

*Machine learning*

---

# Classification & KNN

Lecture II

---

פיתוח:  
ד"ר יהונתן שלר  
משה פרידמן

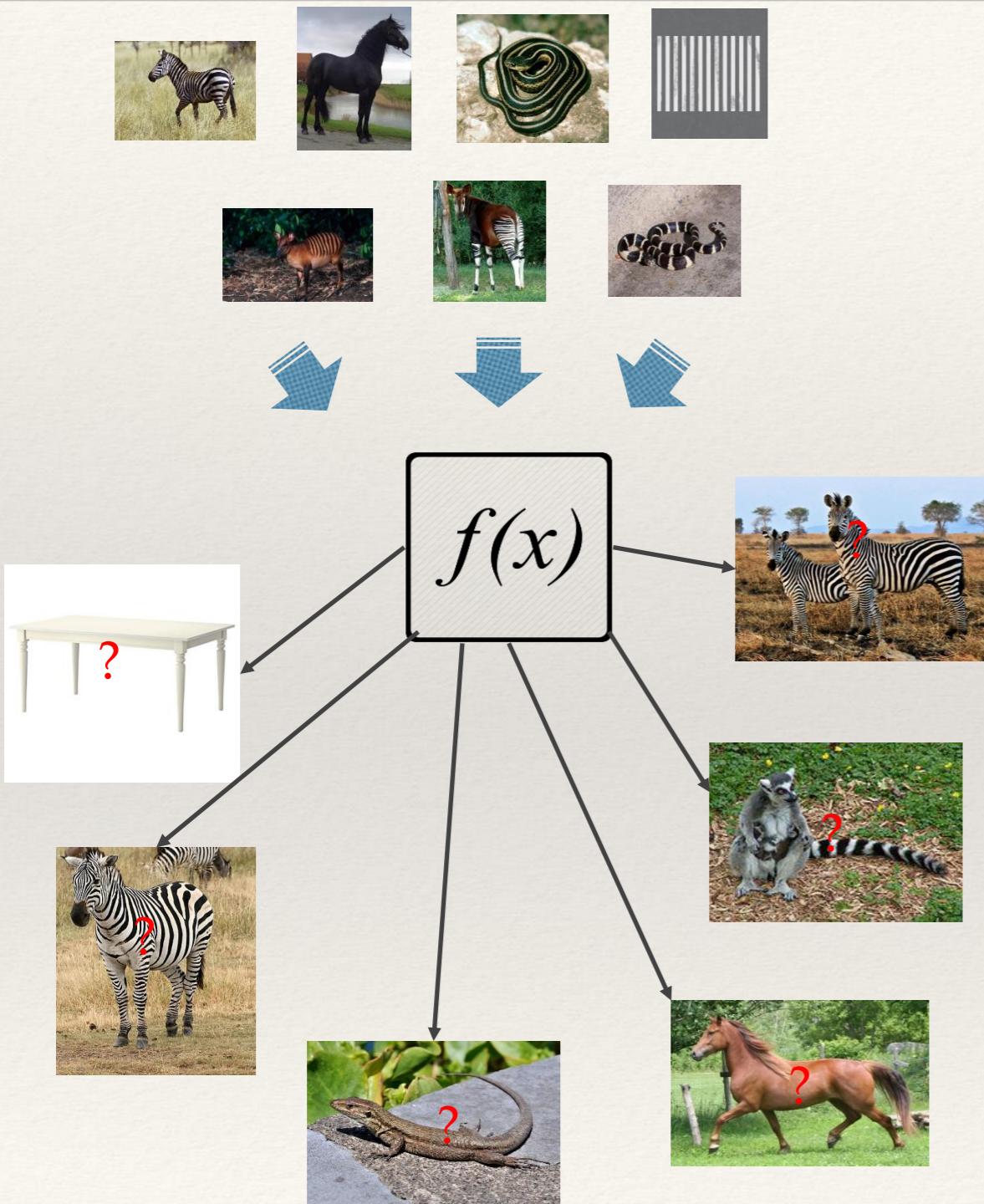
# **בעיות סיוג – שלבי למידה - תזכורת**

- א. מידול (Modeling)
- ב. הכנות Dataset
- ג. בניית מודל סיוג (ע"י למידה (Learning))
- ד. שיעורך (evaluation)

**שימוש במודל:**

**סיוג/ניבוי (classification/prediction)**

# מהו מודל סיוג - תזכורת



**מודל סיוג:**

הינו פונקציית סיוג - האמורה להחליט,  
עבור *instance*, מה התשובה ( מבין  
התשובות שהגדנו ) לשאלת המוגדרת

**מהי פונקציות סיוג?**

- ❖ פונקציה שתשקל את המאפיינים ב-*feature vector*, כדי לקבל החלטה (יוסבר בהמשך)
- ❖ בדו' לעיל מדובר בעז החלטה (יוסבר עוד בהמשך)

# KNN – מודל הסיווג הראשון

טוב שכן קרוב מאח רחוק (משל)



# מודל הסיווג הראשון – KNN

## K- Nearest Neighbors

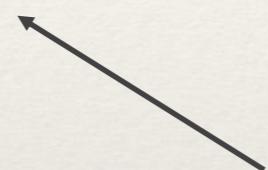
### K-NN – הרעיון הכללי:

- בודקים מהם ה"שכנים"
- נסועג את הקטגוריה, לפי הקטגוריה של ה"שכנים"
- אבל, מה הקשר לעניינו?
- מי הם ה"שכנים"?



# מודל הסיווג הראשון – KNN

## K- Nearest Neighbors



**מיهو שכן?**

- השכנים הם דוגמאות האימון
- הדוגמה אותה בודקים, אינה חלק מהאימון



# – אלגוריתם הסיג – KNN

## Input:

- ❖ k – the number nearest neighbors; the set of training examples.

## The KNN Algorithm:

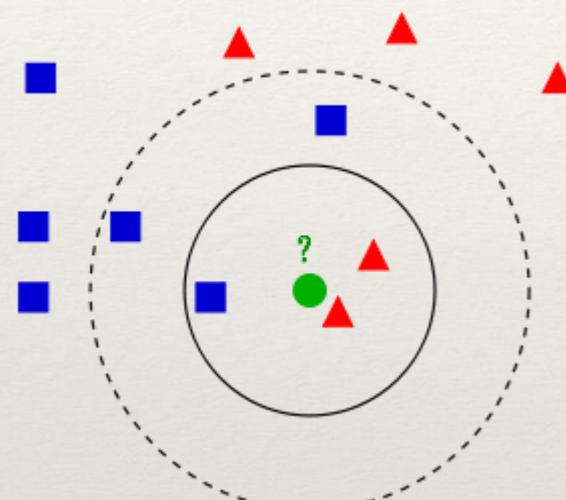
- ❖ for test instance  $x_j$  in the test-set:
  - ❖ Calculate  $d(x_j, x_i)$
  - ❖ Select the k closest training examples,  $d(x_j, x_i)$  sorted
  - ❖ Use majority voting to classify the test examples

## Notations and Terms:

$x_j$  – example (number j) from the test-set  
 $x_i$  – example (number i) from the train-set  
 $d(x_j, x_i)$  – distance function – measures distance between  $x_j$  and  $x_i$ .

# מודל הסיווג הראשון – KNN

- הדוגמה החדשה, מסומנת בעיגול השחור
- הדוגמאות מהאימון מסומנים בעיגולים שחורים ואדומים
- בדוגמה  $K=3$



Demo

תודה לד"ר יונתן רובין

---

# KNN כמודל סיוג

❖ ניקח צעדי לאחור ונברר כמה נושאים קודם ...

# KNN - שאלות עליהם נרצה לענות

- ❖ מהם **סוגי המחלקות**, ומה הקשר שלהם לבעיית הסיווג?
- ❖ מהם **סוגי המאפיינים** ומה הקשר ביניהם ובין ה-**feature vector**?
- ❖ מהו **train-set** ו**test-set** ומהו?
- ❖ את מה אנחנו מכנים שכנים?
- ❖ מהו הוקטור?
- ❖ איך מציגים דוגמה כוקטור?
- ❖ מה הופך את השכנים לקרים? איך מודדים קירבה?
- ❖ זכרים את פונקציית הסילום? מה הקשר ל-KNN?
- ❖ איך בסוף מקבלים את ההכרעה?

# מחלקותesi... ...ו

בביעות סיוג – המטרה היא ללמד מודל שתפקידו להזות מהי הקטגוריה/מחלקה (category/label/tag/class) של דוגמה חדשה.

- ❖ **ערכי המחלקה** – ערך המחלקה יכול להיות קטgoriy (כמו בסוגי האירוסים) או מספרי בד"ד (כמו בערך הטלת קוביה)
- ❖ **מחלקות עם ערכיים קטgoriyim** – עבור מקרים בהם הערכים קטgoriyim, בד"כ נרצה להמיר את ערך המחלקה לערך מספרי בד"ד.

a. מידול (Modeling

▪ שאלה סיוג; קטgoriyת התשובה;

מאפיינים

מחלקות – ערכיים קטgoriyim.  
לרוב - נצרכ להמיר בערכיים  
**בדידים מספריים**

סוגי אירוסים



Versicolor



Setosa



Virginica

# המאפיינים – מרחב המאפיינים לוקטור המאפיינים

בעית האירוסים



על גביע (sepal):

❖ אורך, רוחב

על כותרת (petal):

❖ אורך, רוחב

sepal		petal	
length	width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2



בעית הזברות

Is animal	Vertical stripes	Black & white	4 legs	large
0	1	1	0	0
1	0	1	0	0
1	0	0	1	1
1	1	1	1	1

מאפיינים –  
ערכיהם בדים  
(מספריים)

א. מידול (Modeling):  
שאלת סיווג; קטגוריה התשובה; **מאפיינים**: (features/attribute)

בלמידת מכונה, מתייחסים לחלק מהמאפיינים כדי למדל דוגמא אחת (צפויות).

נתיחס לכל דוגמה Feature Set (מרחב המאפיינים): המאפיינים שבאמת עבור

דוגמה מסוימת Feature Vector (instance)

חיה: (כן, לא)

פסים אנכיים: (אנכיים, אופקיים,  
לא)

צבעים: (שחור, לבן, חום, ...)

