

אלגוריתם א-מרכזים K-Means Algorithm



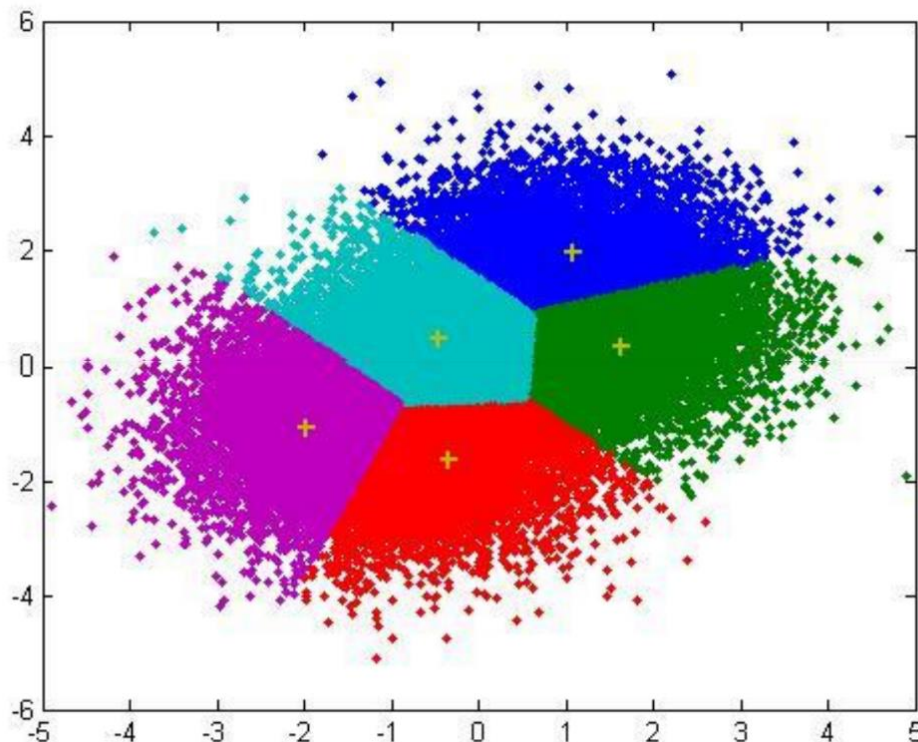
בפני חברי האיגוד הישראלי למדעני נתונים מקצועיים

רועי פולניצר, FRM, F.I.L.A.V.F.A., PDS,
בעלים ומדען הנתונים הראשי של פרדיקציות יועצים
מייסד ומנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA)

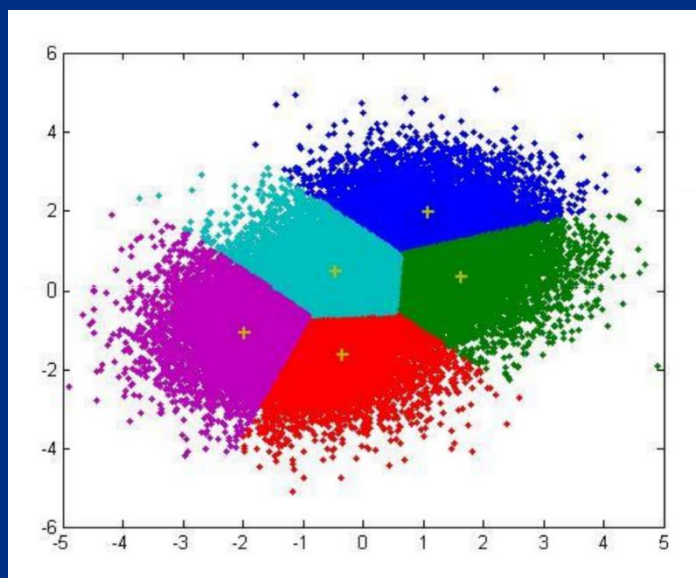
תל אביב, 30 בנובמבר 2021

- ❖ האלגוריתם הנפוץ ביותר של Clustering הוא אלגוריתם ה- K-Means המכונה בעברית א-מרכזים.
- ❖ K מלשון מספר מסוים, K קלאסטרים.
- ❖ כמו שיש לנו את אלגוריתם ה- K-Nearest Neighbors (אלגוריתם k השכן הקרוב ביותר) שבו אנחנו מחשבים או בעצם מבקשים K, למשל 5 שכנים קרובים, אז ב- K-Means אנחנו מבקשים K, למשל 5 קלאסטרים.

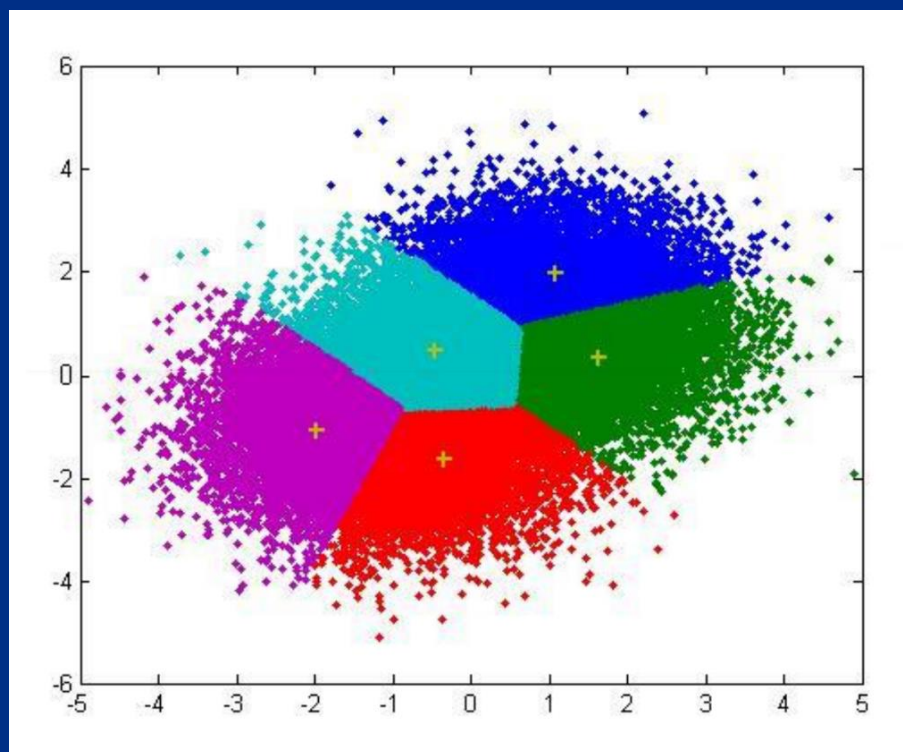
❖ כאן יש לנו דוגמא של שני פיצ'רים X_1 ו- X_2 שחולקו ל- 5 קלאסטרים, כפי שניתן לראות ב- Scatter Plot של הפיצ'רים עצמם.



- ❖ ב-K-Means Clustering אנחנו ניקח את הפיצ'רים שלנו ונחלק אותם לפי ה-K.
- ❖ למשל, אם K שווה 5 אז נחלק אותם ל-5 קלאסטרים.
- ❖ ניקח נקודה שנמצאת מתחת לקלאסטר האדום ונגיד שהיא שייכת לקלאסטר האדום, למשל לקלאסטר 3.



❖ ניקח למשל נקודה שנמצאת מעל הקלאסטר הכחול
($X_1=2.5$ ו- $X_2=5$) ונגיד שהנקודה הזו שייכת
לקלאסטר כחול, למשל לקלאסטר 4.

