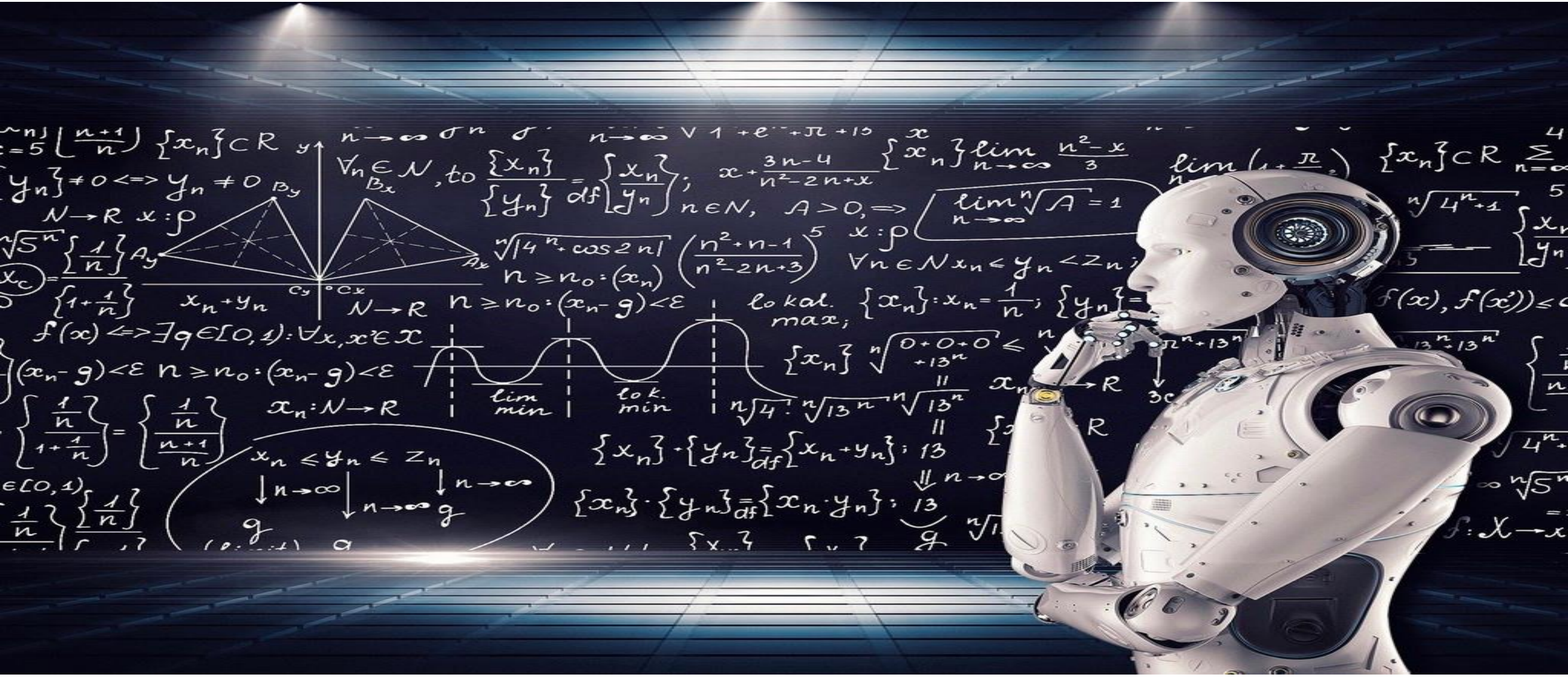
Representation Learning



Agenda

- K-Means
  - Reminder
  - Example exam question (last year moed A)
- PCA & LDA
  - PCA – reminder
  - LDA
    - Principles
    - Numeric Example
- Performance comparison (notebook) - Vanilla, LDA, PCA

Agenda

- K-Means
  - Reminder
  - Example exam question (last year moed A)
- PCA & LDA
  - PCA – reminder
  - LDA
    - Principles
    - Numeric Example
- Performance comparison (notebook) - Vanilla, LDA, PCA

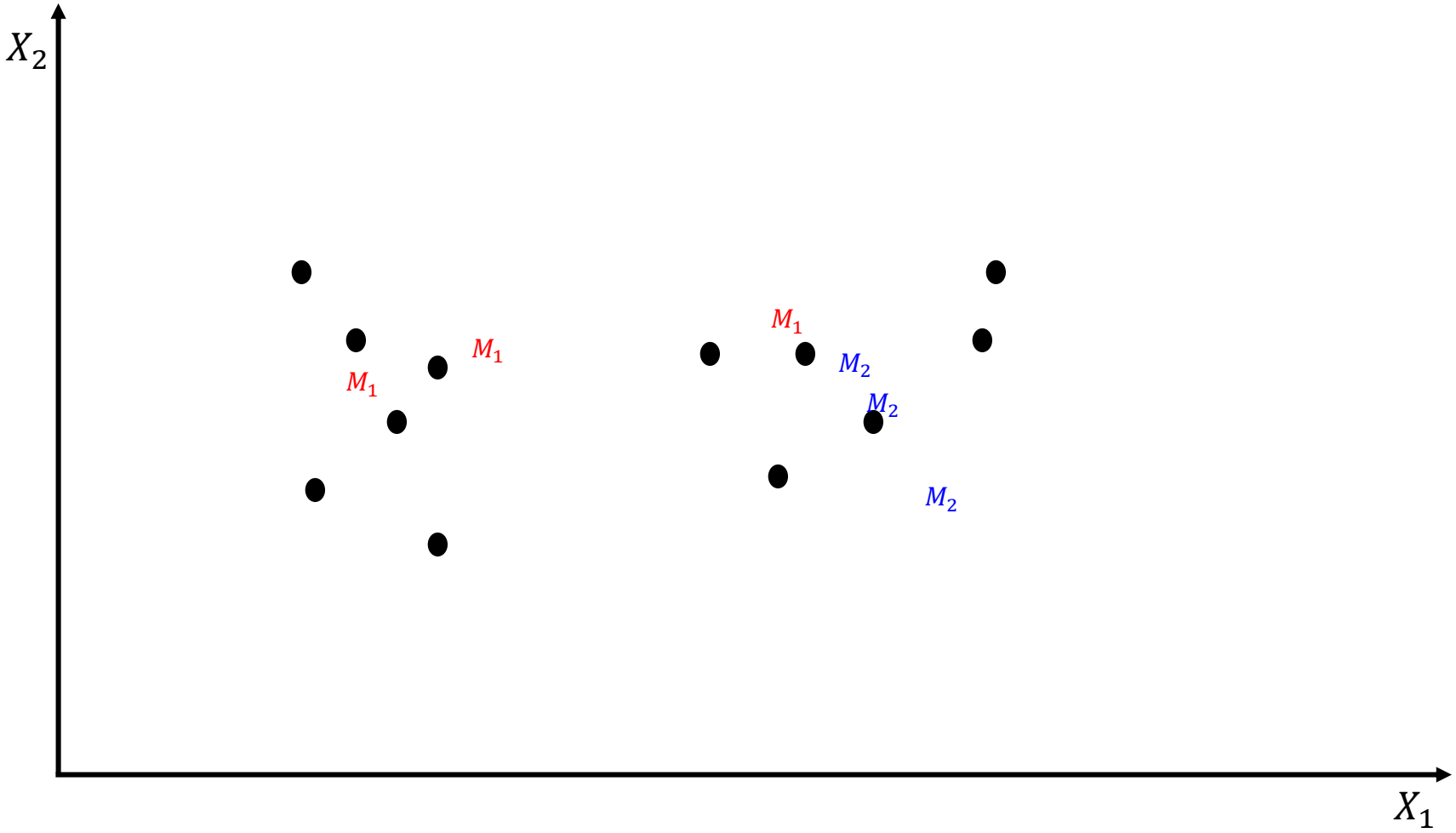
IDC HERZLIYA

K-Means

- Initialize randomly the k-means $\mu_1, \cdots, \mu_k$
- Repeat
  - For each instance
    - Assign it to the nearest cluster w.r.t its mean μ_i
  - Re-computes μ_i for each cluster
- Until no change in $\mu_1, \cdots, \mu_k$ (or any other stopping condition)
- Return $\mu_1, \cdots, \mu_k$