רגליים: (2, 4, לא)

גודל החיה: (גדולה, בינונית,  
קטנה)

**ערך המאפיין:**  
ערך המאפיין יכול להיות  
קטורי, מספר בדיד או מספרי  
רציף  
עבור ערכיהם קטgorים, בד"כ  
(וגם ב-KNN), נרצה להמיר  
את ערך המאפיין לערך מספרי  
בדיד.

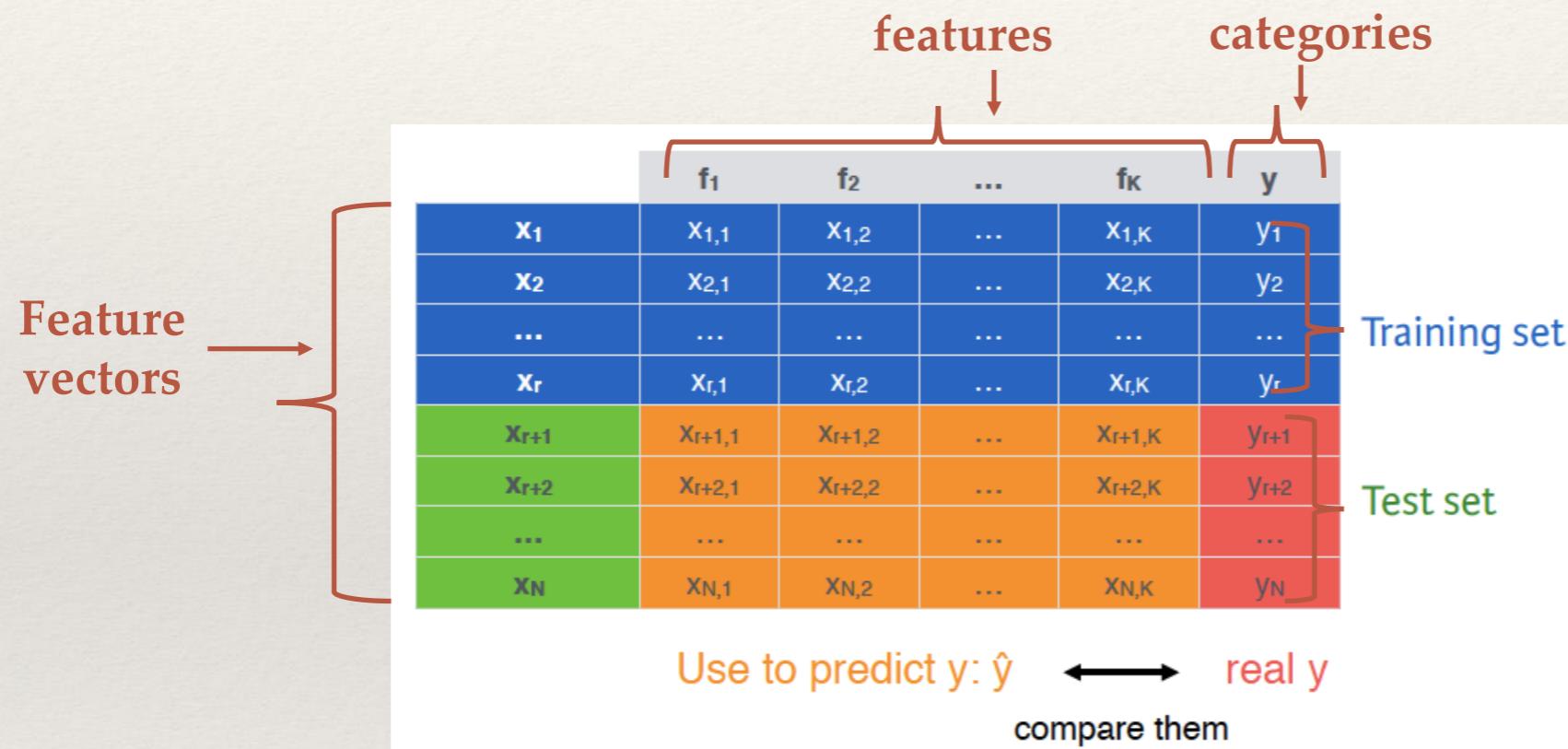
מאפיינים –  
ערכיהם בדים  
(מספריים)

מאפיינים –  
ערכיהם  
קטgorים.

# dataset – train-set and test-set

Training Dataset: The dataset that we use to train the model

- ❖ The model *sees* and *learns* from this data.



Test dataset: The dataset that provides the gold standard used to evaluate the model.

- ❖ It is only used once a model is completely trained.

# שאלות ביניים

שאלה 1:

מה הקשר בין בעיית הסיווג למחלקות האירוסים? האם סוגי האירוסים הם ערכיים קטגוריים או מספריים בדידים? כיצד מmirים ערכים קטgorיים למספריים בדידים?

שאלה 2:

מהו מאפיין? מה הקשר בין המאפיין לדוגמאות האימון? מה הקשר ל-  
feature vectors וול-feature set

שאלה 3:

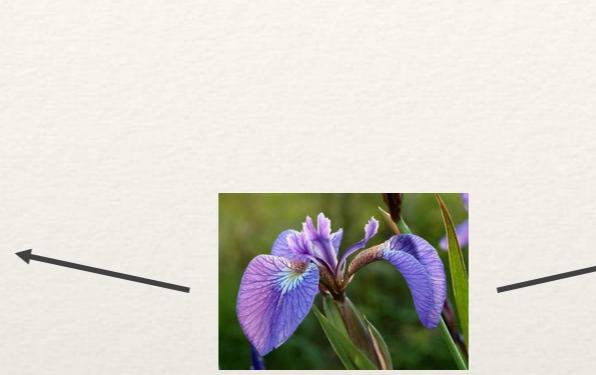
מהם ?test set ו-training-set משמשים בהם? מה הקשר לקטgorיות,  
ול-feature vectors וול-feature set

# KNN - שאלות עליהם נרצה לענות

- ❖ מהם סוגים המחלקות, ומה הקשר שלהם לבעיית הסיווג?
- ❖ מהם סוגים המאפיינים ומה הקשר ביניהם ובין ה-feature vector?
- ❖ מהו test-set ומהו train-set?
- ❖ את מה אנחנו מכנים שכנים?
- ❖ מהו הוקטור?
- ❖ איך מציגים דוגמה כוקטור?
- ❖ מה הופך את השכנים לקרובים? איך מודדים קירבה?
- ❖ איך בסוף מקבלים את ההכרעה?

# מודל הסיווג הראשון – KNN

## K- Nearest Neighbors



**מה מגדיר אותם כשכנים?**

- בהנתן דוגמה חדשה, נקבע את ערכי המאפיינים מה-set feature- $\vec{x}_j$  וניתור את הווקטור  $\vec{x}$
- שכן – וקטור  $\vec{x}_j$  השיך לדוגמאות האימון, הנמצא בשכנות ל-  $\vec{x}$



**הדוגמאות**

- כל דוגמה מיוצגת ע"י וקטור  $\vec{x}$
- וקטור מכיל ערכים מספריים של המאפיינים
- לכל דוגמא מושגת ע"י וקטור  $\vec{x}$
- לא כל דוגמא מושגת ע"י וקטור  $\vec{x}$

אבל מהו בעצם אותו וקטור?

---

---

**מושגים מתמטיים -  
וקטוריים וסקלריים**

# מושגים – סקלר

סקלר – איבר בשדה, במקרה שלנו מספר למרחב  $\mathbb{R}$

משתנה (variable) – האות  $a, b, \lambda \in \mathbb{R}$  (האות למדה)

למשל:  $a = 5.5, b = -\sqrt{2}, \lambda = 3$

$$f: A \subset \mathbb{R} \rightarrow \mathbb{R}$$

פונקציות – פועלות על סקלרים, התוצאה היא סקלר באותו שדה

למשל:

$$3+4=7$$

$$\sqrt[3]{8} = 2$$

$$3 \cdot 1.5 = 4.5$$

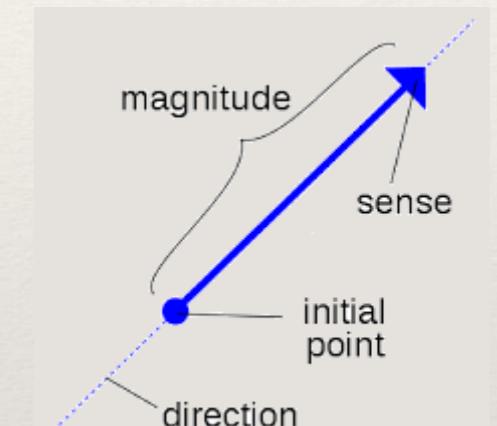
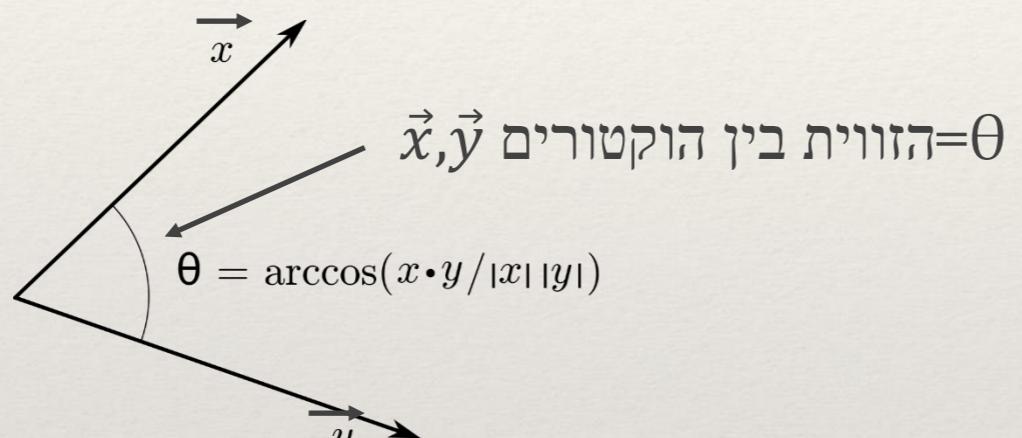
$$|-5.5| = 5.5$$

$$-1 / 10 = -0.1$$

# מושגים – וקטור (פרשפקטיבית גאומטרית)

וקטור (vector) - נסמן  $\vec{x}, \vec{y} \in \mathbb{R}^n$

❖ הסבר פיזיקלי/גאומטרי: ישות מתמטית בעלת גודל וכיון



נורמה של וקטור - הכללה של מושג ה"אורך" (magnitude) – נסמן  $\|\vec{x}\|$

מכפלה סקלרית (dot product) של וקטורים – נסמן  $\vec{y} \cdot \vec{x}$

$$\vec{x} \cdot \vec{y} = \|\vec{x}\| \cdot \|\vec{y}\| \cdot \cos \theta$$

❖ לכן, מכפלה סקלרית של שני וקטורים היא 0 או אין הם ניצבים, כיוון ש- $\cos 90^\circ = 0$