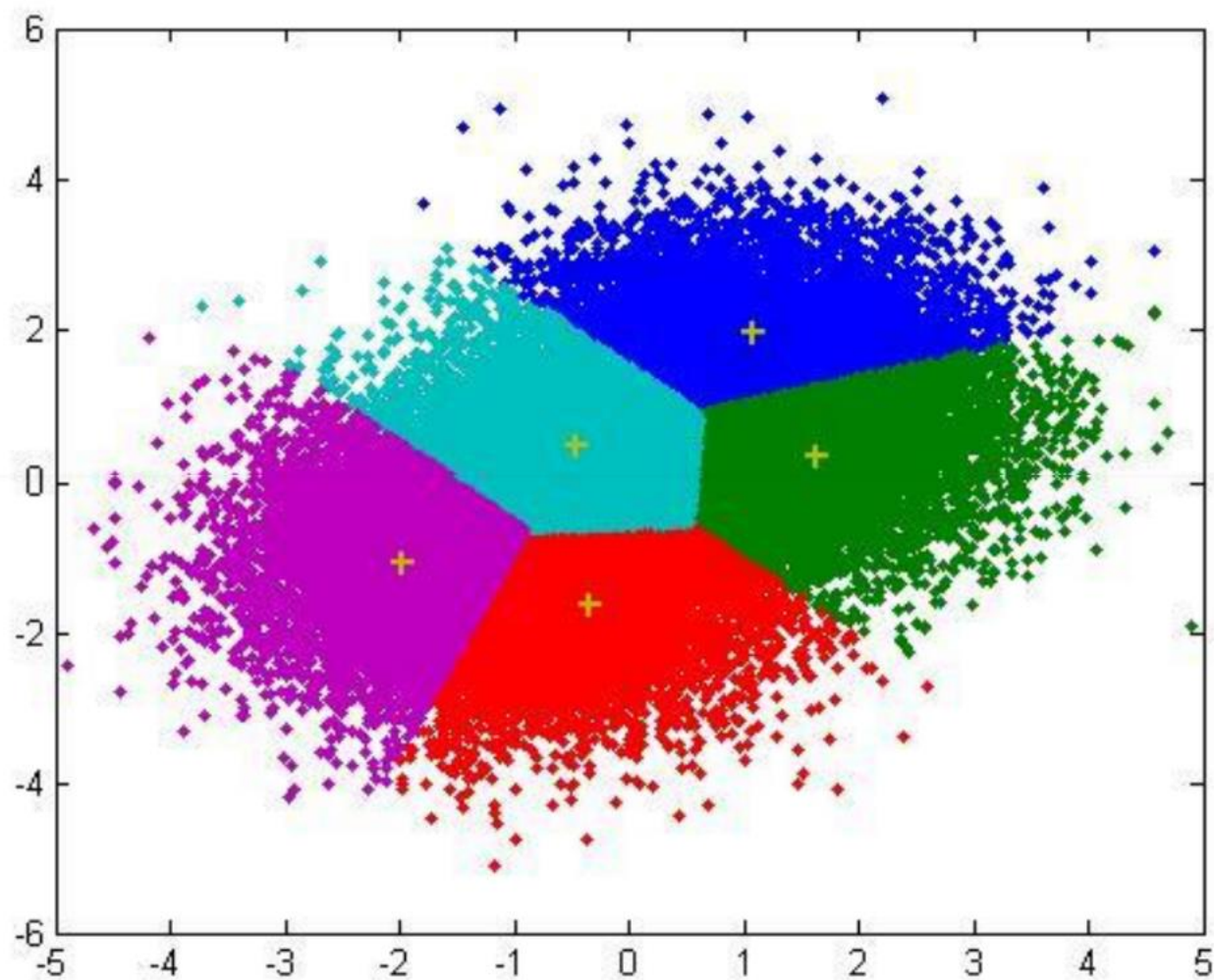


❖ מספר הקלאסטר עצמו הוא משהו שרירותי כמובן, אין לזה משמעות אם נקרא לזה 1 או 4, העיקר שנהיה עקביים.

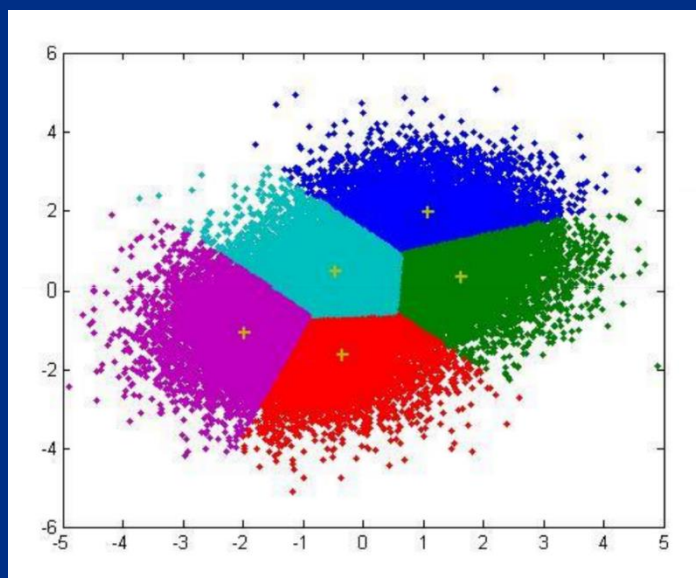
❖ אז ב- K-Means Clustering הרעיון הוא לקחת את כל הפיצ'רים ולחלק אותם לקלאסטרים.

❖ למה Means? כי Means זה מלשון ממוצעים.

❖ תכף נסביר איך האלגוריתם הזה עובד, הוא מחפש את המרכזים, בתרשים שלנו אלו הם הצלבים הצהובים.



❖ הצלב הצהוב הוא בעצם המרכז של כל קלאסטר.
❖ אם ניקח למשל את כל הנקודות הכחולות ונחשב את מרכז הכובד של כל הנקודות, אז נמצא שהצלב הצהוב שבתוך הקלאסטר הכחול הוא מרכז הכובד של כל הנקודות הכחולות.