* Usually uses simple Euclidean distance in feature space

k-Means clustering algorithm



Agenda

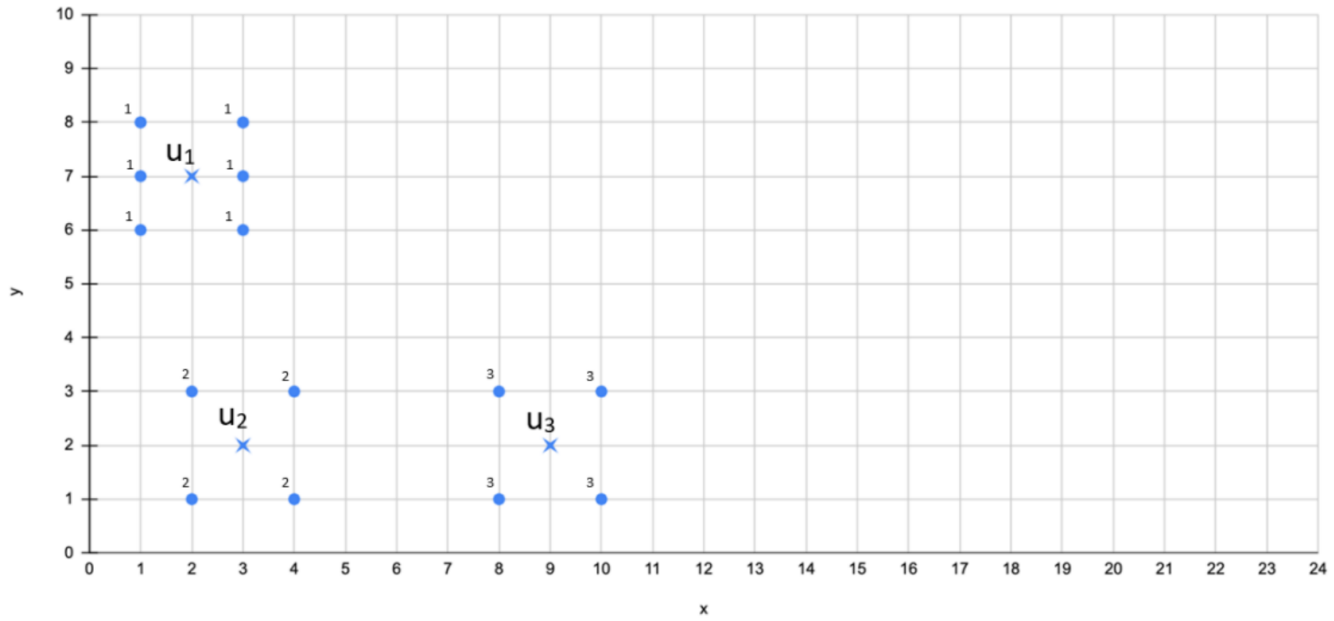
- K-Means
  - Reminder
  - Example exam question (last year moed A)
- PCA & LDA
  - PCA – reminder
  - LDA
    - Principles
    - Numeric Example
- Performance comparison (notebook) - Vanilla, LDA, PCA

IDC HERZLIYA

KMEANS – exam question

Assume the use of Euclidian distance in all parts of this question.

1. Consider the following dataset with 14 data points in $\mathbb{R}^2$, and that this is the current stage of running k-Means with $k = 3$:



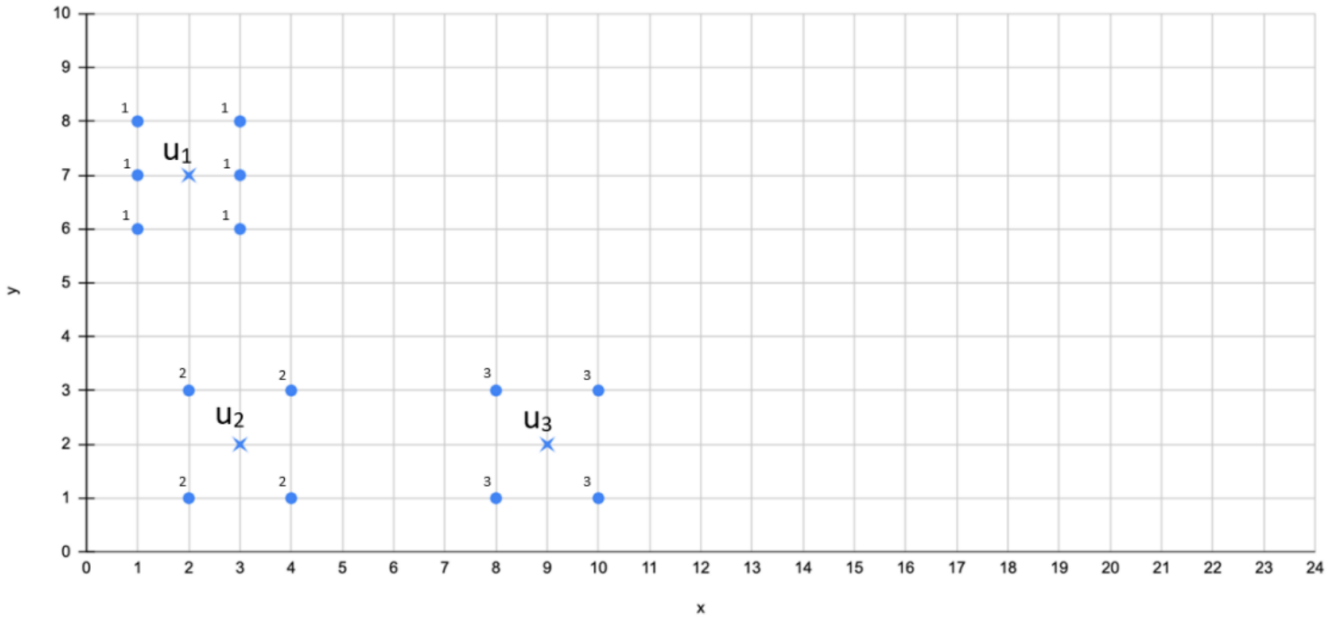
The numbers near each point represent the cluster index to which the point is currently assigned and the X symbols represent the centroids u_1, u_2, u_3.

a) (3 pts) State the formula for the loss function of the k-Means algorithm.
b) (3 pts) Is this a stable state for the algorithm? That is – will running another iteration lead to a new assignment? Explain your answer.

KMEANS – exam question

Assume the use of Euclidian distance in all parts of this question.

1. Consider the following dataset with 14 data points in $\mathbb{R}^2$, and that this is the current stage of running k-Means with $k = 3$:



The numbers near each point represent the cluster index to which the point is currently assigned and the X symbols represent the centroids u_1, u_2, u_3.

a) (3 pts) State the formula for the loss function of the k-Means algorithm.
b) (3 pts) Is this a stable state for the algorithm? That is – will running another iteration lead to a new assignment? Explain your answer.