# מושגים – וקטור (פרספקטיבה אלגברית)

וקטור (vector) - נסמן  $\vec{x}, \vec{y} \in \mathbb{R}^n$

❖ מבחן אלגברית:  $(\vec{x}, \vec{y}) = (x_1, x_2, \dots, x_n), (\vec{y}, \vec{y}) = (y_1, y_2, \dots, y_n)$

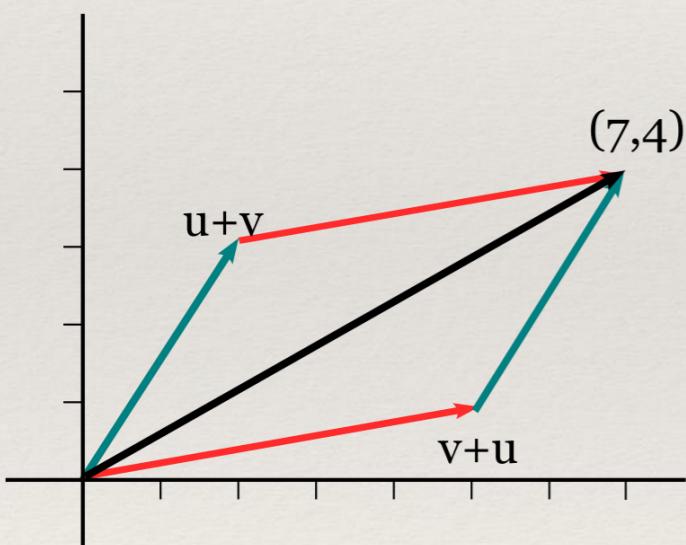
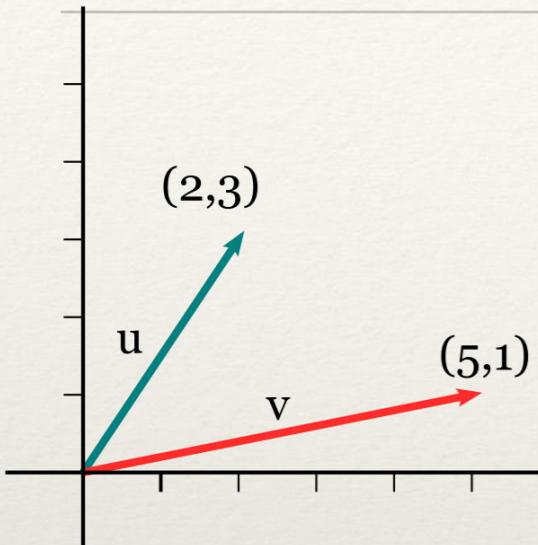
כפל (של וקטור) בסקלר -  $\vec{x} \cdot \lambda = \lambda \cdot \vec{x} = (\lambda \cdot x_1, \lambda \cdot x_2, \dots, \lambda \cdot x_n)$

מכפלה סקלרית (dot product) של וקטורים – נסמן  $\vec{x} \cdot \vec{y}$

מבחן אלגברית -  $\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^n x_i \cdot y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + \dots + x_n \cdot y_n$

❖ בצורה מטריציונית -  $\langle \vec{x}, \vec{y} \rangle = \vec{x} \cdot \vec{y}^T = \sum_{i=1}^n x_i \cdot y_i$

# חיבור וקטורי



חיבור של וקטורים – נסמן  $\vec{x} + \vec{y}$

מבחן אלגברית  $(x_1+y_1, x_2+y_2, \dots, x_n+y_n)$

תרגיל – בדוגמה שבתמונה  $\vec{u} + \vec{v}$

❖ תשובה –  $\vec{u} + \vec{v} = (2,3) + (5,1) = (7,4)$

נסמן  $\vec{u} - \vec{v}$  – מתקיים  $\vec{z} - \vec{u} = \vec{z} + -\vec{u} = \vec{v}$ ,  $\vec{z} = \vec{u} + \vec{v}$

אבל מהו הקשר בין  
הוקטורים לדוגמאות?  
(נראה בהמשך)

# שאלות ביניים

שאלה 1:

?test-set training-set מי? מה הקשר בין ה- מיהם השכנים?

שאלה 2:

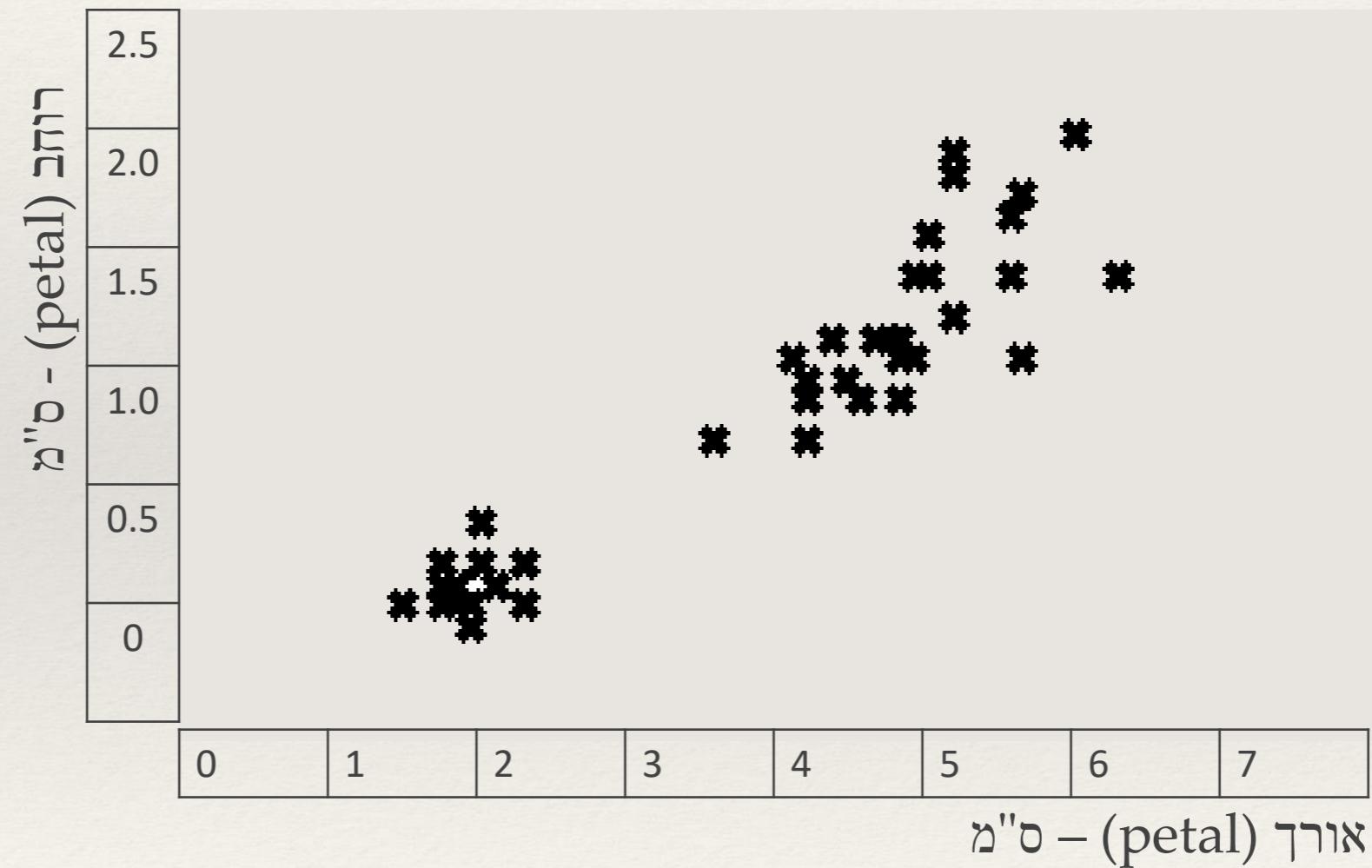
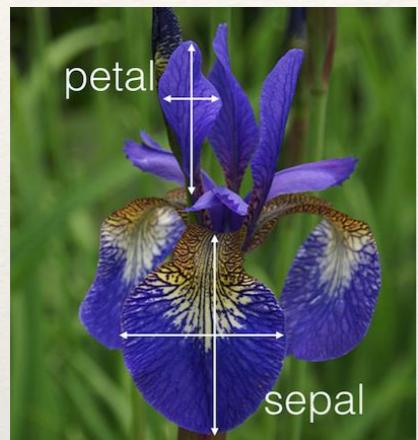
מהו סקלר? מהו וקטור בתצורה גאומטרית? מהו וקטור בתצורה אלגברית?

# KNN - שאלות עליהם נרצה לענות

- ❖ מהם סוגים המחלקות, ומה הקשר שלהם לבעיית הסיווג?
- ❖ מהם סוגים המאפיינים ומה הקשר ביניהם ובין ה-feature vector?
- ❖ מהו test-set ומהו train-set?
- ❖ את מה אנחנו מכנים שכנים?
- ❖ מהו הוקטור?
- ❖ איך מציגים דוגמה כוקטור?
- ❖ מה הופך את השכנים לקרובים? איך מודדים קירבה?
- ❖ זוכרים את פונקציית הסילום? מה הקשר ל-KNN?
- ❖ איך בסוף מקבלים את ההכרעה?