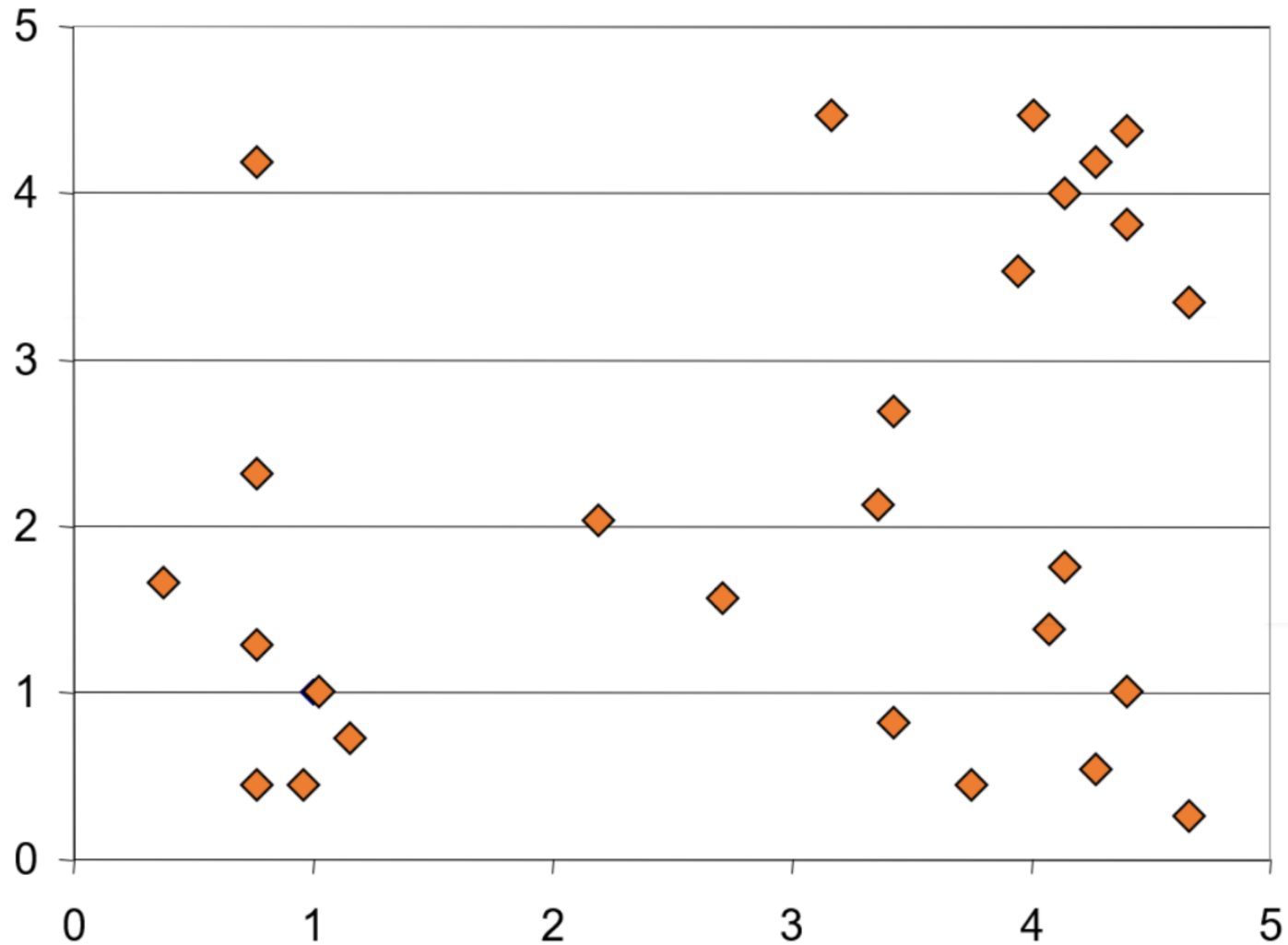
איך מוצאים מרכז כובד?

❖ פשוט עושים ממוצע על כל הנקודות ואז מוצאים את מרכז הכובד.

איך האלגוריתם עובד?

- ❖ עכשיו בוא נראה איך האלגוריתם הזה עובד.
- ❖ קודם כל מדובר באלגוריתם איטרטיבי, כלומר באלגוריתם הפועל באיטרציות.
- ❖ אז באיטרציה הראשונה בואו ניקח דוגמא קצת יותר קלילה.
- ❖ יש לי פה 2 פיצ'רים: X_1 ו- X_2 .
- ❖ X_1 נע בין 0 ל- 5 וגם X_2 נע בין 0 ל- 5.
- ❖ נציג את ה- Scatter Plot של הפיצ'רים.

איך האלגוריתם עובד?



איך האלגוריתם עובד?

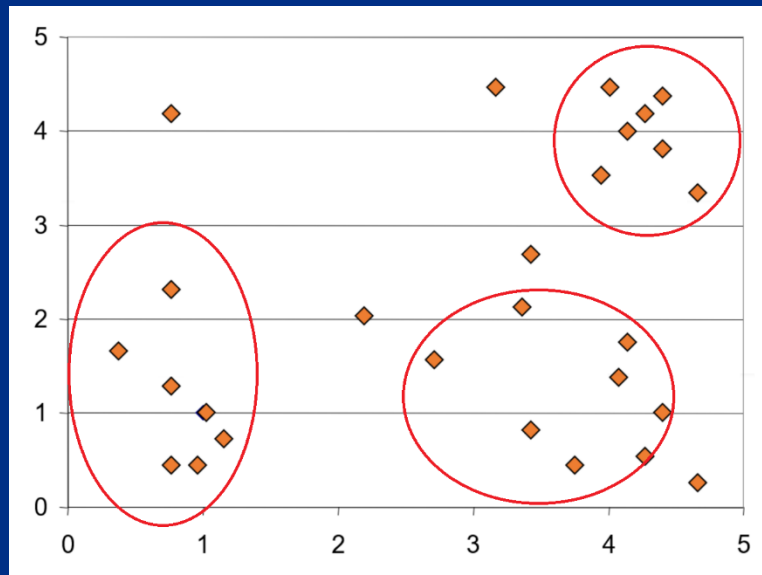
- ❖ זהו בעצם דאטה סט שבו יש לי עמודה אחת בשם X_1 , עמודה שנייה בשם X_2 .
- ❖ מה עם Y ? אין לי Y , כי זה Unsupervised Learning ולכן אני לא יודע לאן כל רשומה שייכת.
- ❖ המטרה שלנו כרגע היא לייצר את ה- Y הזה, והוא יהיה לאיזה מספר קלאסטר כל רשומה שייכת.

איך האלגוריתם עובד?

- ❖ אז אנחנו מתחילים מה- Scatter Plot של $X1$ ו- $X2$.
- ❖ זוהי בעצם הדגמה ויזואלית לאיך האלגוריתם הזה עובד.
- ❖ קודם כל אנחנו צריכים לקבוע כמה קלאסטרים אנחנו בכלל רוצים, שפה זה גם שאלה כבדה בפני עצמה.
- ❖ בואו נניח שאני רוצה לקחת את הפיצ'רים שלי ולחלק אותם ל- 3 קלאסטרים.

איך האלגוריתם עובד?

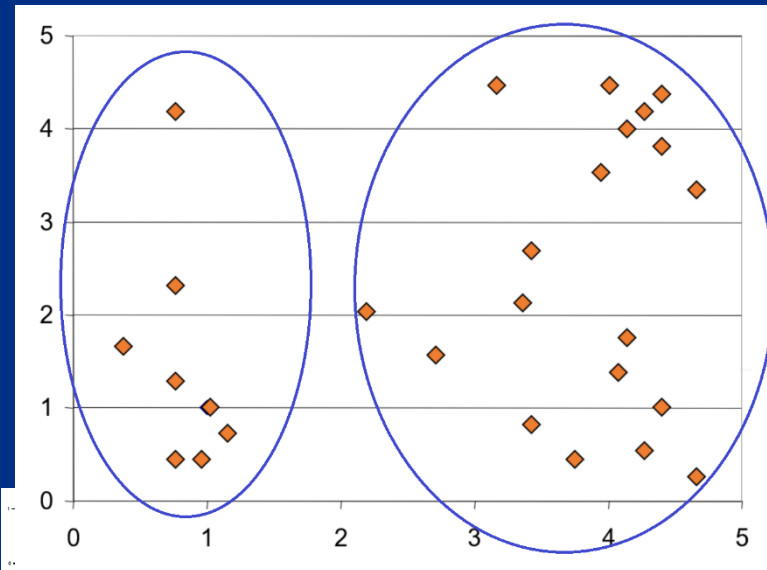
❖ אם אני מחלק את הדאטה שלי ל- 3 קלאסטרים, אז אני מצפה שקלאסטר אחד יהיה באזור הימני העליון של ה- Scatter Plot, הקלאסטר השני יהיה באזור הימני התחתון של ה- Scatter Plot והקלאסטר השלישי יהיה באזור השמאלי התחתון של ה- Scatter Plot.



איך האלגוריתם עובד?

❖ מה שה- K-Means ינסה למצוא זה מרכזים של קלאסטרים, כך שיהיו כמה שיותר נקודות שיהיו קרובות אליו. שזה שלושה.

❖ אם אני מבקש למשל K שווה 2, כלומר, 2 קלאסטרים, אז נקבל את התרשים הבא:



❖ נניח שאני מבקש 3 קלאסטרים – למעשה אני קובע לו מראש כמה קלאסטרים אני רוצה.

