תרגול 4: אשכול

# אשכול

המטרה:

להפריד את ה- data שלנו לקבוצות (clusters) כאשר לכל קבוצה איזשהן תכונות דומות.

דוגמאות למקרים בהם נרצה לבצע אשכול:

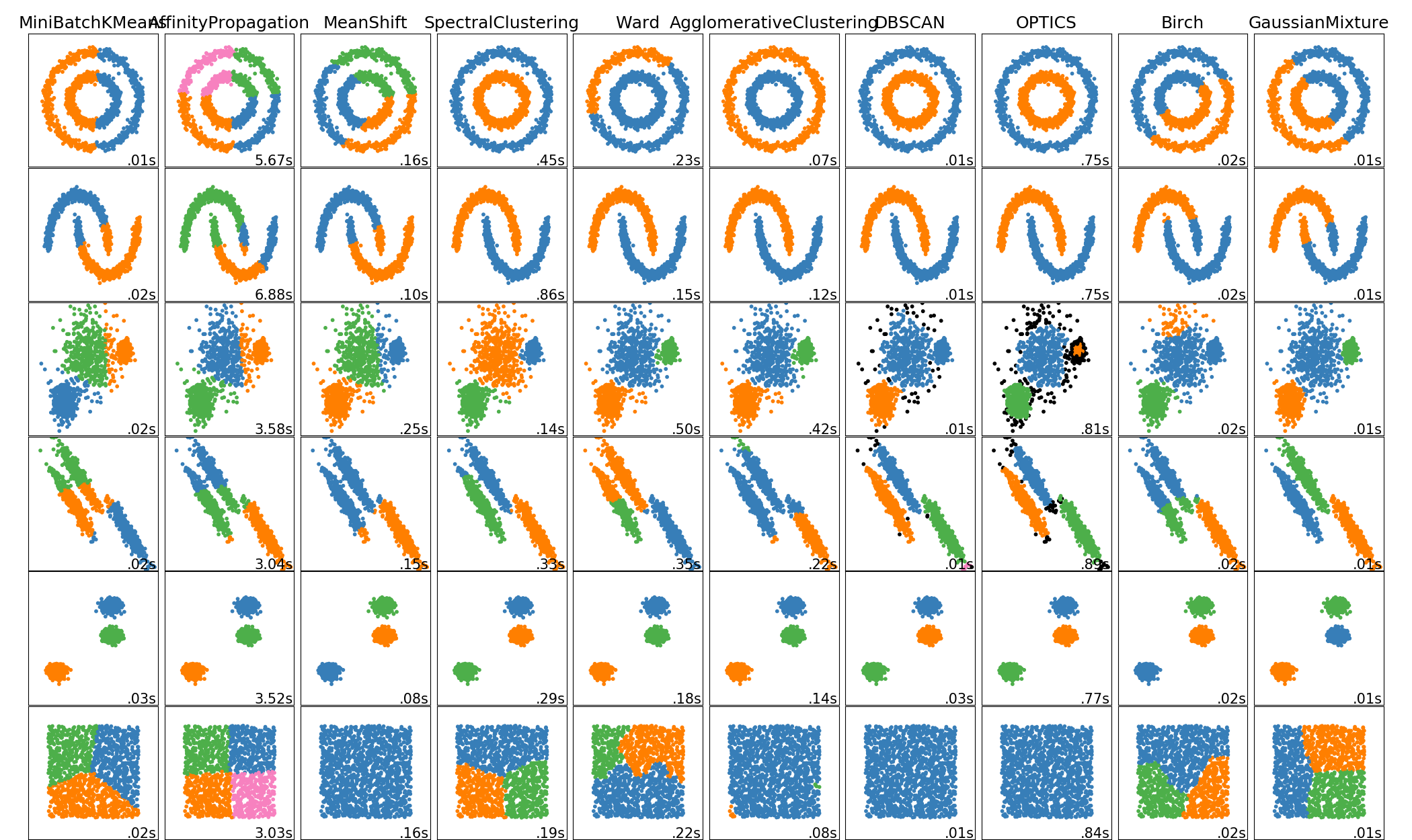
1. נרצה למצוא מבנה סטטיסטי שמתאים לנתונים.
2. ברצוננו למצוא איזושהי חוקיות/מבנה של ה-data, אשר יאפשר לנו להסיק תכונות של דוגמה חדשה ששייכת לאיזושהי קבוצה (cluster) על סמך תכונות של דוגמאות אחרות ששייכות לאותה קבוצה.