# מודל הסיווג הראשון - KNN

נחזיר לסוגי האירוסים:

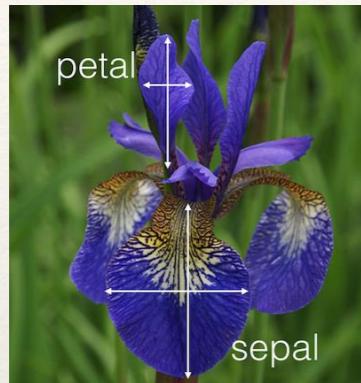


---

**אמור לי מי חבריך ואומר לך מי אתה** (מיגל דה סרוואנטס)

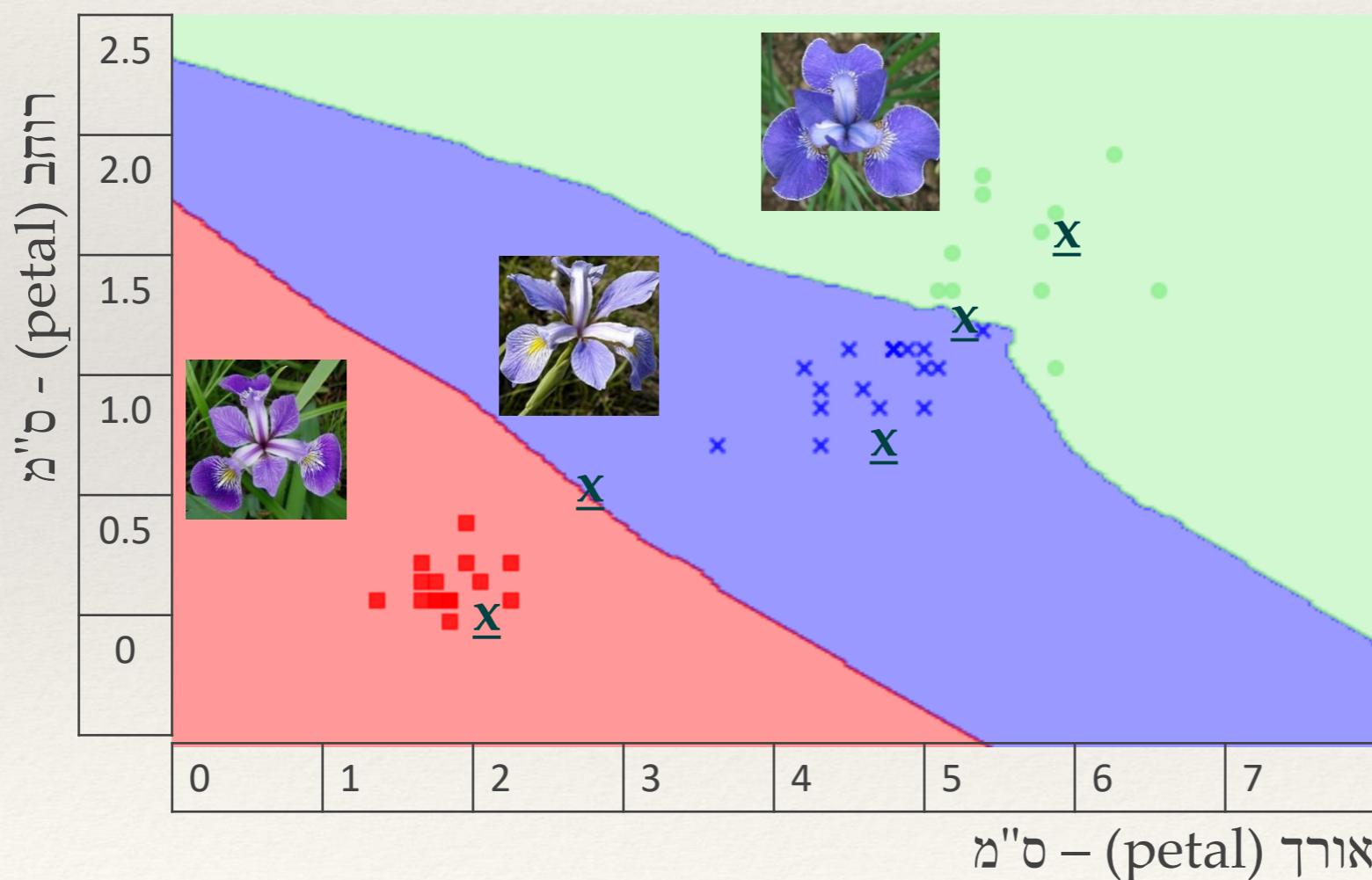


# מודל הסיווג הראשון - KNN



המטרה – להחליט מהו סוג האירוס, ע"י ערכי המאפיינים

- וקטורים עם 2 מאפיינים: אורך ורוחב של עלי גביע (petal)
- הרעיון – נחליט מהי הקטגוריה של הדוגמה החדשה בהתאם לשכנים "הקרובים"



- מקרים פשוטים יותר
- מקרים מורכבים יותר

---

# פונקציות מרחק והקשר ל-KNN

# פונקציות מרחק - הגדמה

- ❖ נסמן ב-  $\vec{x}_j$  את הוקטור החדש אותו נרצה לסוג train set
  - ❖ נסמן ב-  $\vec{x}_i$  אחד הוקטוריים השبيיכים לו
  - ❖ נסמן ב-  $\text{dist}(\vec{x}_j, \vec{x}_i)$  את פונקציית המרחק בין ב-  $\vec{x}_j$  ל-  $\vec{x}_i$
- הגדרה מויקיפדיה (של פונקציית מרחק):

*Distance metric uses distance function which provides a relationship metric between each elements in the dataset.*

- ❖ נוכל להגיד את השם הקרוב ביותר, ככל שהוא רחוק, בעזרת הפונקציה  $\text{dist}(\vec{x}_j, \vec{x}_i)$ :

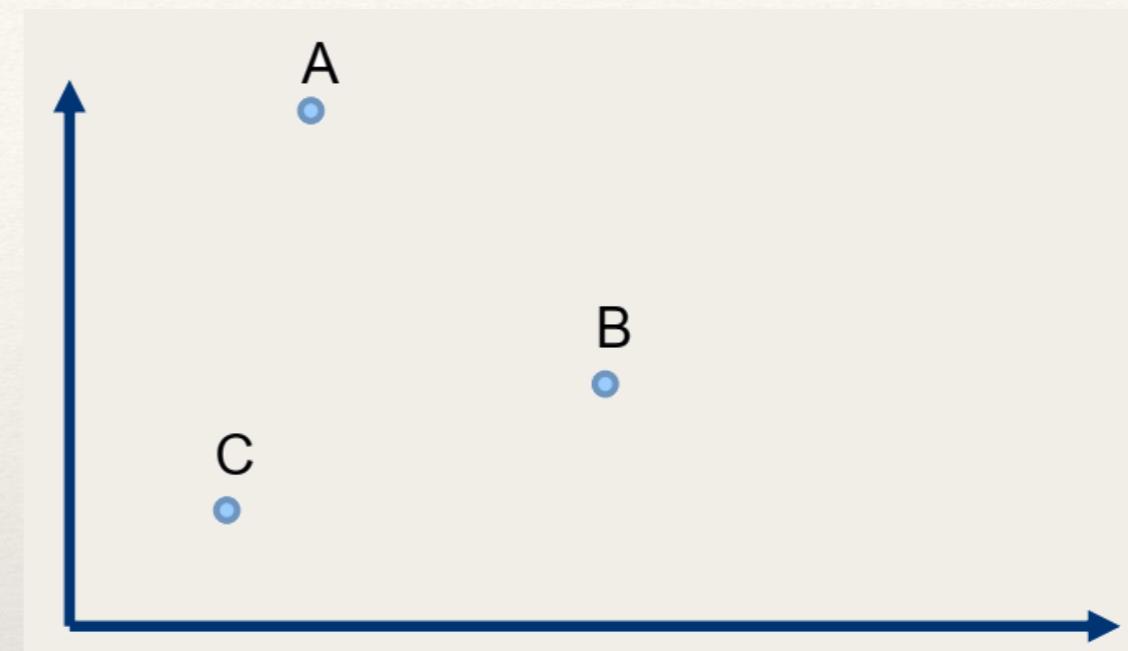
$$1\text{-NN} = \arg\min_i \text{dist}(\vec{x}_j, \vec{x}_i)$$

- ❖ כיצד נчисב את הפונקציה  $\text{dist}(\vec{x}_j, \vec{x}_i)$ ? (נראה עוד מעט)

# פונקציית מרחק - תכונות

Metric distances:

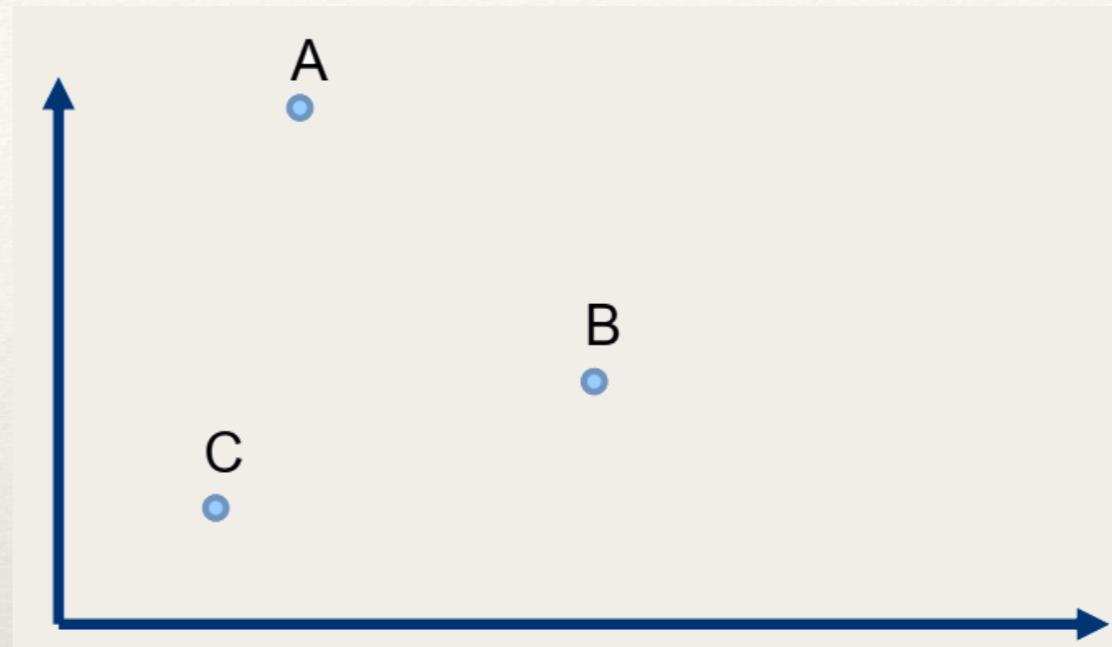
1.  $d_{ab} \geq 0$  (מרחק תמיד אי שלילי)
2.  $d_{ab} = d_{ba}$  (המרחק מקיים סימטריות)
3.  $d_{aa} = 0$  (מרחק עצמי)
4.  $d_{ab} \leq d_{ac} + d_{cb}$  (אי שיוויון המשולש)