❖ בשלב הראשון, האלגוריתם מגריל 3 נקודות אקראיות.

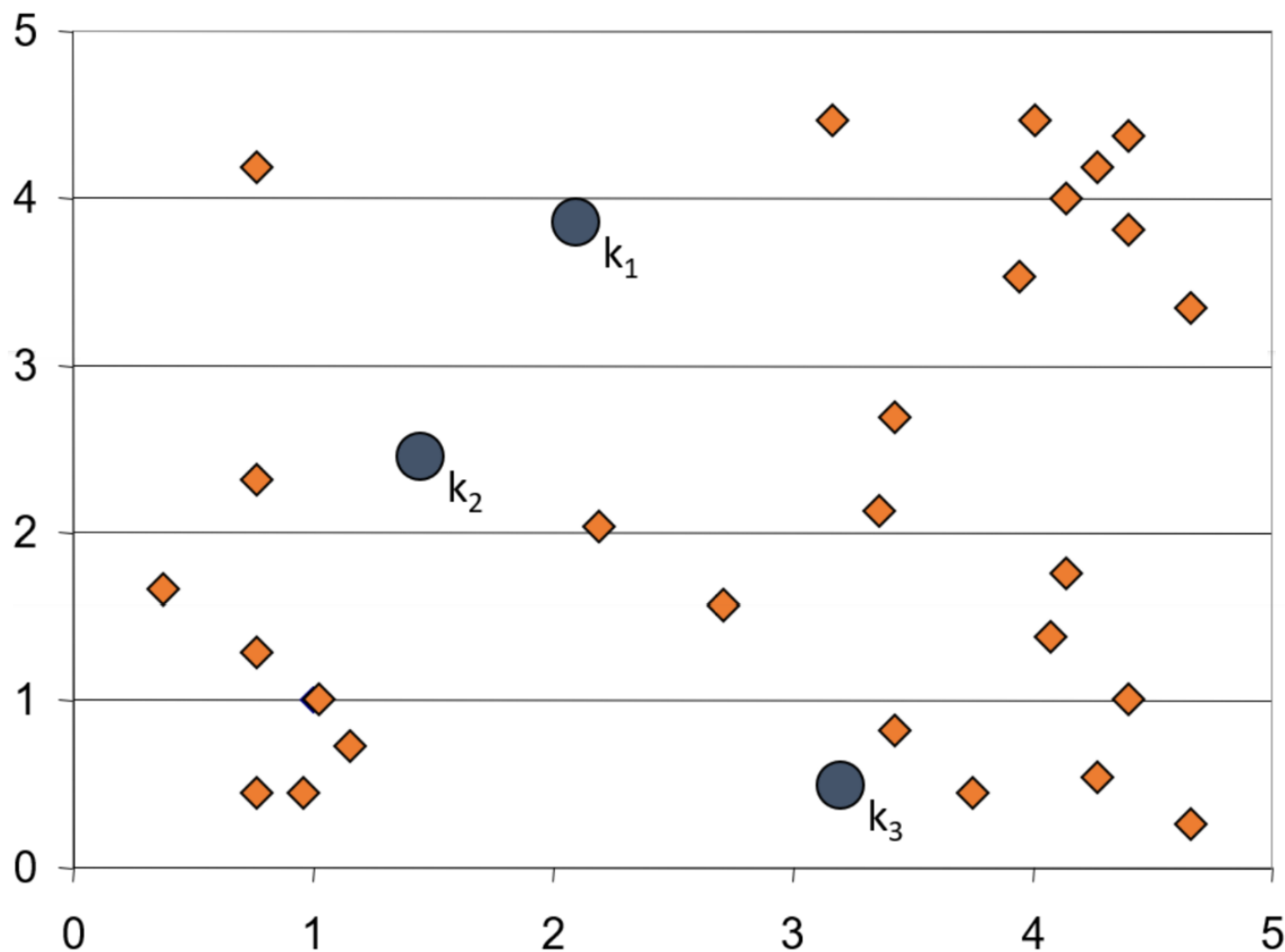
❖ עכשיו פה יש כל מיני אלגוריתמים.

❖ אפשר לבקש אלגוריתם שלא מגריל נקודות אקראיות אלא בוחר רשומות מתוך הדאטה סט ומתחיל איתן.

❖ אבל זה לא משנה, כי בסוף ההתכנסות היא מאוד מאוד דומה.

❖ אז בשלב הראשון הגרלנו 3 נקודות אקראיות: k_1 , k_2 ו- k_3 .

❖ זה השלב הראשון, שלב ההתחלה. הגרלתי 3 נקודות אפורות.

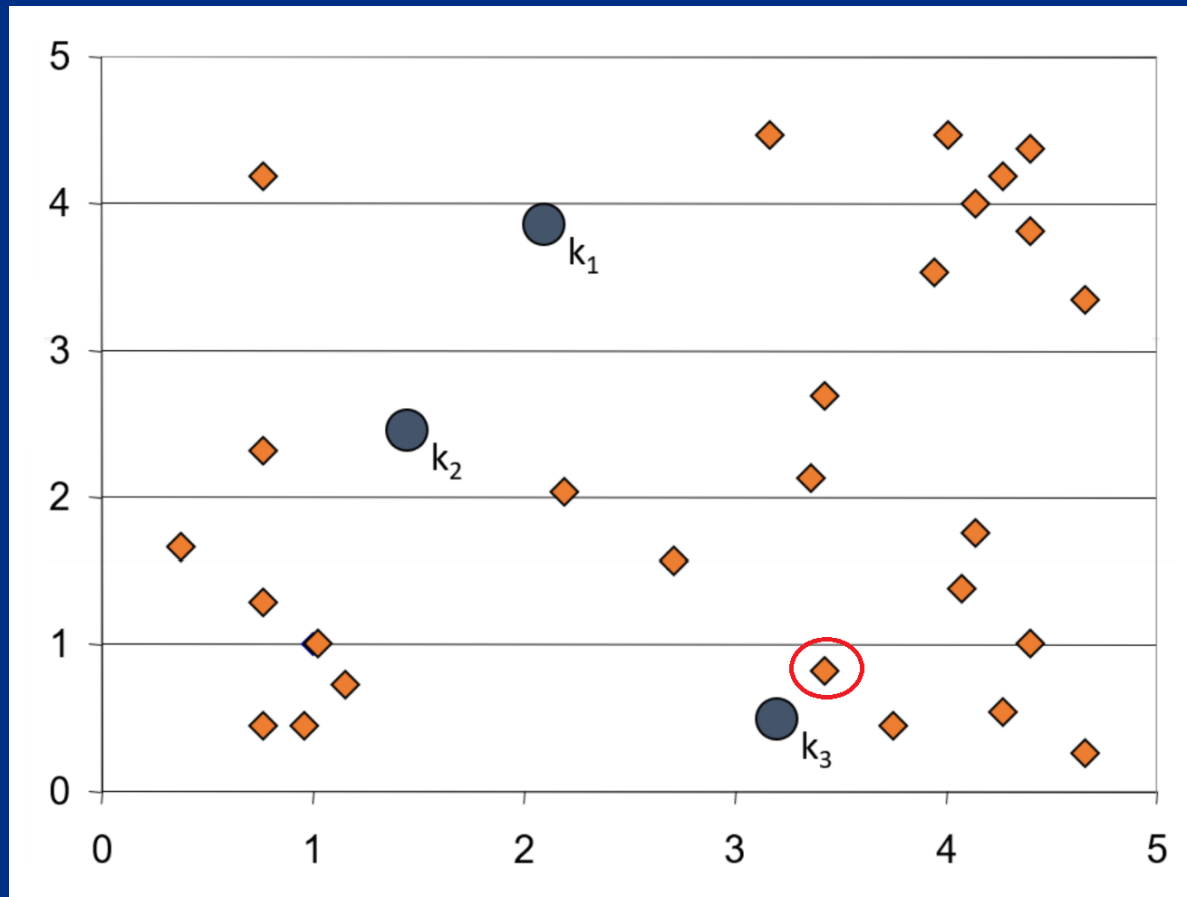


❖ בשלב השני אני לוקח כל אחת מהנקודות בדאטה סט שלי, שכרגע אף אחת מהן לא משוייכת לאף קלאסטר כי אין להן label, ואני מתחיל לחשב עבור כל נקודה, לאיזה מרכז כובד היא הכי קרובה.