1.a. The formula for the loss function for k-Means is given by

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - C_i\|^2$$

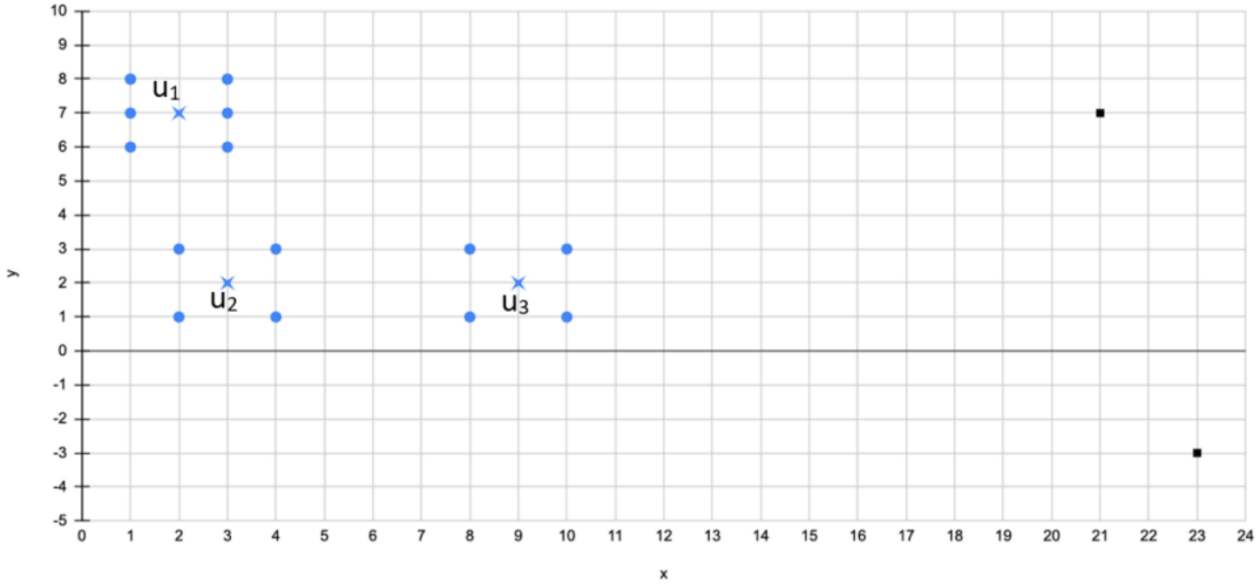
Where:
($\mathbf{x}_1$, …, $\mathbf{x}_n$) are *d*-dimensional vectors
(C_1, …, C_k) are the k cluster centroids

1.b. Yes, this is a stable assignment. If we do another iteration and update the assignments of points to centroids, it will not change.

KMEANS – exam question

After running the algorithm, you found out that there were 2 data points that were accidently omitted from the dataset given to you.
Instead of re-running the algorithm, you decided to add these points and do one more iteration of k-Means, starting with assignments. Running the k-Means single step resulted in a new assignment and in new centroids.

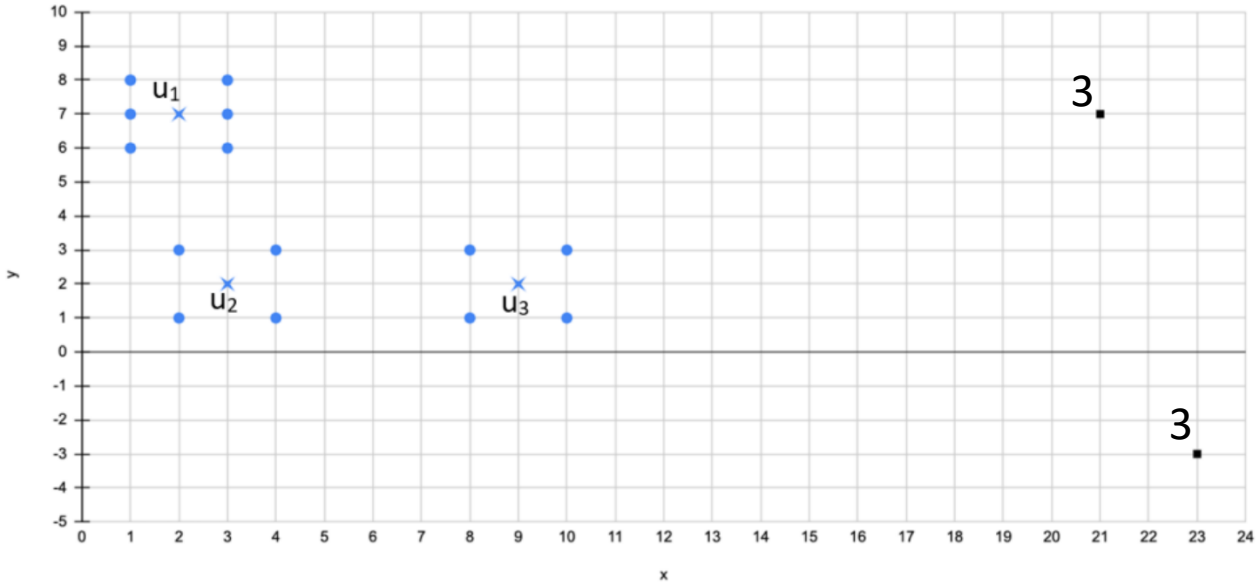
c) (2 pts) Which clusters would the new points be assigned to?
In the plot in the next page, write next to each new point the cluster it would be assigned to.



KMEANS – exam question

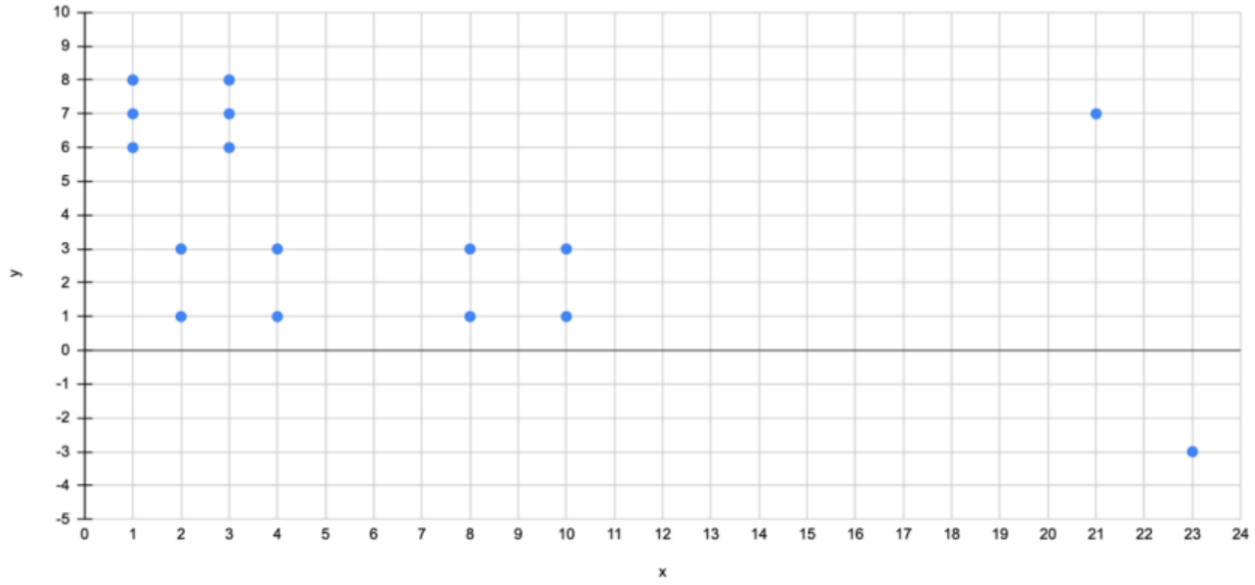
After running the algorithm, you found out that there were 2 data points that were accidently omitted from the dataset given to you.
Instead of re-running the algorithm, you decided to add these points and do one more iteration of k-Means, starting with assignments. Running the k-Means single step resulted in a new assignment and in new centroids.

c) (2 pts) Which clusters would the new points be assigned to?
In the plot in the next page, write next to each new point the cluster it would be assigned to.