# פונקציית מרחק Minkowski Distance

## Minkowski Distance:

- ❖ Formula:  
$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$
- ❖ Minkowski distance is the generalized distance metric.
- ❖ It means that we can manipulate the formula to calculate the distance between two data points in different ways



# פונקציית מרחק Minkowski Distance

## Minkowski Distance:

- ❖ Formula:

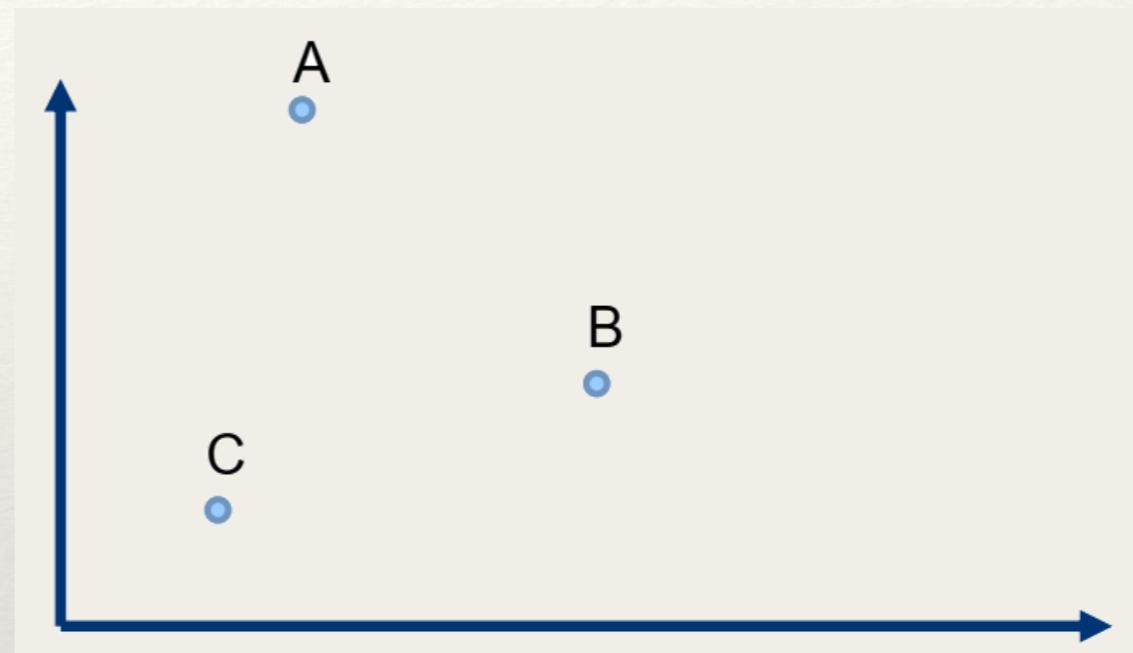
$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

Manipulating the value of  $p$ , assists in calculating three different methods:

$p = 1$ , Manhattan Distance

$p = 2$ , Euclidean Distance

$p = \infty$ , Chebyshev Distance



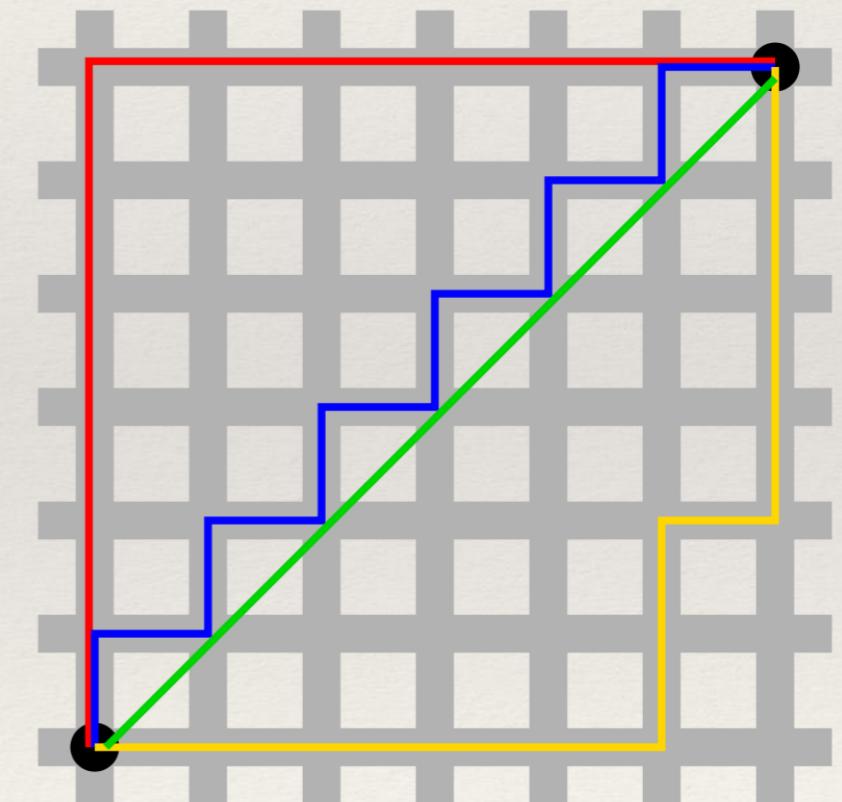
# פונקציית מרחק Manhattan



# פונקציית מרחק Manhattan Distance - מרכז

## Manhattan Distance:

The sum of the horizontal and vertical distances between points on a grid  
(also called Taxicab Geometry)



# פונקציית מרחק Manhattan Distance - מינקובסקי

Minkowski Distance:

- ❖ Formula:

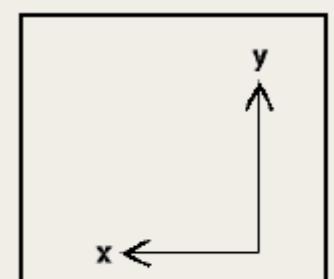
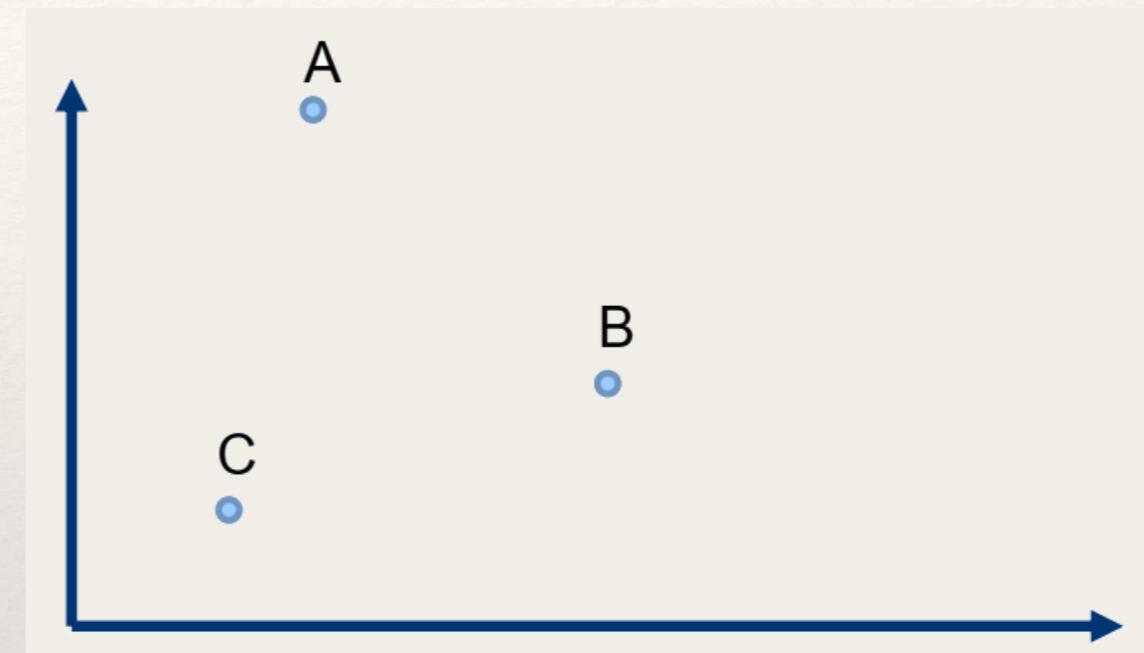
$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

Manhattan Distance: we use

Minkowski distance with  $p=1$

$$\rightarrow d = \sum_{i=1}^n |x_i - y_i|$$

$$= |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$



Manhattan

# מרחק מנהטן - תרגיל

לפי פונקציית מרחוק מנהטן, מהו המרחק בין זוגות הוקטורים הבאים?

$$v1=(1,0,1) \quad v2=(1,0,4) \quad v3=(2,1,5)$$

$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

תשובה:

א.  $3 = \text{ה מרחק} - v1, v2$ .

ב.  $3 = \text{ה מרחק} - v2, v3$ .

ג.  $6 = \text{ה מרחק} - v1, v3$ .