Who is your favorite x men? character?

Wolverine  
Storm

קיימים אלגוריתמים רבים לביצוע אשכול. כל אלגוריתם יוביל לתוצאת אשכול שונה.

בתיעוד של חבילת scikit-learn של פיתון, ניתן לראות השוואה בין התוצאות שאליהם מגיעים אלגוריתמי אשכול שונים על dataset פשוט דו מימדי:



התמונה לקוחה מתוך:

<https://scikit-learn.org/stable/_images/sphx_glr_plot_cluster_comparison_001.png>

בקורס אנחנו נתמקד באלגוריתם אשכול ספציפי: אלגוריתם K-Means.

# אלגוריתם K-Means

זהו אלגוריתם איטרטיבי אשר מחלק וקטורים נתונים ל- קבוצות, ולכל קבוצה מוצא נקודת "מרכז מסה" שמייצגת את הקבוצה כולה (נקודה מרכזית או סנטרואיד, centroid). בינתיים נניח כי מספר הקבוצות  נתון מראש.

סימונים:  הוא הסנטרואיד של קבוצה  ().

האלגוריתם:

* אתחול : בחירת  נקודות מרכזיות  עבור הזמן.
* התהליך:
  1. סיווג הנקודות הקיימות באמצעות אלגוריתם  ביחס לסנטרואידים. כלומר, נקודה  שייכת לקבוצה  אם

.

נניח כי במקרה של שוויון תמיד נבחר את הקבוצה עם האינדקס הקטן.

* 1. מציאת הסנטרואידים החדשים: ,

כאשר  הוא מספר האיברים בקבוצה ה-.

אם , .

* 1.  וחזרה לשלב 1 עד להתכנסות ( לכל ).

האלגוריתם שואף למזער את סכום השגיאות הריבועיות: .

לאלגוריתם זה מובטחת התכנסות למינימום מקומי. בניסיון מתקבל שתהליך זה עמיד  לתנאי ההתחלה, אך מומלץ לבחור את תנאי ההתחלה על פי מחשבה ושימוש במידע מקדים במידת האפשר.

# קביעת מספר הקבוצות ע"פ המידע עצמו

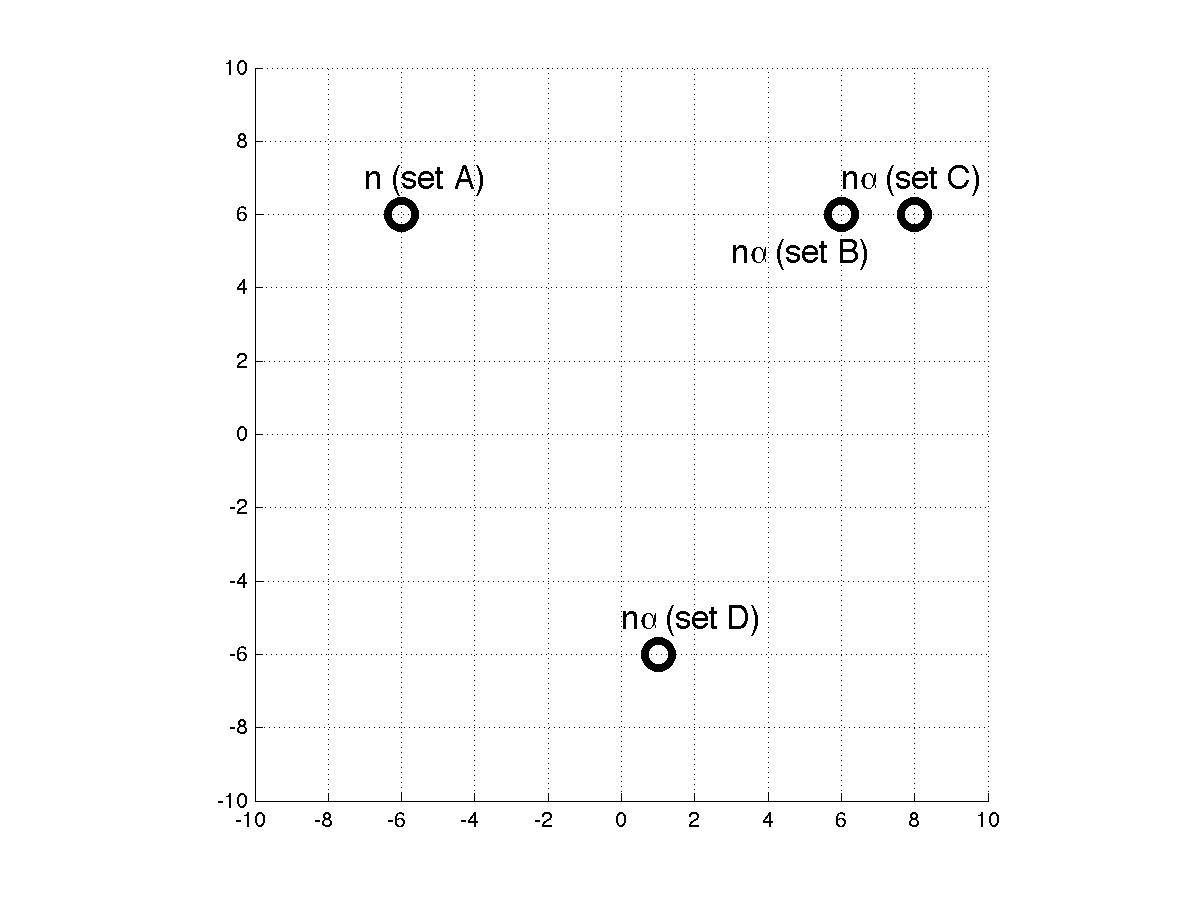
לעתים קרובות, לא יודעים מראש את ערכו של **** ומעוניינים לקבוע **** "טבעי", תוך כדי ריצת האלגוריתם.