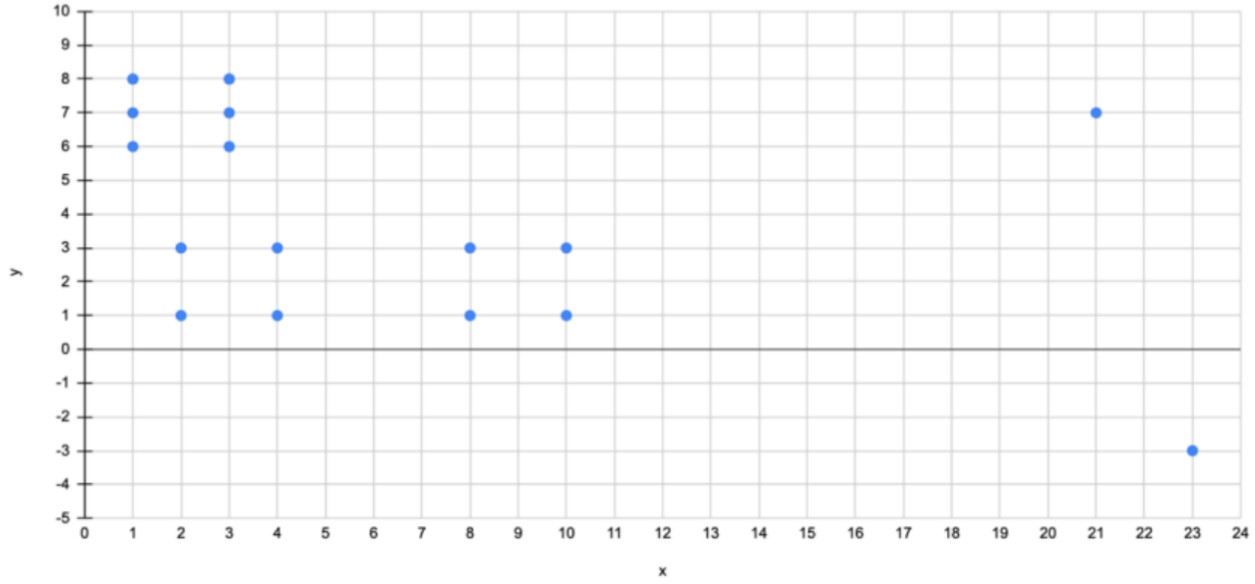
KMEANS – exam question

d) (8 pts) In the plot below, indicate the new cluster centroids using X marks and the new assignment to clusters as a number next to each data point.



KMEANS – exam question

d) (8 pts) In the plot below, indicate the new cluster centroids using X marks and the new assignment to clusters as a number next to each data point.



1.d.

u_1 and u_2 will remain the same as their corresponding points are the same as before.

u_3 now has points: [(8,1), (8,3), (10,1), (10,3), (21,7), (23,-3)]

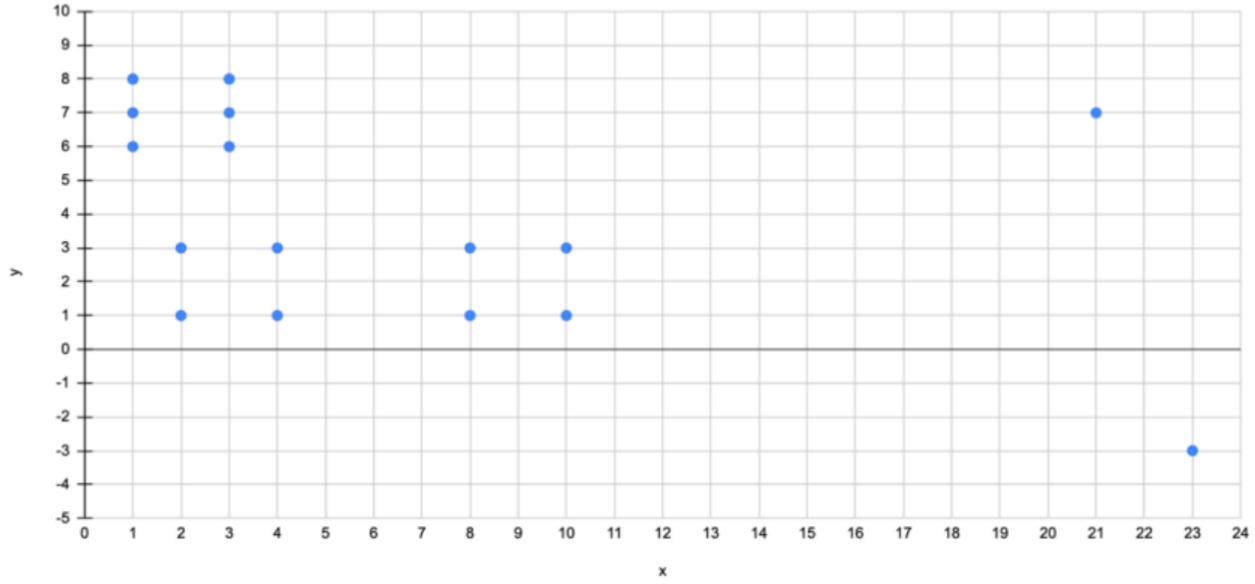
Therefore, u_3 will be:

$$u_3 = \left(\frac{8+8+10+10+21+23}{6}, \frac{(1+3+1+3+7+(-3))}{6}\right) = (13.33, 2)$$

IDC HERZLIYA

KMEANS – exam question

d) (8 pts) In the plot below, indicate the new cluster centroids using X marks and the new assignment to clusters as a number next to each data point.



1.d.

u_1 and u_2 will remain the same as their corresponding points are the same as before.

u_3 now has points: [(8,1), (8,3), (10,1), (10,3), (21,7), (23,-3)]

Therefore, u_3 will be:

$$u_3 = \left(\frac{8+8+10+10+21+23}{6}, \frac{(1+3+1+3+7+(-3))}{6}\right) = (13.33, 2)$$

Updating assignments:

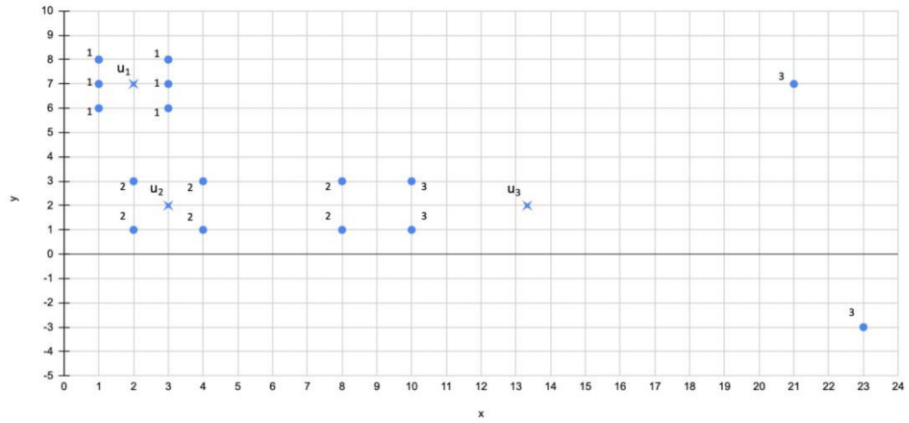
Notice that now points [(8,1), (8,3)] belong to cluster u_2 and not u_3 since the distance to u_2 is smaller.