❖ למעשה לכל נקודה בדאטה סט שלי אני מחשב את 3 מרחקים אוקלידיים (לפי משפט פיתגורס) – אחת מכל מרכז כובד לנקודה, כאשר בסוף התהליך הזה אני משייך את הנקודה למרכז הכובד בעל המרחק הקצר ביותר ממנה.

❖ כך למשל, הנקודה המוקפת בעיגול אדום, זה מקרה קל, היא שייכת ל- k_3 כי היא הכי קרובה ל- k_3 .

❖ כך למשל, הנקודה המוקפת בעיגול אדום, זה מקרה קל, היא שייכת ל- k_3 כי היא הכי קרובה ל- k_3 .

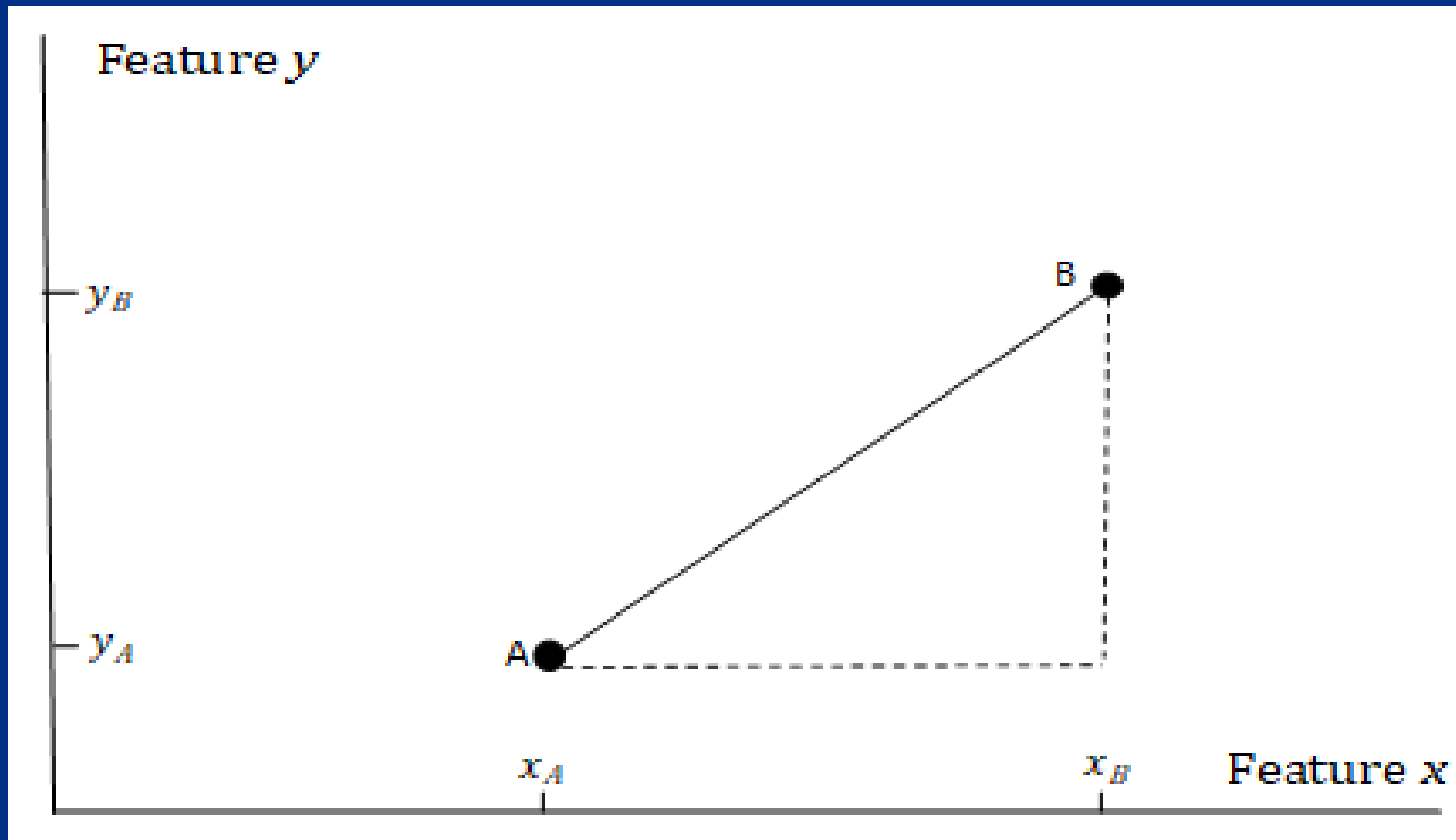


מה זה מרחק אוקלידי?

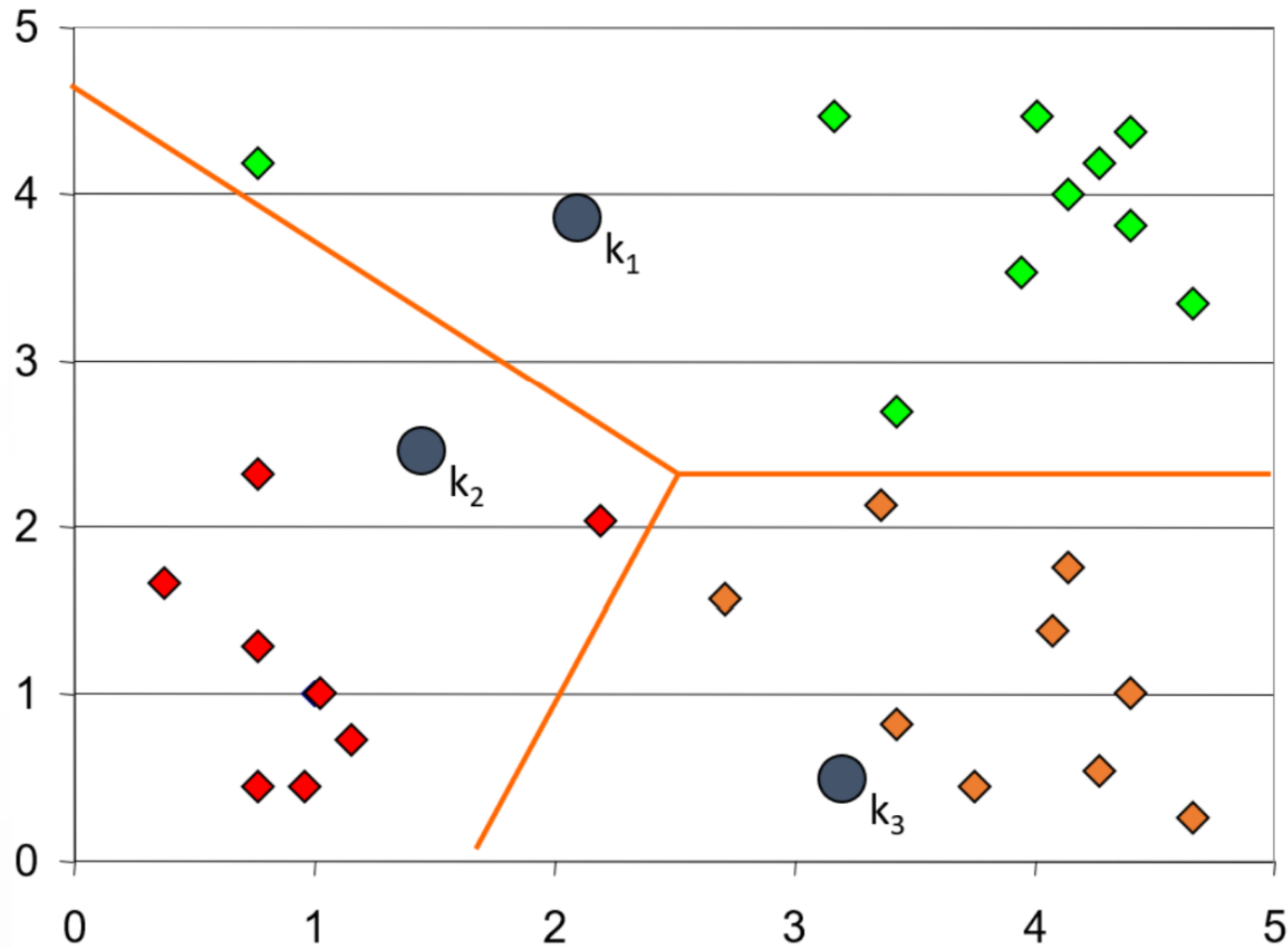
- ❖ מרחק אוקלידי, כלומר $X1$ פחות מרכז הכובד בריבוע ועוד $X2$ פחות מרכז הכובד בריבוע ועל כל זה שורש.
- ❖ כלומר, אני מחשב את המרחק האוקלידי של כל נקודה ונקודה בדאטה סט מכל אחד מ- 3 מרכזי הכובד.
- ❖ ואז אני לוקח כל נקודה ובודק לאיזה מרכז כובד היא הכי קרובה ואז אני מסמן אותה ב- label המתאים.