# תרגיל בית – Manhattan Distance

חשבו מרחק מנהאטן בין 2 שורות בטבלה מסוימת

Manhattan Distance:

$$d = \sum_{i=1}^n |x_i - y_i|$$

$$= |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$

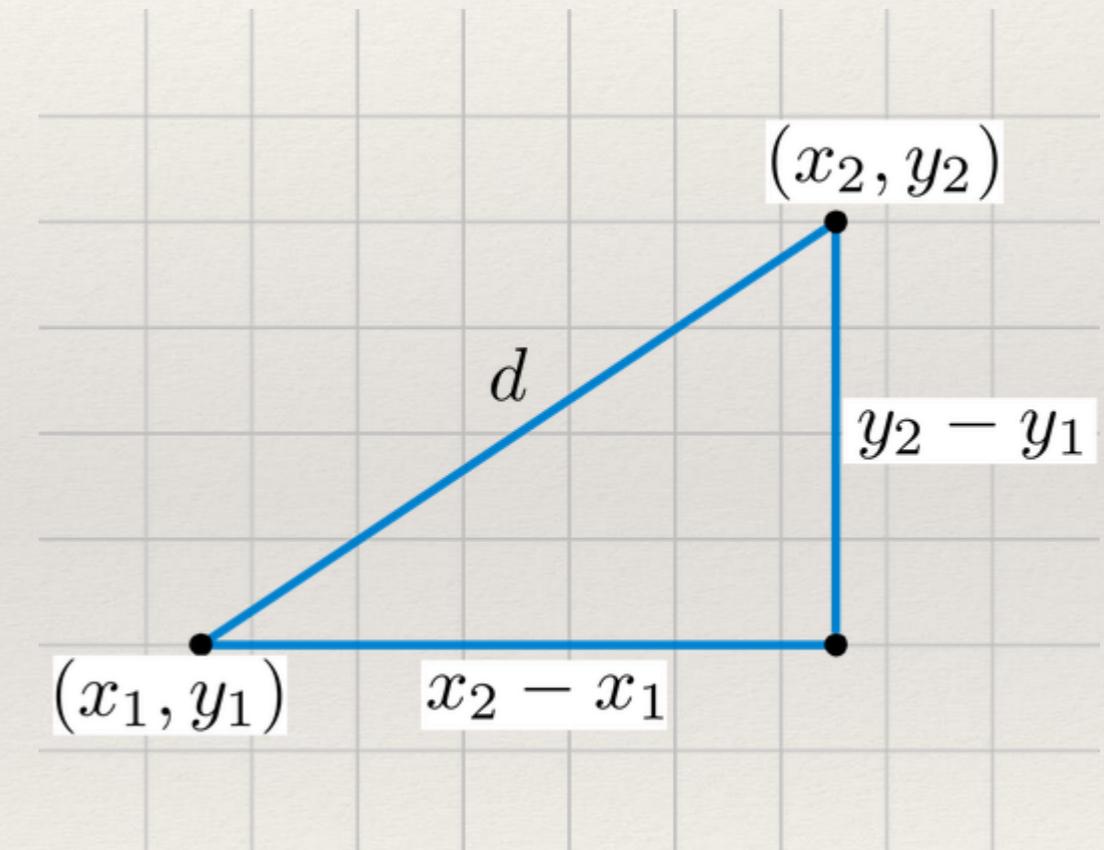
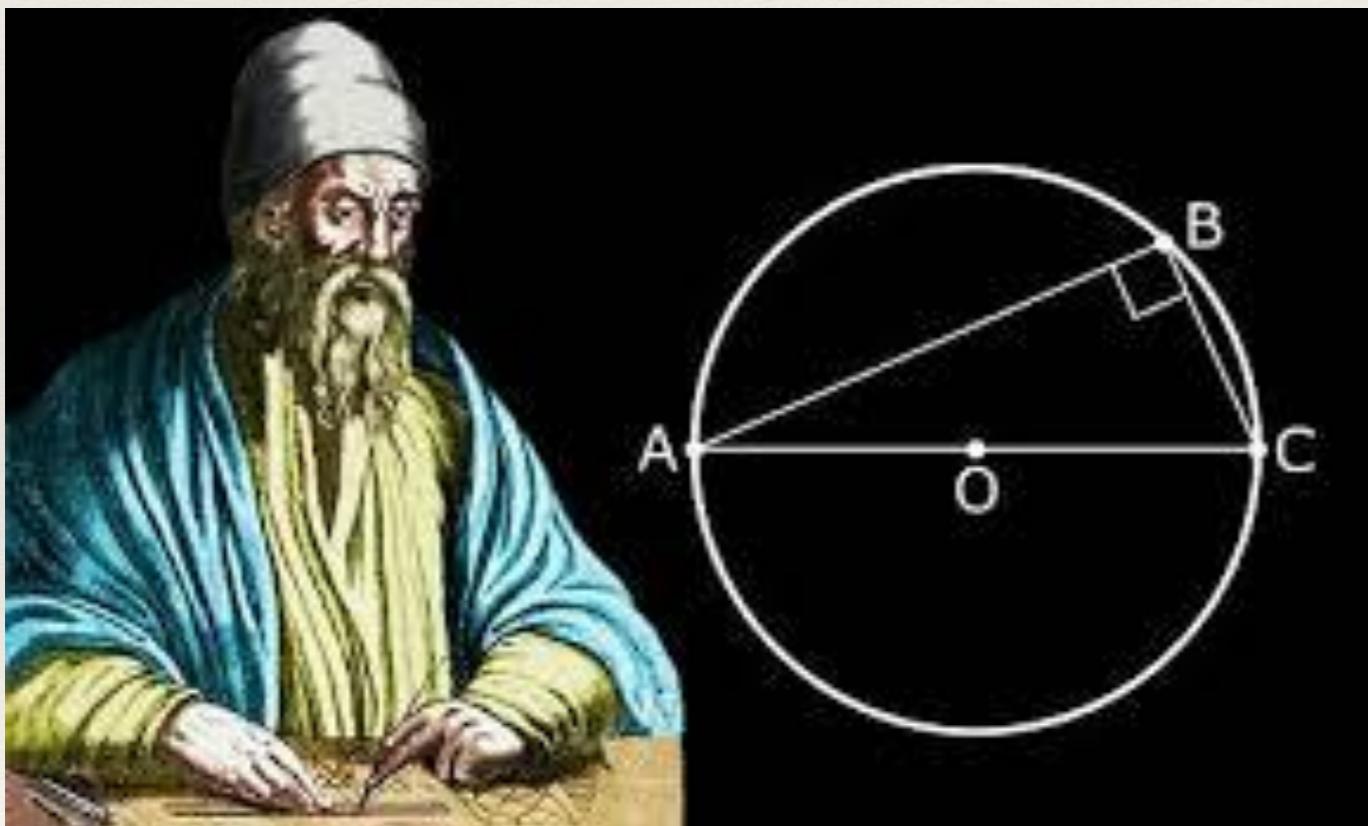
sepal		petal	
length	Width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

is_animal	vertical_stripes	black_&_white	4_legs	large
0	1		1	0
1	0		1	0
1	0		0	1
1	1		1	1

# פונקציות מרחק - פונקציית מרחק

## Euclidean Distance:

Used in Euclidean Geometry



# פונקציות מרחק - Euclidean Distance

Minkowski Distance:

- ❖ Formula:

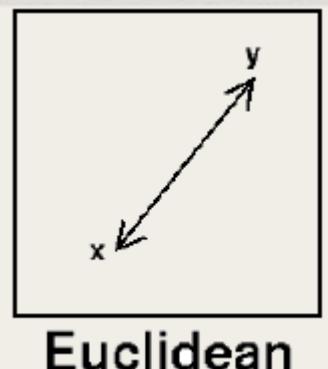
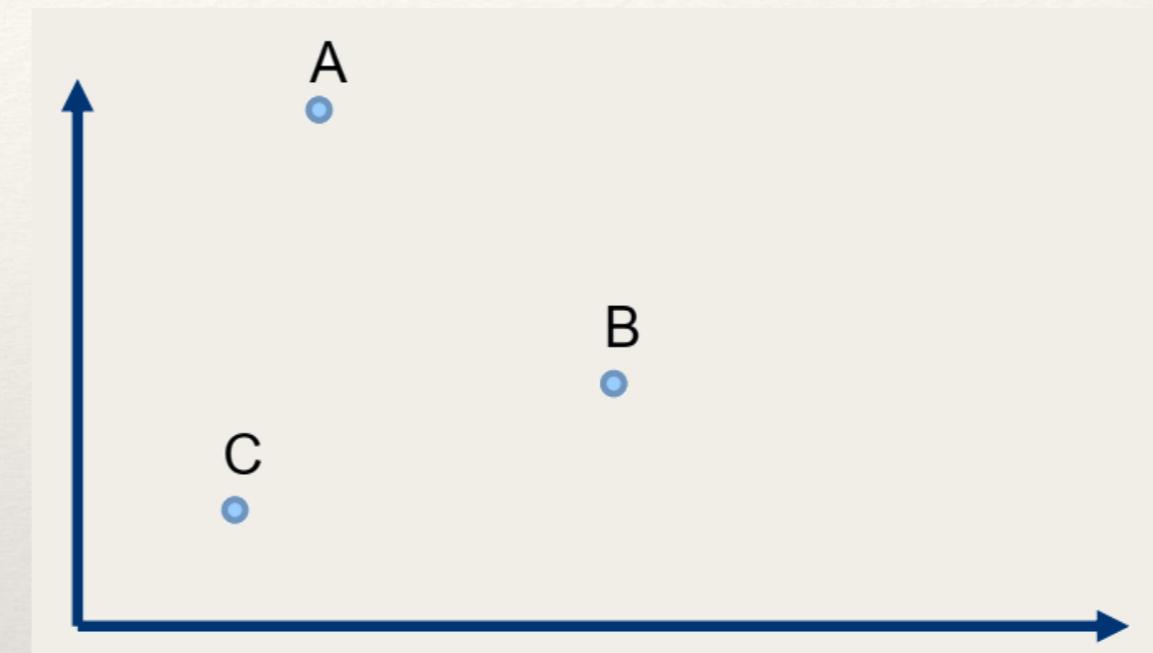
$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

Euclidean Distance: we use

Minkowski distance with  $p=2$



$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$



# מרחק אוקלידי - תרגיל

לפי פונקציית מרחוק אוקלידי, מהו המרחק בין זוגות הוקטורים הבאים?

$$v1=(1,0,1) \quad v2=(1,0,4) \quad v3=(2,1,5)$$

$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

תשובה:

א.  $\sqrt{3} = \text{ה מרחק } v1, v2$

ב.  $\sqrt{1.73} = \text{ה מרחק } v2, v3$

ג.  $\sqrt{4.24} = \text{ה מרחק } v1, v3$

# תרגיל בית – Euclidean Distance

חשבו מרחק אוקלידי בין 2 שורות בטבלה מסוימת

Euclidean Distance:

$$\begin{aligned}d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) &= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\&= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}.\end{aligned}$$

sepal		petal	
length	Width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

is_animal	vertical_stripes	black_&_white	4_legs	large
0	1		1	0
1	0		1	0
1	0		0	1
1	1		1	1

# פונקציות מרחק - Chebyshev Distance

## Chebyshev Distance:

Also called *chessboard distance*



	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	1	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1

# מרחק צ'בישב - תרגיל

לפי פונקציה מרחק צ'בישב, מהו המרחק בין זוגות הוקטורים הבאים?

$$v1=(1,0,1) \quad v2=(1,0,4) \quad v3=(2,1,5)$$

$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

תשובה:

.א.  $3 = \text{המרחק} - v1, v2$

.ב.  $1 = \text{המרחק} - v2, v3$

.ג.  $4 = \text{המרחק} - v1, v3$

# תרגיל בית - Chebyshev Distance

חשבו מרחק צ'בישב בין 2 שורות בטבלה מסוימת

Chebyshev Distance:  $d(\vec{x}_j, \vec{x}_i) = \max_{1 \leq m \leq d} |x_{jm} - x_{im}|$

sepal		petal	
length	Width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

is_animal	vertical_stripes	black_&_white	4_legs	large
0	1	1	0	0
1	0	1	0	0
1	0	0	1	1
1	1	1	1	1

# שאלות ביניים

שאלה 1: מה הקשר בין מרחק, קרבה ושכנות? מה הקשר לקביעת הקטגוריה של הדוגמה החדשה (בנייה בדוגמה של השכן הקרוב ביותר)?