For example, the distance between (8,1) and u_2 is

$$d\big((3,2),(8,1)\big) = \sqrt{5^2+1^2} = \sqrt{26}$$

While the distance between (8,1) and u_3 is

$$d\big((13.33,2),(8,1)\big) = \sqrt{5.33^2+1^2} = \sqrt{29.41}$$



KMEANS – exam question

2. The loss function for k-Means, which you stated above is also called the *inertia*.
   You performed k-Means on data with 16 distinct points and with k = range(16).
   As output, you recorded the respective values of the inertia for the cluster structure obtained for each k.
   a. (2 pts) Which column (A-F) of the table below potentially represents the resulting output? (Assume that for each k the algorithm has converged to the global optimum).

k	A Inertia	B Inertia	C Inertia	D Inertia	E Inertia	F Inertia
1	847	847	847	847	847	847
2	290	535	535	290	535	290
3	140	377	377	140	377	140
4	78	180	180	78	180	78
5	26	110	110	26	110	26
6	20	75	75	20	75	20
7	16	20	20	16	20	16
8	12	16	16	12	16	16
9	10	10	10	10	10	10
10	8	8	8	8	8	8
11	6	6	6	6	6	10
12	4	4	4	4	4	4
13	2.5	2.5	2.5	2.5	2.5	2.5
14	1	-1	0.5	1	1	1
15	0.5	-0.1	0	0.5	0.5	0.5
16	0	0	0	0.1	0	0

   b. (5 pts) For each of the other columns indicate why it cannot be the output of the process.
   c. (2 pts) According to the "knee/elbow" method learnt in class to find the optimal k, which k would you chose to work with?

KMEANS – exam question

2. The loss function for k-Means, which you stated above is also called the *inertia*.
   You performed k-Means on data with 16 distinct points and with k = range(16).
   As output, you recorded the respective values of the inertia for the cluster structure obtained for each k.
   a. (2 pts) Which column (A-F) of the table below potentially represents the resulting output? (Assume that for each k the algorithm has converged to the global optimum).

k	A Inertia	B Inertia	C Inertia	D Inertia	E Inertia	F Inertia
1	847	847	847	847	847	847
2	290	535	535	290	535	290
3	140	377	377	140	377	140
4	78	180	180	78	180	78
5	26	110	110	26	110	26
6	20	75	75	20	75	20
7	16	20	20	16	20	16
8	12	16	16	12	16	16
9	10	10	10	10	10	10
10	8	8	8	8	8	8
11	6	6	6	6	6	10
12	4	4	4	4	4	4
13	2.5	2.5	2.5	2.5	2.5	2.5
14	1	-1	0.5	1	1	1
15	0.5	-0.1	0	0.5	0.5	0.5
16	0	0	0	0.1	0	0

   b. (5 pts) For each of the other columns indicate why it cannot be the output of the process.
   c. (2 pts) According to the "knee/elbow" method learnt in class to find the optimal k, which k would you chose to work with?

Answers:

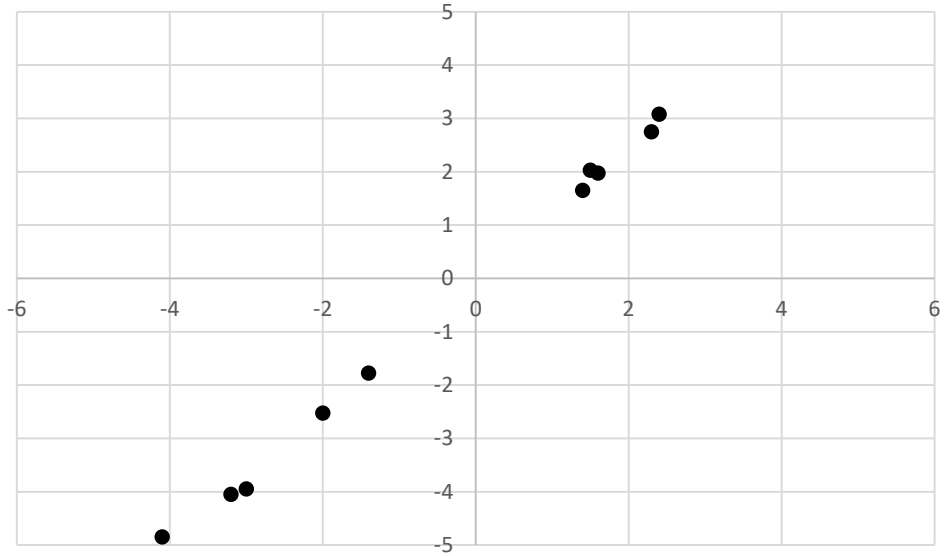
- A- Correct, the "knee" is at k=5 by analyzing the slope of the graph.
- B- Has a negative inertia value, inertia is always positive.
- C- We should get zero inertia only at 16 centroids, not 15 (if you mentioned all points might not be different, this was also accepted).
- D- At 16 clusters, each point is a centroid, so inertia should be 0 and not 0.1.
- E- Correct, the "knee" is at k=7.
  (Large decrease in % from 75 to 20, followed by a small decrease from 20 to 16)
- F- Inertia should always be decreasing, here it's going up when we go from k=10 to k=11.

IDC HERZLIYA

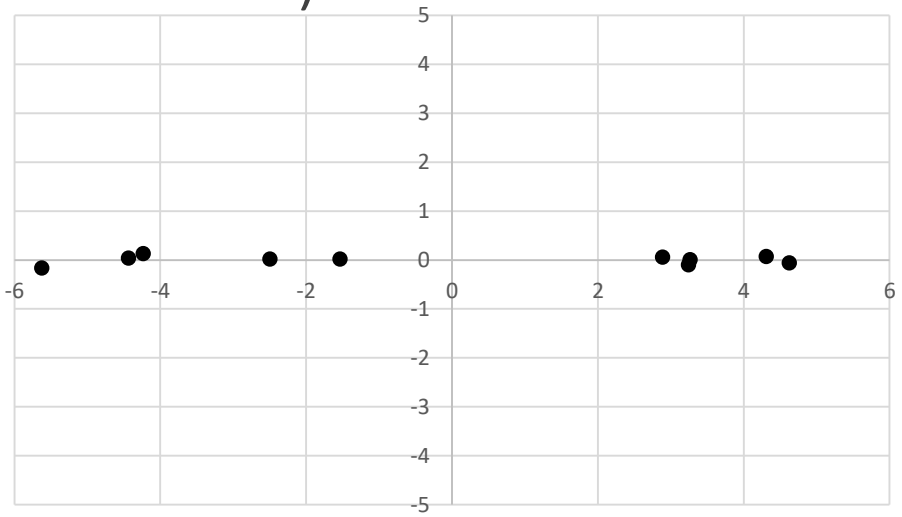
Agenda

- K-Means
  - Reminder
  - Example exam question (last year moed A)
- PCA & LDA
  - PCA – reminder
  - LDA
    - Principles
    - Numeric Example
- Performance comparison (notebook) - Vanilla, LDA, PCA

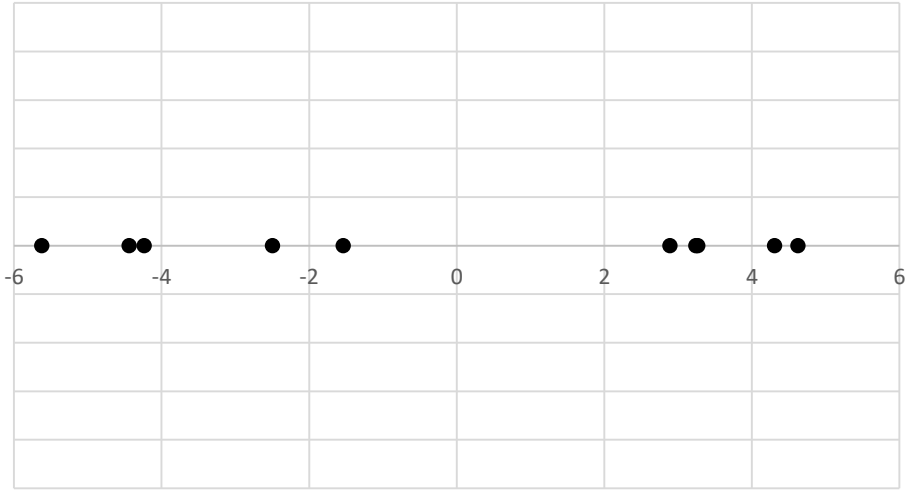
PCA – example visualization (end result)