תזכורת למרחק אוקלידי (משפט פיתגורס)

$$Distance = \sqrt{(y_B - y_A)^2 + (x_B - x_A)^2}$$



- ❖ למשל, נקודה שהמרחק האוקלידי הכי קצר שלה הוא למרכז הכובד k_2 אז אני מסמן אותה כ- $label\ 2$.
- ❖ נקודה אחרת שהמרחק האוקלידי הכי קצר שלה הוא למרכז הכובד k_1 אז אני מסמן אותה כ- $label\ 1$ וכך הלאה.
- ❖ מה ה- $label$ הזה אומר? $class1$, $class2$ או $class3$. או במקרה שלנו, אדום, ירוק או כתום.



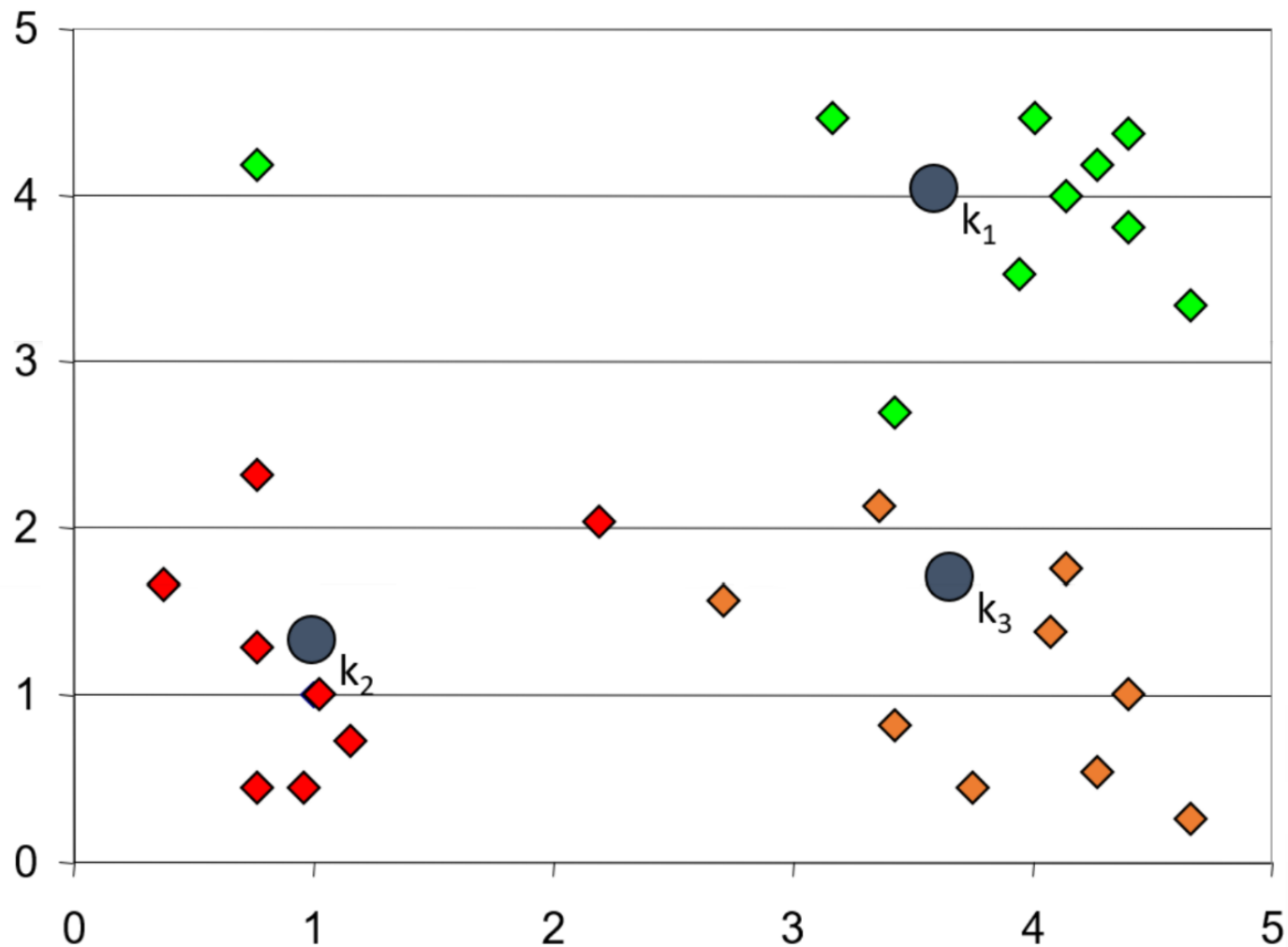
❖ שני השלבים שתוארו עד כאן הם רק האיטרציה הראשונה.

❖ חשוב להגיד שהאיטרציה הראשונה היא ייחודית, כלומר, היא מתרחשת פעם אחת, הואיל ויש לה אלגוריתם מיוחד משלה וכעת יתר האיטרציות מתרחשות בצורה של לולאה, כאשר כל איטרציה דומה לאיטרציה הקודמת שלה מבחינת האלגוריתם.

- ❖ בהתחלה האלגוריתם שלי בחר 3 נקודות אקראיות ובדק כל נקודה לאיזה class היא שייכת.
- ❖ כעת אני עובד בלולאה, כלומר, אני מתחיל להפעיל את האלגוריתם שאמור להתכנס.
- ❖ איך האלגוריתם הזה עובד? פשוט מאוד, יש לי כאן את מרכזי הכובד של כל אחד מהקלאסטרים.
- ❖ משעה שקבעתי את מרכז הכובד של כל אחד מהקלאסטרים, אני לוקח את הנקודות ששייכות לאותו מרכז כובד ומחשב את מרכז הכובד החדש של כל אחד מהקלאסטרים.