שאלה 2: האם יכול להיות שמרחב צ'בישוב גדול מרחק מנהטו?

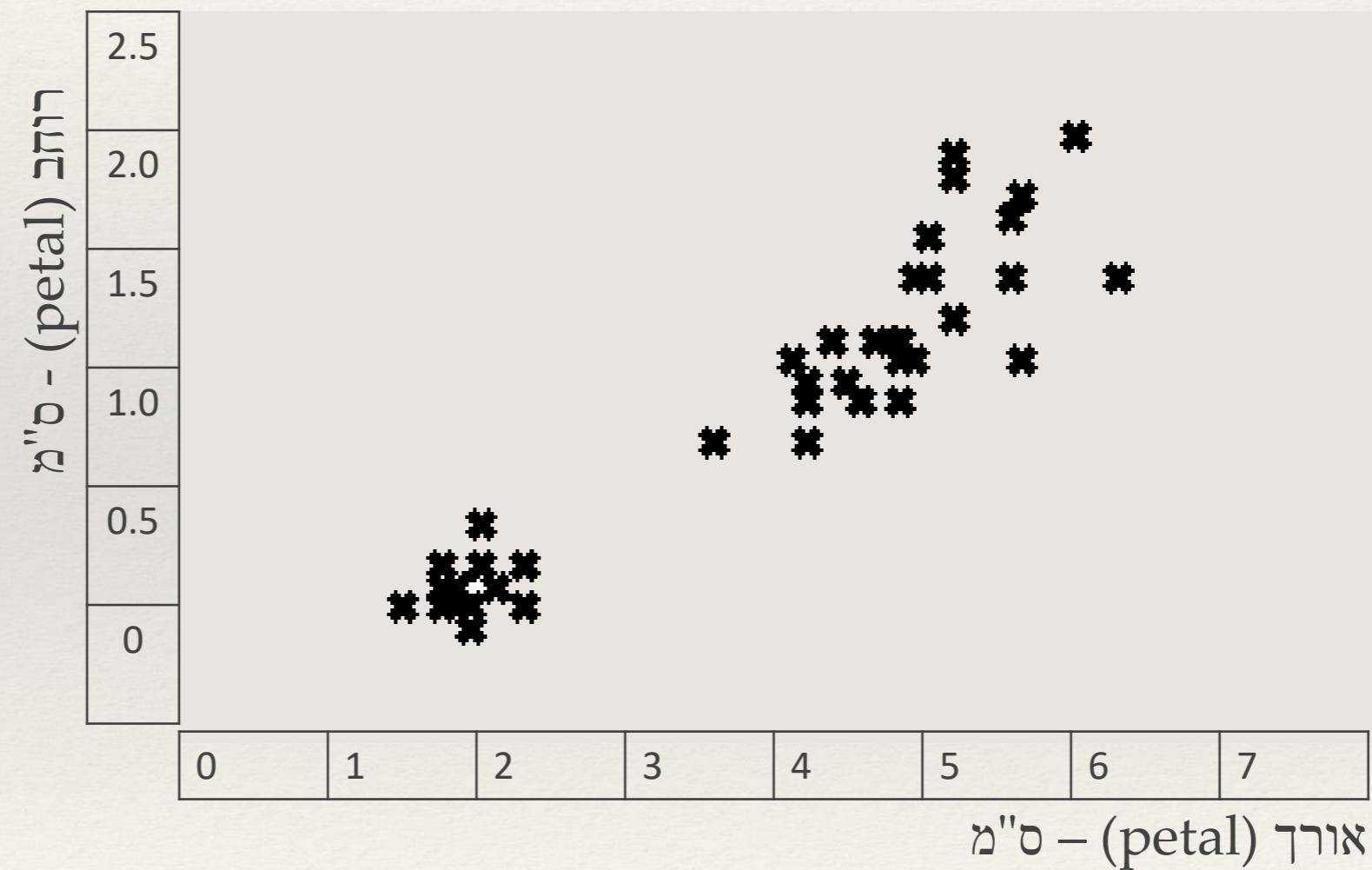
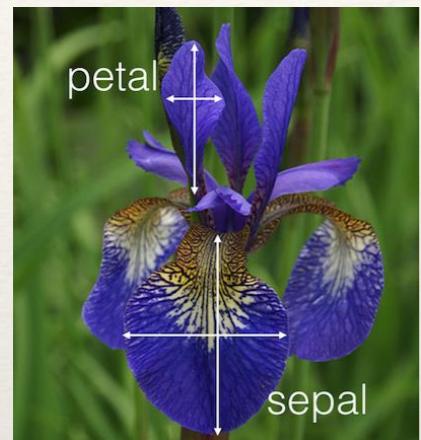
שאלה 3: מה הקשר אוקלידי למרחק בין נקודות למרחב?

# KNN - שאלות עליהם נרצה לענות

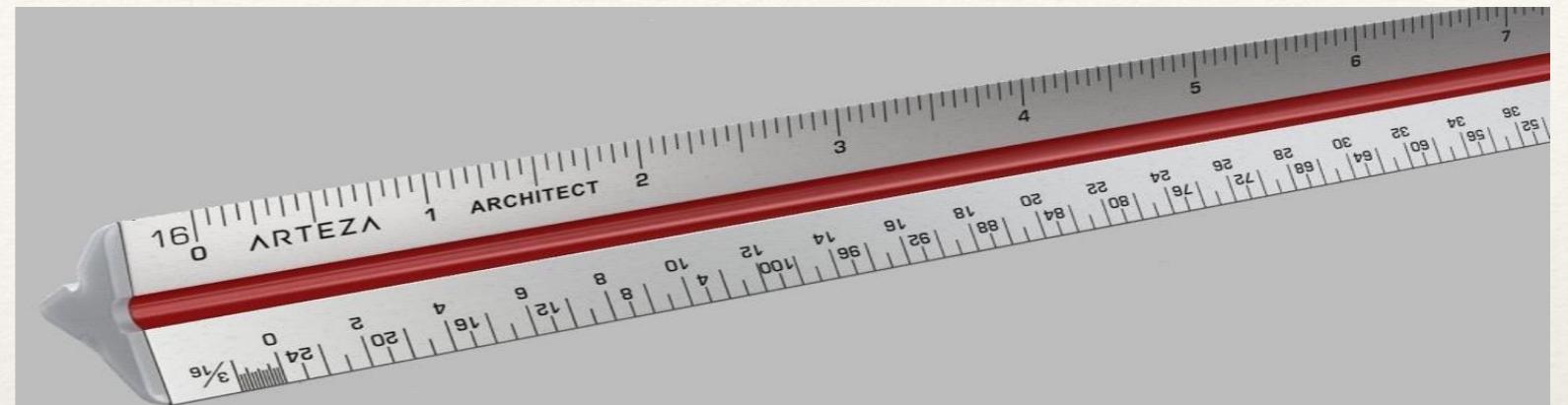
- ❖ מהם סוגים המחלקות, ומה הקשר שלהם לבעיית הסיווג?
- ❖ מהם סוגים המאפיינים ומה הקשר ביניהם ובין ה-feature vector?
- ❖ מהו test-set ומהו train-set?
- ❖ את מה אנחנו מכנים שכנים?
- ❖ מהו הוקטור?
- ❖ איך מציגים דוגמה כוקטור?
- ❖ מה הופך את השכנים לקרים? איך מודדים קירבה?
- ❖ זוכרים את פונקציית הסילום? מה הקשר ל-KNN?
- ❖ איך בסוף מקבלים את ההכרעה?

# תצוגה וקטוריית של הדוגמאות - חזרה

נחזור לסוגי האירוסים:



# סילום (Scaling) של מאפיינים - תזכורת



**סילום (Scaling):**

**סילום מאפיינים** – הוא שיטה המשמשת לנורמליזציה של טווח המאפיינים.

המטרה: סילום מחדש, כמו במעבר מאינץ' לס"מ

standardization – הופכים את הממוצע החדש ל-0, וסטיית התקן, הופכת ל-1

- משתמשים בהתקלגות t

minmax normalization – הסילום מתבצע כך שערך המינימום והמקסימום החדש, הינם 0 ו-1 בהתאם.

מה הקשר בין סילום ל-KNN?

# KNN וסילום (Scaling) - מוטיבציה



## מוטיבציה –

- ❖ למאפיינים שונים פונקציית התפלגות שונה
- ❖ KNN לא מניה איזשאם הנחות על התפלגות הנתונים
- ❖ סולם (scale) שונה עלול להוביל לעיוות המרחק – מדוע?
- ❖ סולם (scale) שונה עלול לתת משקל שונה למאפיינים שונים, רק בגלל הסולם השונה (במקרה של KNN, זה די דומה) – מדוע?

# (t-distribution) standardization- ו-KNN

## מוטיבציה –

- ❖ למאפיינים שונים פונקציית התפלגות שונה
- ❖ KNN לא מניח איזושם הנחות על התפלגות הנתונים

## פתרון ע"י שיטת הסילום (t-distribution) standardization:

- ❖ בסטטיסטיקה – כל התפלגות ניתנת להפוך לתפלגות  $t$ , אם ידועות הממוצע וסטיית התקן במדגם (התפלגות  $t$ , היא קירוב לתפלגות  $z$  שהינה סוג של התפלגות נורמלית)
- ❖ במקרה שלנו – המדגם הוא ה- train set
- ❖ ב-KNN, מהויה תרומה למידים השונים של פונקציית המרחק

# minmax normalization - KNN

מotiveציה –

- ❖ סולם (scale) שונה עלול להוביל לעיוות המרחק – מדוע?
- ❖ סולם (scale) שונה עלול לחת משקל שונה למאפיינים שונים, רק בגלל הסולם השונה (במקרה של KNN, זה די דומה) – מדוע?