Original data

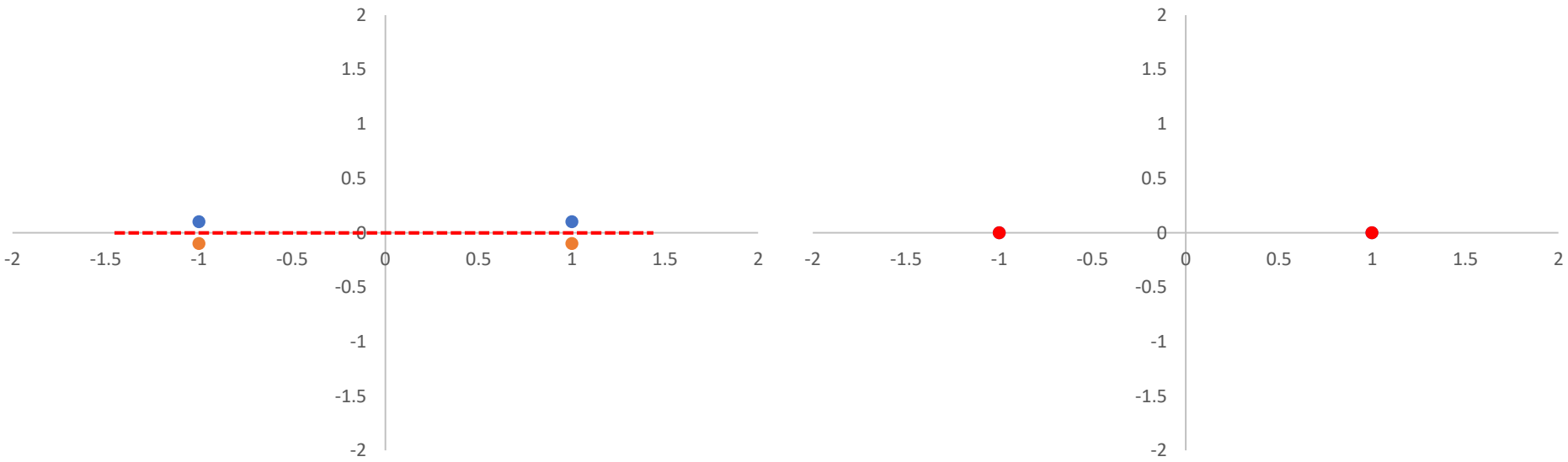


Rotated (using PCA)



Rotated + feature selection (only 1 PC)

PCA – classification problem



LDA

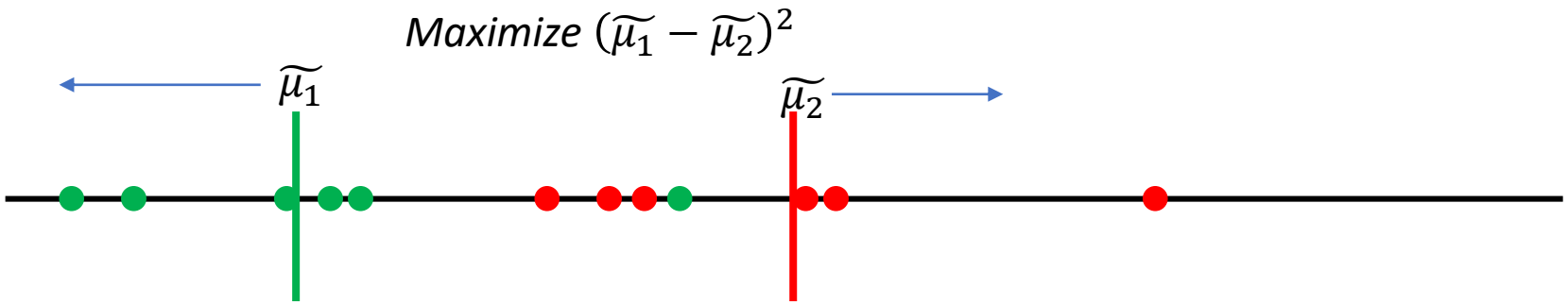
- Principles
- Example

LDA - Principles

- A linear mapping to a lower dim that maximizes class seperability
- Principles
  - Maximize the distance between class means
  - Minimize the variation within each class

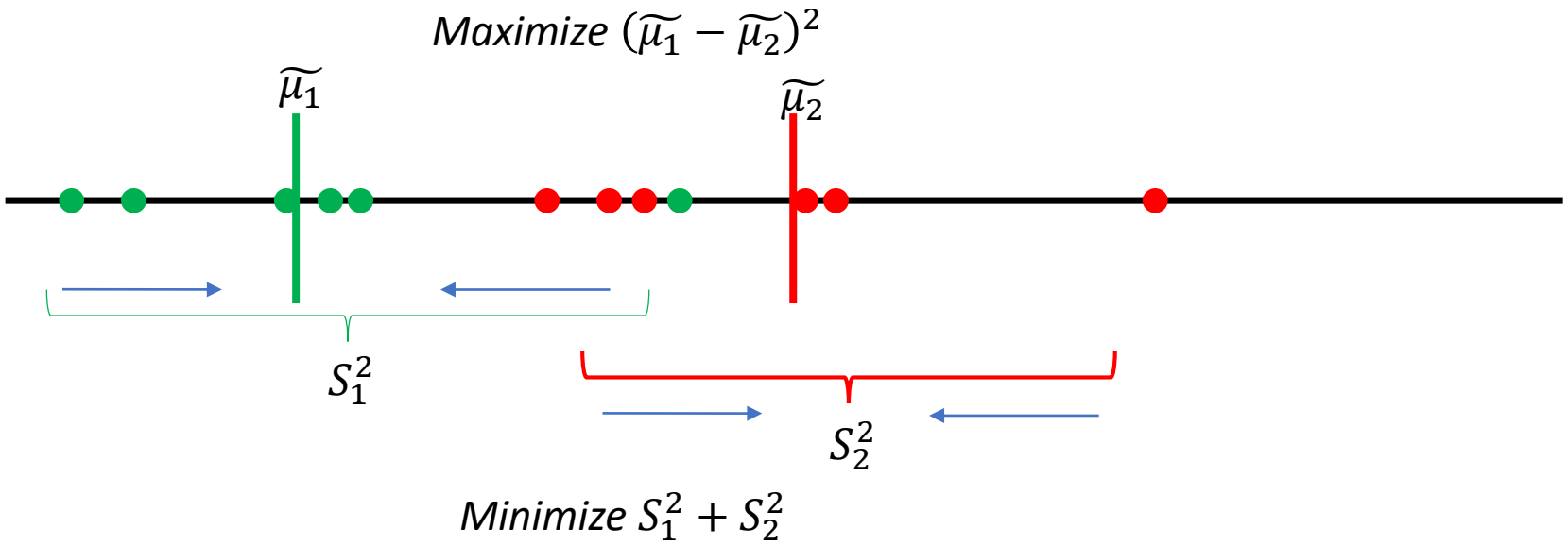
LDA - Principles

- A linear mapping to a lower dim that maximizes class seperability
- Principles
  - **Maximize the distance between class means**
  - Minimize the variation within each class



