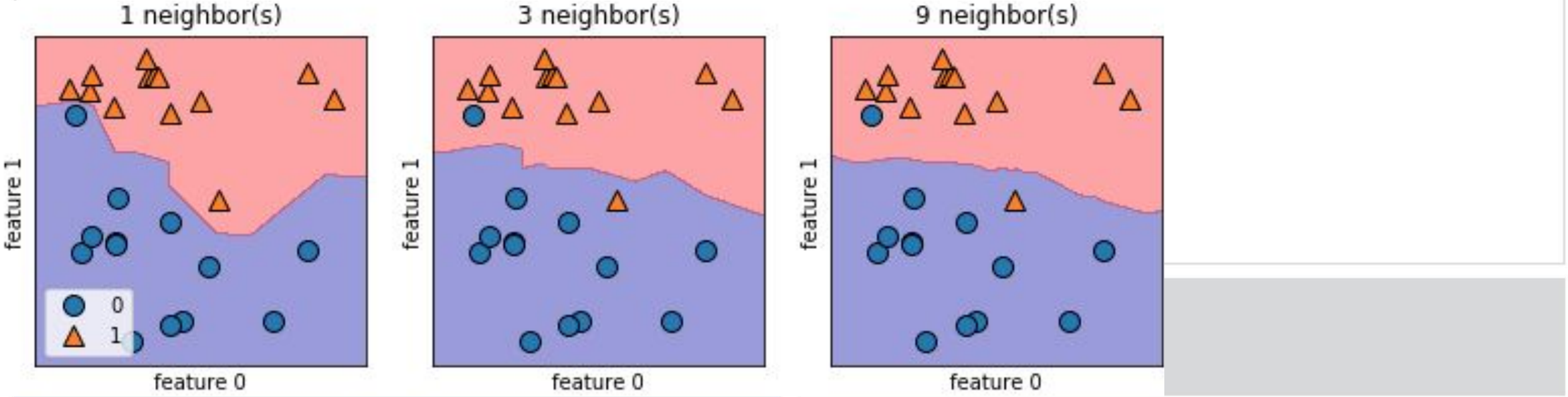
KNN - הדגמה



http://vision.stanford.edu/teaching/cs231n-demos/knn/

השפעת הפרמטר K

K קטן
קו החלטה פחות רציף – מודל יותר מורכב – הכללה פחותה
K גדול
קו החלטה חלק יותר – מודל פשוט יותר – הכללה טובה יותר



הטכניון
מכון טכנולוגי
לישראל

תודה על ההשתתפות