פתרון ע"י שיטה הסילום :minmax normalization

- ❖ השוואת פשטה של הסולם, ע"י קביעת סולם בטוח אחיד.
- ❖ מכונה גם נרמול מינימום ומקסימום.

מקובלים למשל הטוחחים:

- ❖  $[0,1]$  – כפי שהיא לכם בתרגיל –  $\text{Min}$  החדש הופך ל-0, והמקסימום ל-1
- ❖  $[-1,1]$  – לעיתים מסיע, בדומה למרחק cosine

---

---

**KNN – איך קיבל החלטה לגבי דוגמה חדשה?**

# – אלגוריתם השמירה – KNN

## Input:

- ❖ k – the number nearest neighbors; the set of training examples.

## The KNN Algorithm:

- ❖ for test instance  $x_j$  in the test-set:
  - ❖ Calculate  $d(x_j, x_i)$
  - ❖ Select the k closest training examples,  $d(x_j, x_i)$  sorted
  - ❖ Use majority voting to classify the test examples

## Notations and Terms:

$x_j$  – example (number j) from the test-set  
 $x_i$  – example (number i) from the train-set  
 $d(x_j, x_i)$  – distance function – measures distance between  $x_j$  and  $x_i$ .

# KNN – תרגיל

סימוניים:

נסמן וקטור (feature vector) עבור ערכי שני מאפיינים  $X_1, X_2$  כך:  $(x_1, x_2)$   
וקטוריים, עבורם ידועה הקטגוריה שלהם  $c$ , נסמן כך:  $(x_1, x_2 | c)$

נתונים הווקטוריים הבאים:

$(0,0|1), (1,0|1), (1,1|-1), (4,2|1), (3,5|-1), (1,4|1), (3,1|-1)$

מצאו באמצעות אלגוריתם KNN את הסיווג של הווקטור  $(3,2)$

הערות:

- ❖ לפונקציית מרחק, השתמשו בשיטת מרחק אוקלידית
- ❖ דלו כרגע על שלב הסילום
- ❖ בchnerו את הפתרון עבור  $k=3,7$

# KNN – תרגיל

קודם כל, נחשב את המרחקים (לפי הוראות התרגיל משתמשים בפונקציה מרחק אוקלידי)

ווקטור	סיג	מרחק מ-(3,2)	
(0,0)	1	$\sqrt{(3-0)^2 + (2-0)^2} = \sqrt{13}$	3.6
(1,0)	1	$\sqrt{(3-1)^2 + (2-0)^2} = \sqrt{8}$	2.8
(1,1)	-1	$\sqrt{(3-1)^2 + (2-1)^2} = \sqrt{5}$	2.2
(4,2)	1	$\sqrt{(3-4)^2 + (2-2)^2} = \sqrt{1}$	1
(3,5)	-1	$\sqrt{(3-3)^2 + (2-5)^2} = \sqrt{9}$	3
(1,4)	1	$\sqrt{(3-1)^2 + (2-4)^2} = \sqrt{8}$	2.8
(3,1)	-1	$\sqrt{(3-3)^2 + (2-1)^2} = \sqrt{1}$	1

איזו קטגוריה נבחר עבור  $k=3$ ? איזו קטגוריה נבחר עבור  $k=7$ ?

- ❖ עבור –  $k=3$  - שני שכנים מצבעים 1-, ושכן אחד מצבע 1, לפי הרוב – נסוג -1
- ❖ עבור –  $k=7$  - הצבעת הרוב – נסוג 1

# KNN – בחרת הקטגוריה בזמן סיג כיצד נבחר את הקטגוריה עבור דוגמה חדשה?

- ❖ **1-NN** - Given a new point  $x$ , we wish to find it's nearest point and return it's classification.
- ❖ **K-NN** - Given a new point  $x$ , we wish to find it's k nearest points and return their average classification.
- ❖ **Weighted** - Given a new point  $x$  , we assign weights to all the sample points according to the distance from  $x$  and classify  $x$  according to the weighted average.

# משמעות של ערכי K שונים

- ❖ המשמעות של ערכי K מאוד קטנים, עלולה להיות החלטה המושפעת מאוד מרעש (מדוע?)
- ❖ המשמעות של ערכי K מאוד גדולים, משמעה, עליה משמעותית בסיבוכיות של אלגוריתם KNN בו מרכז הקובד הוא בזמן אמת / בזמן הבדיקה.
- ❖ שאלה: מה יקרה אם K שואף ל- $\infty$  (כאשר  $\infty$  מסמנת את מספר הדוגמאות באימון)?
- ❖ תשובה: התשובה בעצם תשאף להתפלגות הקטגוריות ב-training-set (מדוע?)

# משמעות של ערכי K שונים (המשר)

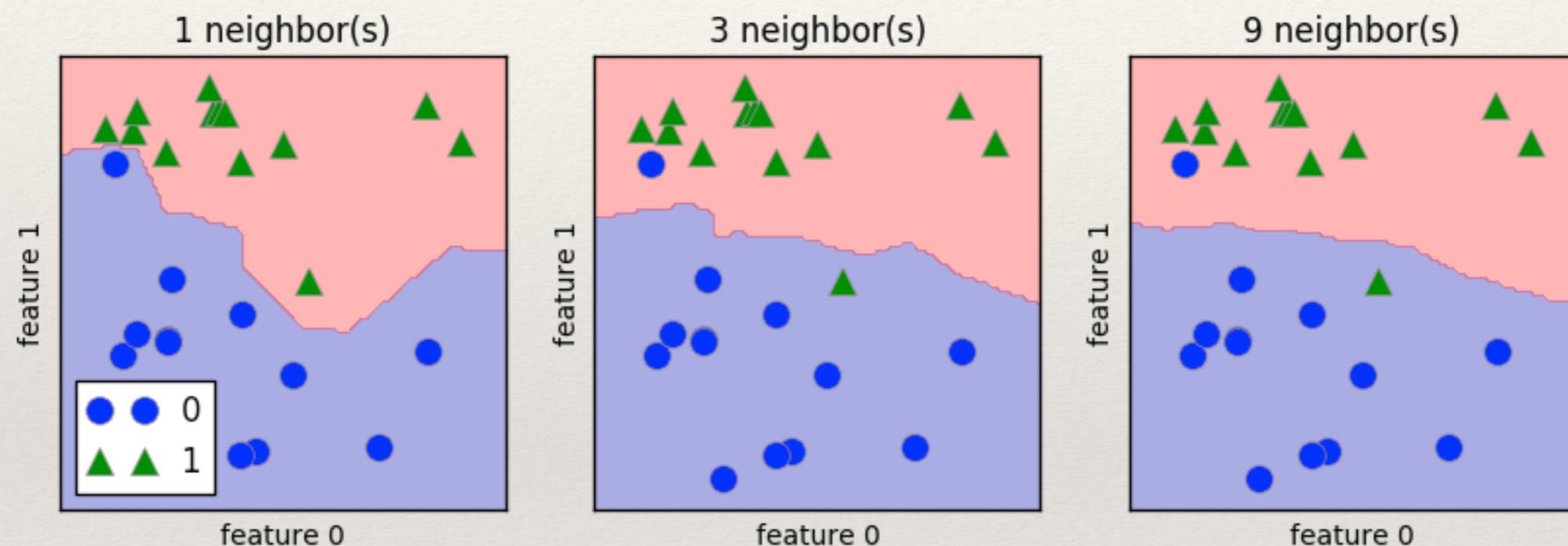
- ❖ אם מספר המחלקות הוא 2, נשאף למספר K אי זוגי (מדוע?)
- ❖ באופן דומה ננסה לבחור את K באופן שיעזר להכריע ו לבחור את המחלקה מבין השכנים הקרובים.
- ❖ אחד מכללי האצבע הוא לבחור  $O(\sqrt{n})$ , אבל זה תלוי בבעיה.

## שיפורים נוספים:

- ❖ בהמשך הקורס, נלמד דרך לבחור את ערכו של K
- ❖ בחירת פונקציית המרחק (ברירת המחדל – היא פונקציה מרחק אוקלידית) המיטבית לבעיה
- ❖ הכרעה, כאשר יש שיוויון בין המחלקות

# משמעות של ערכי K שונים (המשך)

דוגמה להשפעה של ערכי K שונים:



שאלה: האם לא נעדיף את הגרף עבור  $K=1$ , הרי נראה שהוא יותר מתאים לנ נתונים?

תשובה: יש חשש להטמת יתר ל-trainig (overfitting), וריגישות גבוהה מדי לרעיש (נרחיב על התאמת יתר עוד במהלך הקורס).

# שאלות ביניים

שאלה 1:

כיצד סילום ע"י  $t$ -distribution) standardization מסיע ל-KNN?

שאלה 2:

כיצד סילום ע"י minmax normalization מסיע ל-KNN?

שאלה 3:

כיצד קיבל החלטה לגבי הקטגוריה של דוגמה חדשה ע"י KNN? מה פירוש NN-1 בעצם? מה ההבדל בין בחירה לפי הרוב, לבין בחירה ממושקלת?

שאלה 4:

מה המשמעות של ערכי  $K$  שונים (קטנים וגדולים)? מדוע נשאף לבחור  $K$  אי זוגי אם מספר הקטגוריות הוא 2?

## KNN – תכונות

- ❖ KNN הינו אלגוריתם עצמן - עיבוד הנתונים מתרבצע רק עם קבלת נקודה חדשה לסיווג
- ❖ דוחים את רוב העבודה לזמן הסיווג
- ❖ KNN לא מניח איזיהם הנחות על התפלגות הנתונים
- ❖ KNN עובד מידית גם על "ריבוי מחלקות" (בדון על ריבוי מחלקות עוד במהלך הקורס)

# KNN – סיכון השיטה

- .1. הגדרת הבעיה כבעיית סיווג ו-modeling
  - .2. איסוף דוגמאות, vectorization
  - .3. חלוקה ל-(validation-) train-test cleansing
  - .4. ניקוי ה-data, וסילום הערכאים האפשריים.
- ❖ – אלגוריתם עצמן – לא עושה (כמעט) כלום בשלב זה
  - ❖ שיעורוד (משתמש בסיווג) – בהמשך
  - ❖ סיווג (דוגמאות לא ידועות)

נתראה בשבוע הבא ☺