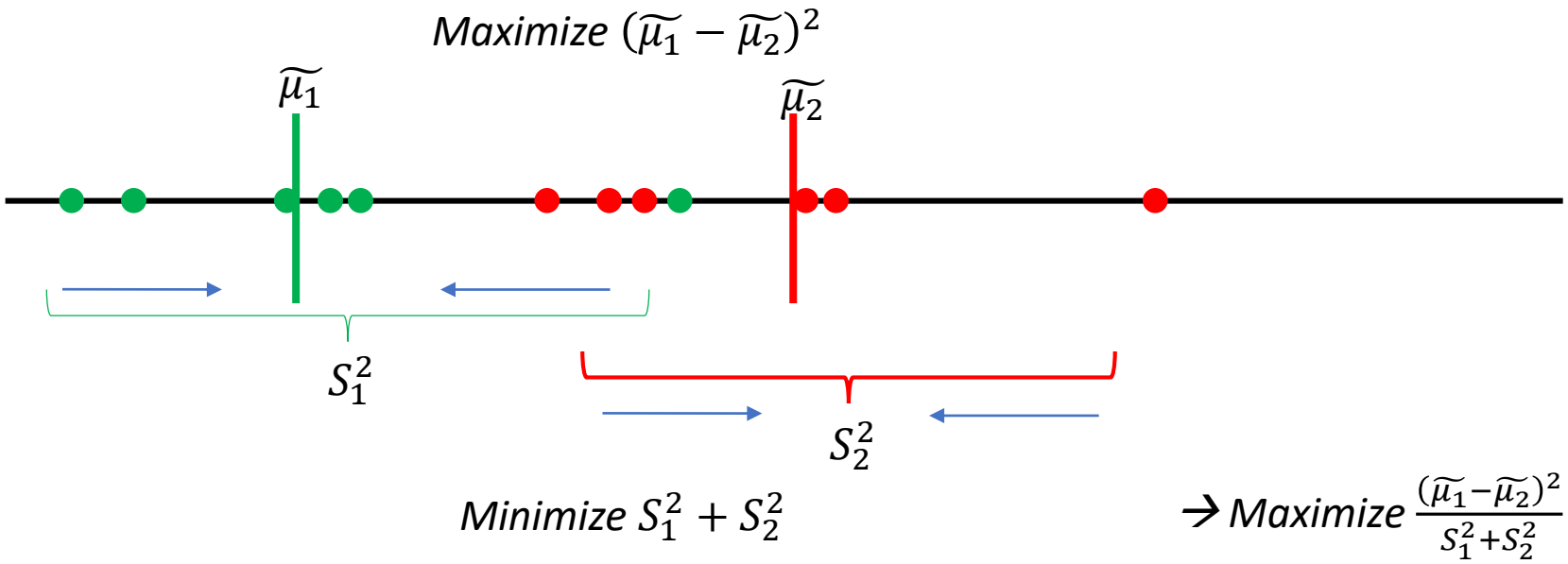
LDA - Principles

- A linear mapping to a lower dim that maximizes class seperability
- Principles
  - Maximize the distance between class means
  - **Minimize the variation within each class**



LDA - Principles

- A linear mapping to a lower dim that maximizes class seperability
- Principles
  - Maximize the distance between class means
  - **Minimize the variation within each class**



LDA - Solution

1. Define the within class scatter (compactne ss) : $S_C = \sum_i S_i$

$$S_i = \sum_{x \in D_i} (x - \mu_i)(x - \mu_i)^T$$

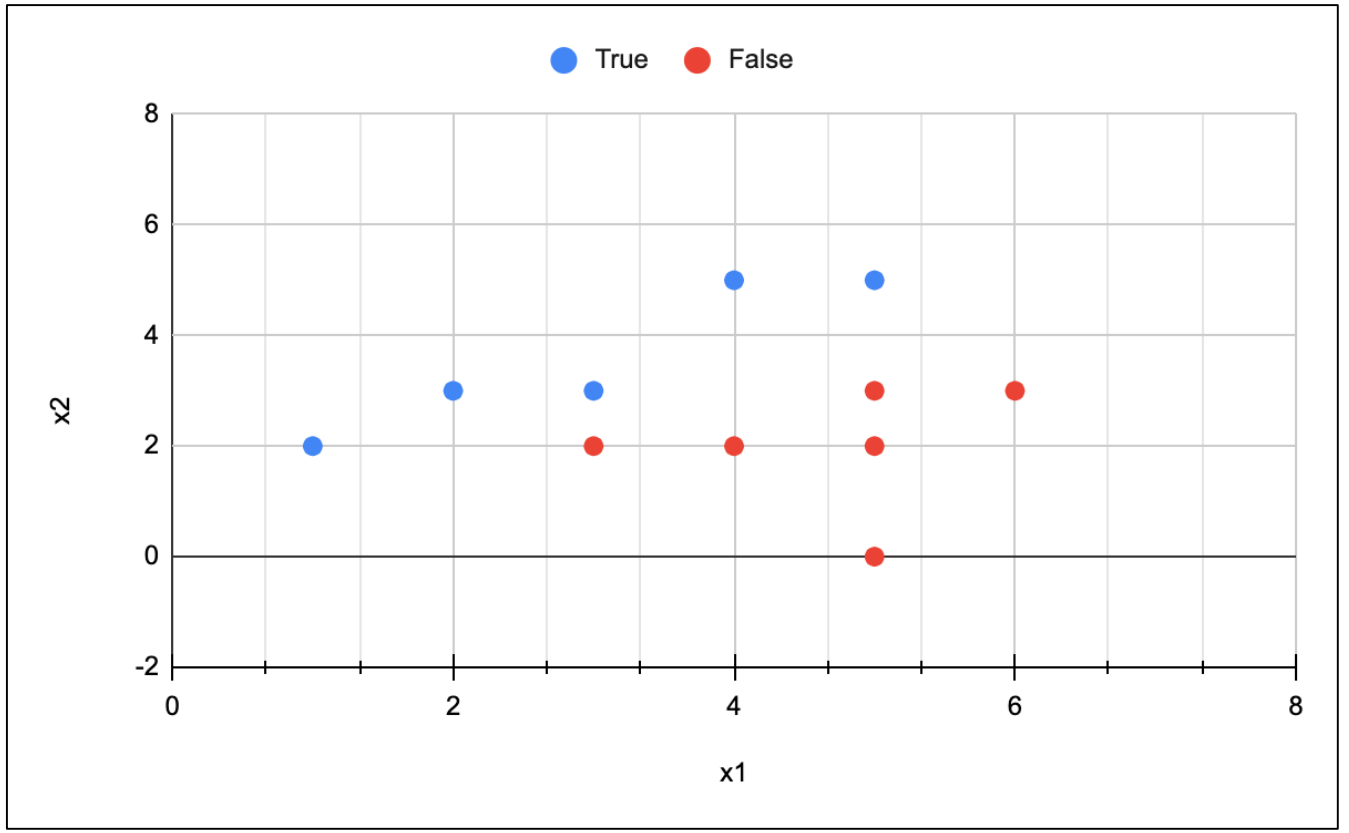
2. Define the between class scatter :

$$S_B = \sum_i N_i (\mu_i - \mu)(\mu_i - \mu)^T$$

3. *F* ind eigenvecto rs of $S_C^{-1} S_B$
4. *S* ort them and transform instances to a subspace
5. Find classifier in this subspace

LDA - Example

x1	x2	Label
1	2	TRUE
2	3	TRUE
3	3	TRUE
4	5	TRUE
5	5	TRUE
4	2	FALSE
5	0	FALSE
5	2	FALSE
3	2	FALSE
5	3	FALSE
6	3	FALSE



LDA - Example

x1	x2	Label	$x - \mu_i$	
1	2	TRUE	-2	-1.6
2	3	TRUE	-1	-0.6
3	3	TRUE	0	-0.6
4	5	TRUE	1	1.4
5	5	TRUE	2	1.4
4	2	FALSE	-0.67	0
5	0	FALSE	0.33	-2
5	2	FALSE	0.33	0
3	2	FALSE	-1.67	0
5	3	FALSE	0.33	1
6	3	FALSE	1.33	1