- ❖ למשל, אני לוקח את כל הנקודות האדומות ומחשב מהו מרכז הכובד החדש של הנקודות האדומות.
- ❖ איך אני עושה את זה? ממוצעים. K-Means זה ממוצע אז פשוט עושים ממוצעים.
- ❖ כלומר, אני עושה ממוצע על כל הנקודות שמאותו הצבע ואז אני מקבל את מרכז הכובד החדש שלהן.

השלב השלישי



❖ שימו לב לאנימציה.

❖ $k2$ היה בנקודה $X1=2.5$ ו- $X2=2.5$, לפי ההקצאה הראשונית, והוא עבר לנקודה $X1=1.0$ ו- $X2=1.33$.

❖ למה הוא עבר לשם? כי לקחתי את התוצאה של ה- Clustering מהאיטרציה הראשונה ועכשיו אני מחשב את מרכז הכובד החדש של אותן רשומות שהיו שייכות אליו. ולכן $k2$ זז ימינה ולמטה.

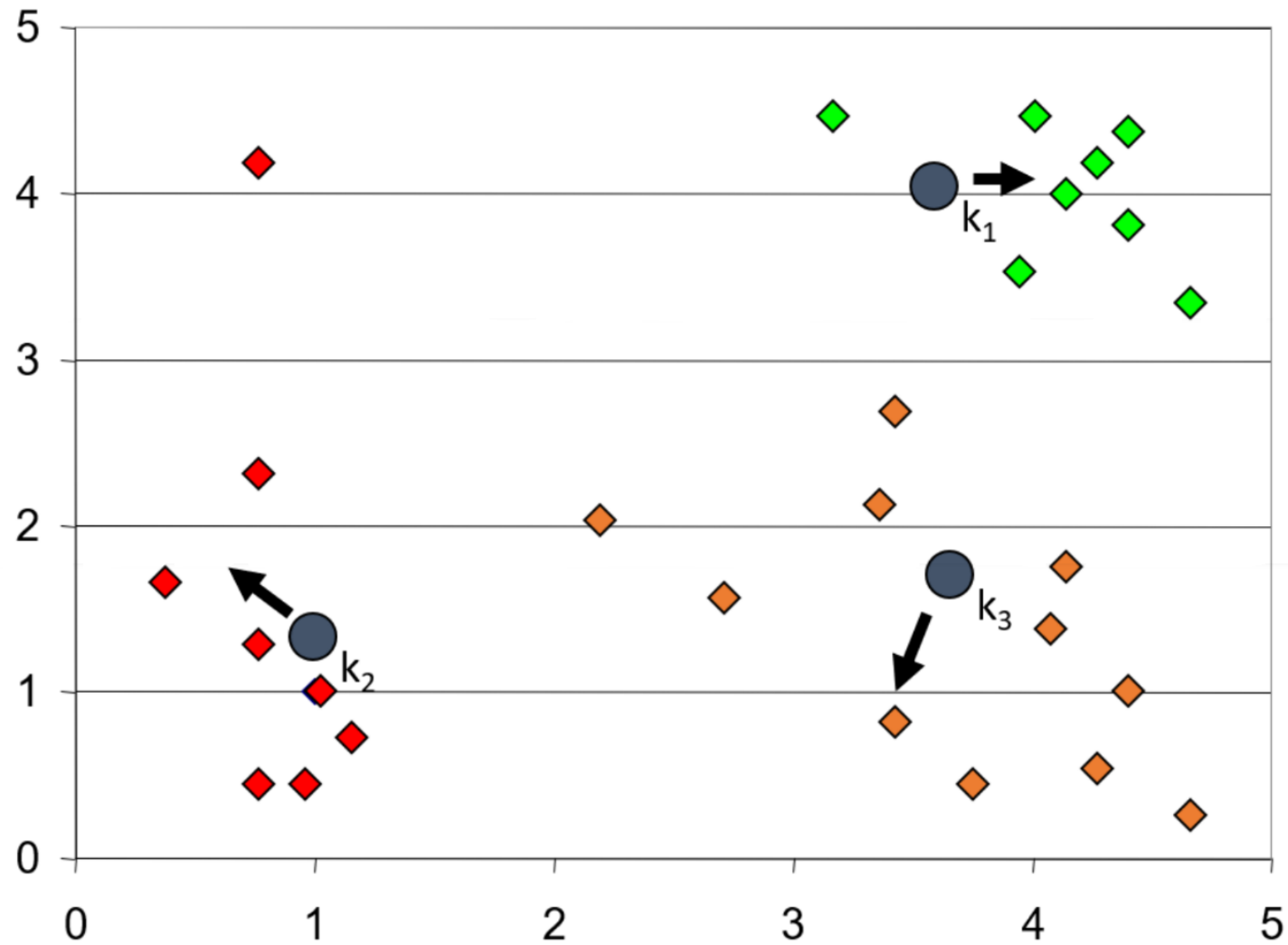
❖ $k1$ שהיה מרכז הכובד של הנקודות הירוקות, היה הגיוני שהוא יזוז קצת ימינה, למרכז הכובד של הנקודות הירוקות, והוא אכן זז ימינה.

❖ מה שהאלגוריתם עושה, זה פשוט ממוצע אריתמטי על הפיצ'רים.

❖ $k3$ שנבחר אקראית להיות איפה שהוא למטה, אחרי שעשיתי ממוצע על הנקודות הכתומות הוא זז ימינה ולמעלה, כי זה פחות או יותר מרכז הכובד של הנקודות הכתומות.