Calculating class means:

$$\mu_1 = (3, 3.6)$$
$$\mu_2 = (4.67, 2)$$

Calculating total (weighted) mean

$$\mu = \frac{5}{11}\mu_1 + \frac{6}{11}\mu_2 = (3.91, 2.727)$$

Calculating within class scatter: (compactness)

$$S_i = \sum_{x \in D_i} (x - \mu_i)(x - \mu_i)^T \qquad S_C = \sum_i S_i$$

$$S_1 = \begin{pmatrix} -2 \\ -1.6 \end{pmatrix} * (-2 \quad -1.6) + \begin{pmatrix} -1 \\ -0.6 \end{pmatrix} * (-1 \quad -0.6) + \begin{pmatrix} 0 \\ -0.6 \end{pmatrix} * (0 \quad -0.6) + \begin{pmatrix} 1 \\ 1.4 \end{pmatrix} * (1 \quad 1.4) + \begin{pmatrix} 2 \\ 1.4 \end{pmatrix} * (2 \quad 1.4) = \begin{pmatrix} 10 & 8 \\ 8 & 7.2 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 5.33 & 1 \\ 1 & 6 \end{pmatrix}$$

$$S_C = S_1 + S_2 = \begin{pmatrix} 15.33 & 9 \\ 9 & 13.2 \end{pmatrix}$$

LDA - Example

x1	x2	Label
1	2	TRUE
2	3	TRUE
3	3	TRUE
4	5	TRUE
5	5	TRUE
4	2	FALSE
5	0	FALSE
5	2	FALSE
3	2	FALSE
5	3	FALSE
6	3	FALSE

Calculating within class scatter (compactness):

$$\boldsymbol{S_C = S_1 + S_2 = \begin{pmatrix} 15.33 & 9 \\ 9 & 13.2 \end{pmatrix}}$$

Calculating between class scatter (distance):

$$S_{B_1} = n_1(\mu_1 - \mu)^T * (\mu_1 - \mu) = 5 * \begin{pmatrix} -0.91 \\ 0.87 \end{pmatrix} * (-0.91 \quad 0.87)$$
$$= \begin{pmatrix} 4.14 & -3.96 \\ -3.96 & 3.78 \end{pmatrix}$$

$$S_{B_2} = n_2(\mu_2 - \mu)^T * (\mu_2 - \mu) = \cdots = \begin{pmatrix} 3.44 & -3.31 \\ -3.31 & 3.17 \end{pmatrix}$$

$$\boldsymbol{S_B = S_{B_1} + S_{B_2} = \begin{pmatrix} 7.58 & -7.27 \\ -7.27 & 3.17 \end{pmatrix}}$$

LDA - Example

x1	x2	Label
1	2	TRUE
2	3	TRUE
3	3	TRUE
4	5	TRUE
5	5	TRUE
4	2	FALSE
5	0	FALSE
5	2	FALSE
3	2	FALSE
5	3	FALSE
6	3	FALSE

Calculating transformation matrix ***W***:

$$A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$W = S_c^{-1} * S_B$$

$$S_B = \begin{pmatrix} 7.58 & -7.27 \\ -7.27 & 3.17 \end{pmatrix}$$

$$S_C = \begin{pmatrix} 15.33 & 9 \\ 9 & 13.2 \end{pmatrix} \rightarrow S_C^{-1} = \begin{pmatrix} 0.11 & -0.07 \\ -0.07 & 0.13 \end{pmatrix}$$

$$\rightarrow \text{W} = \begin{pmatrix} 1.36 & -1.31 \\ -1.48 & 1.42 \end{pmatrix}$$

Eigen values:

$$\det(\lambda I - W) = \det\left(\begin{pmatrix} \lambda - 1.36 & 1.31 \\ 1.48 & \lambda - 1.42 \end{pmatrix}\right) = \lambda^2 - 2.78\lambda - 0.0076$$

$$\rightarrow \lambda_1 = 2.78\,;\lambda_2 = -0.002$$

Eigen Vectors:

$$\begin{pmatrix} \lambda - 1.36 & 1.31 \\ 1.48 & \lambda - 1.42 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\rightarrow LD_1 = (-0.92\,, 1)\,; LD_2 = (0.96, 1)$$

After Normalizing:

$$LD_1 = (-0.68\,, 0.74)\,; LD_2 = (0.69, 0.72)$$

IDC HERZLIYA

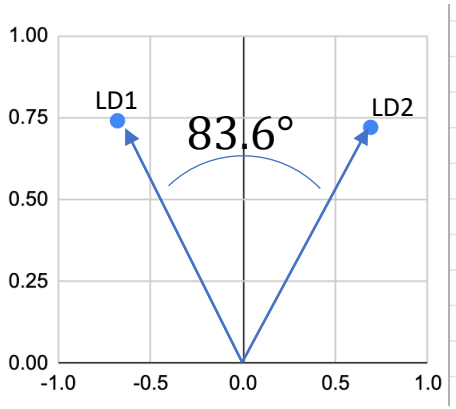
LDA - Example

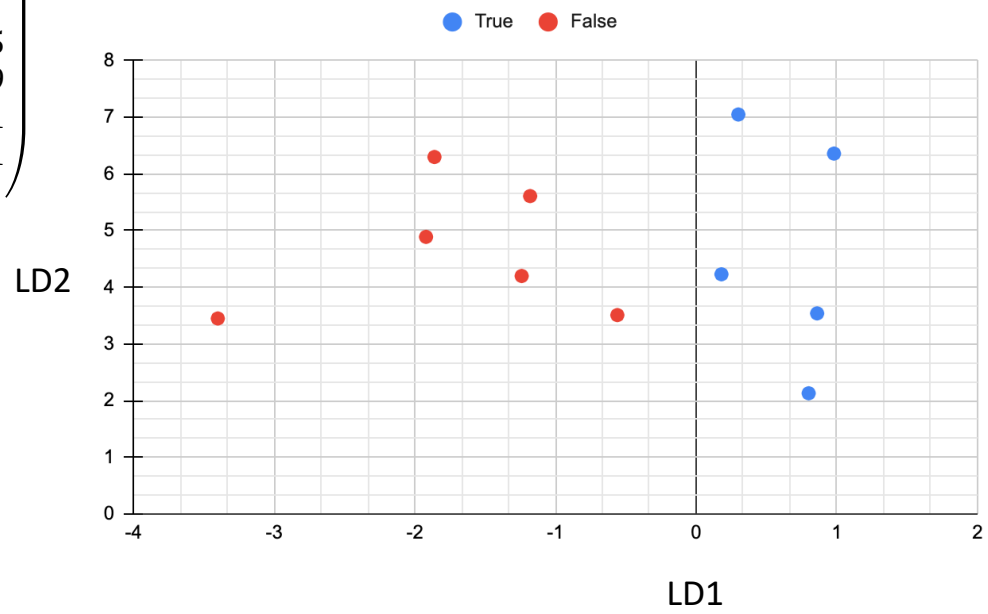
Projecting the points to the new axes

$$LD_1 = (-0.68\,, 0.74)\,; LD_2 = (0.69, 0.72)$$

x1	x2	Label
1	2	TRUE
2	3	TRUE
3	3	TRUE
4	5	TRUE
5	5	TRUE
4	2	FALSE
5	0	FALSE
5	2	FALSE
3	2	FALSE
5	3	FALSE
6	3	FALSE

$$X * \begin{pmatrix} -0.68 & 0.69 \\ 0.74 & 0.72 \end{pmatrix} = \begin{pmatrix} 0.8 & 2.13 \\ 0.86 & 3.54 \\ 0.18 & 4.23 \\ 0.98 & 6.36 \\ 0.3 & 7.05 \\ -1.24 & 4.2 \\ -3.4 & 3.45 \\ -1.92 & 4.89 \\ -0.56 & 3.51 \\ -1.18 & 5.61 \\ -1.86 & 6.3 \end{pmatrix}$$





Questions