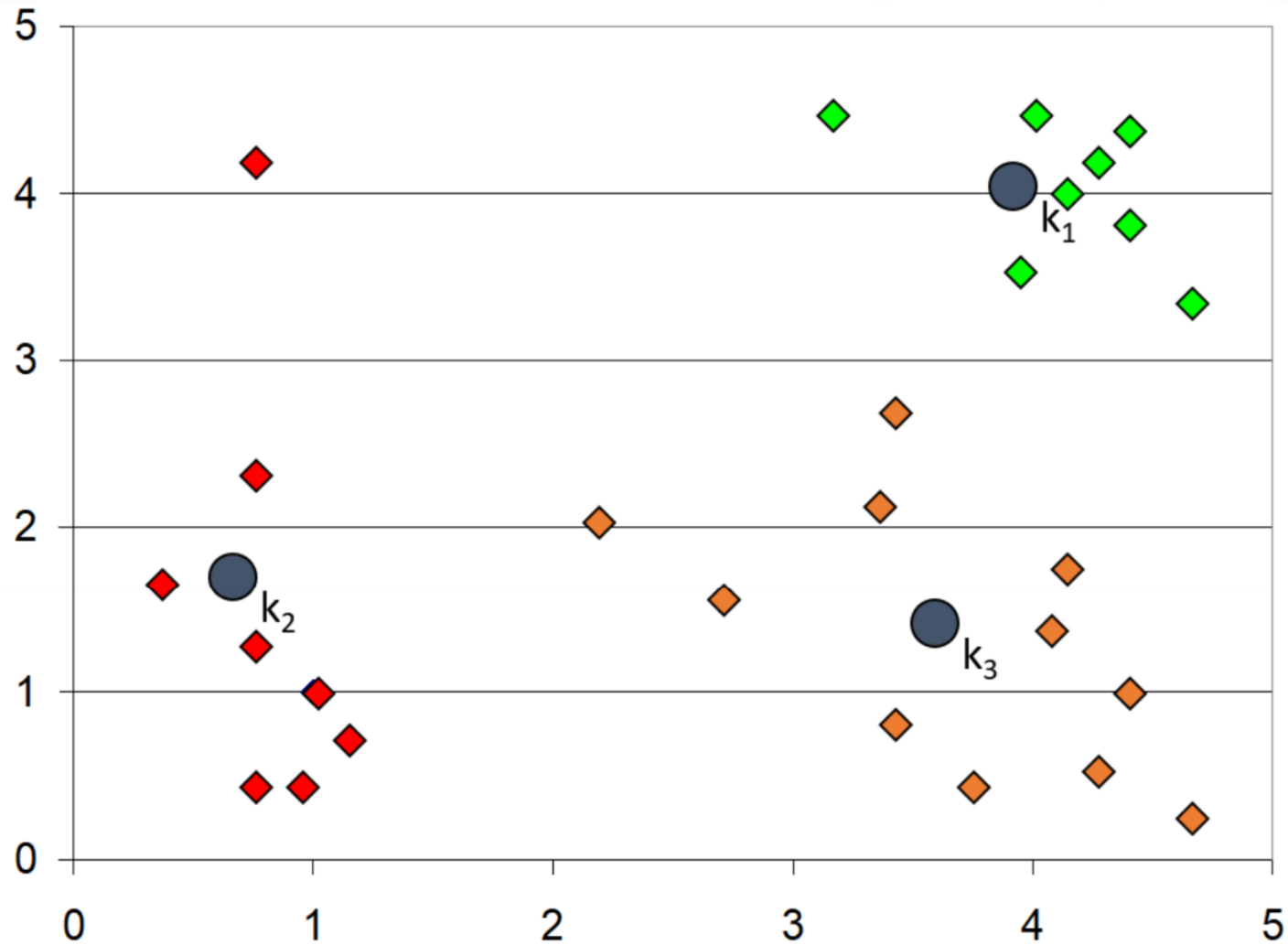
- ❖ עכשיו חשוב לזכור שאני נמצא בלולאה, חישובתי כרגע את מרכזי הכובד. מהו אם כך השלב הבא שאני צריך לעשות? בדיוק כפי שעשיתי קודם.
- ❖ אני לוקח כל נקודה ואני שואל למי היא שייכת, כי הרי שיניתי את מרכזי הכובד שלי.
- ❖ כל נקודה במרחב, אני בודק עכשיו האם היא שייכת לקלאסטר שמקודם שייכתי אותה אליו או שהיא שייכת לקלאסטר אחר, כי הרי שיניתי את מרכזי הכובד שלי.

השלב הרביעי



- ❖ וזה מה שאני עושה באיטרציה הבאה, כלומר, את זה אני עושה כל הזמן, מחשב את מרכזי הכובד ואחרי שחישבתי את מרכזי הכובד אני לוקח כל נקודה או כל רשומה בעצם ונותן לה label חדש, כלומר שם של קלאסטר חדש, לפי הקלאסטר שהיא הכי קרובה אליו.
- ❖ ואחרי שעשיתי את זה, אני עוד פעם מחשב את מרכזי הכובד מחדש, כי הנקודות שלי השתנו, כל אחת קיבלה label מסוג אחר.
- ❖ וזה מה שקרה כאן.

השלב החמישי



- ❖ אז אמרנו ממשיכים, המשכנו לאיטרציה הבאה, $k1$, $k2$ ו- $k3$ זזו עוד קצת.
- ❖ באיטרציה הבאה, אם אני אבדוק ואראה שהם לא זזים בכלל או זזים בשבריר של מרחק – אז אני אפסיק את האיטרציה, כי אין טעם להמשיך.

❖ אז מה קרה פה? יצרתי בהתחלה 3 מרכזי כובד אקראיים במרחב, התחלתי לבדוק אלו נקודות קרובות לכל אחד ממרכזי הכובד, יצרתי קלאסטרים ראשוניים.

❖ אחרי זה חישבתי מחדש את מרכזי הכובד ומרכזי הכובד קצת זזו.

❖ ראיתי אלו נקודות שייכות עכשיו למרכזי הכובד החדשים, שייכתי אותן לקלאסטרים החדשים, בצורה חדשה, וחישבתי עוד פעם את מרכזי הכובד.
❖ אז מרכזי הכובד שוב פעם זזו.

- ❖ לאט לאט מרכזי הכובד הללו יתחילו להיות יציבים.
- ❖ כלומר, לאט לאט הם יזוזו אבל בצורה הרבה הרבה יותר איטית, בצורה הרבה יותר חלשה במרחב והדברים האלה יתחילו להתייצב.
- ❖ זו הכוונה באלגוריתם איטרטיבי, שאנחנו לאט לאט, בכל איטרציה מזיזים את מרכזי הכובד, עד לנקודה שבה הם כבר לא זזים.

❖ רו"ח בנצי ישראל, FRM שואל האם יש תלות בעצם
ב- random state הראשוני לגבי התוצאה הסופית?

❖ אז יש למעשה שני אלגוריתמים.

❖ אלגוריתם אחד שאומר בו נגריל נקודות במרחב ואז

אנחנו בעצם תלויים בהגרלה הראשונית הזו.

❖ אלגוריתם שני אומר בוא נבחר נקודות אקראיות מתוך

הדאטה סט שלי ולא סתם נקודות אקראיות במרחב,

ואז אנחנו תלויים בבחירה הראשונית.

❖ בשני המקרים הראו שזה לא ממש משנה לתהליך

ההתכנסות.

❖ למשל, אם באיטרציה הראשונה אני אבחר את k_1 , k_2 ו- k_3 בטעות בקצה השמאלי העליון, אז פשוט יקח לאלגוריתם קצת יותר זמן עד שהוא יתחיל להתכנס, כלומר, עד שהוא יתחיל להתקדם.

❖ אז זה רק עניין של זמן ביצוע.

❖ מאידך, יכול להיות שאם אני אבחר את 3 מרכזי הכובד הראשוניים להיות באותו אזור (נגיד בקצה השמאלי העליון) אז יכול להיות שהם יתקעו שם, כי הם אחד על השני ולא יוכלו לזוז.

❖ נציין שישנם אלגוריתמים שאם הם רואים שאין שינוי במרכזי הכובד במרחב שלהם אז מגרילים מרכזי כובד מחדש.
❖ אבל זה כבר אופטימיזציה של האלגוריתם.
❖ בגדול:

K-MEANS “out of the box” in scikit learn works well.

פולניצר, ר' (2019), "למידה ללא השגחה
(Unsupervised Learning)", סטטוס - כתב עת לחשיבה
ניהולית ואסטרטגית, נובמבר.



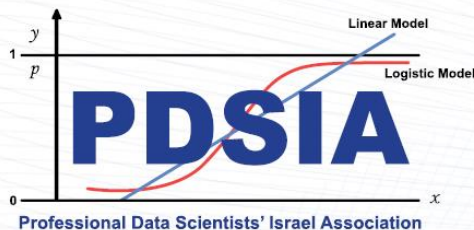
Professional Data Scientist

The Professional Data Scientists' Israel Association, recognizing successful completion of the requirements of the PDS, PDSIA's Professional Data Scientist Certification, has conveyed upon

Roi Polanitzer

The designation of PDS.

This award recognizes the attained knowledge and understanding of the building blocks essential to the successful practice of Professional Data Science. These include Python Development, Data Analysis and Visualization with Python, Machine Learning, and Deep Learning at a level required in the industry in general and in the Data worlds in particular.



Rami Yosef

Prof. Rami Yosef, PDS
Chairman, PDSIA

Granted on this date:
December 26, 2019

PDSIA Certificate Number:
PDS1001

נשמח לעמוד לרשותכם

רועי פולניצר, PDS, F.I.L.A.V.F.A., FRM

בעלים של פרדיקציות יועצים – פיתוחי מודלים ואנליזות מתקדמות

מנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA)

polanitz7@gmail.com

www.pdsia.org