נגדיר את שגיאת האִשכול .

ככל שמגדילים את, השגיאה קטנה. ניתן לראות שאם נהפוך כל נקודה לאשכול עצמאי, השגיאה תהיה אפס, אבל לא הרווחנו שום ידע. אחת השיטות לקביעת מספר קבוצות "טבעי" היא ע"י הגדלה הדרגתית של ****, וחישוב את  בסיום כל שלב. עוצרים את התהליך ב-**** שנותן (למשל) , כאשר  הוא סף מסוים.

# תרגילים

## שאלה 1

נתונות  נקודות שונים. לדוגמא, ישנן n נקודות בקואורדינטות (-6,+6). רוצים לקבץ את הנקודות ל-3 קבוצות.

1. מאתחלים את האלגוריתם על ידי בחירת שלוש מתוך ארבע הקבוצות, וקביעת המרכז להיות מיקום הנקודות בקבוצה. לאילו חלוקות יתכנס האלגוריתם בכל אחד מארבעת האתחולים האפשריים.
2. איזו חלוקה של הנקודות לשלוש קבוצות מקבלת ערך מינימלי של פונקציית המטרה. ניתן להניח שכל הנקודות באותו מיקום משויכות לאותו אשכול. רשמו את הפתרון כפונקציה של הפרמטר .
3. האם קיים אתחול עבורו האלגוריתם לא יתכנס לפתרון בעל הערך המינימלי שמצאתם בסעיף הקודם? הדגימו.

## פתרון

1. אם מאתחלים בנקודות A,B,C אזי הנקודות ב-D ישויכו למרכז שבקבוצה B, מרכז זה יזוז למרכז הכובד של הקבוצות B,D, בשלב הבא הנקודות ב-B ישויכו למרכז ב-C שקרוב יותר ממרכז הכובד, ולכן לנקודות ב-D תהיה Cluster משל עצמן, והנקודות ב-B,C ישויכו למרכז הכובד שלהן.

אם מאתחלים בנקודות A,B,D אזי הנקודות ב-C ישויכו למרכז שבקבוצה B, ולכן לנקודות ב-D תהיה Cluster משל עצמן, והנקודות ב-B,C ישויכו למרכז הכובד שלהן.

אם מאתחלים בנקודות A,C,D אזי הנקודות ב-B ישויכו למרכז שבקבוצה C, ולכן לנקודות ב-D תהיה Cluster משל עצמן, והנקודות ב-B,C ישויכו למרכז הכובד שלהן. כל הפתרונות האחרונים זהים.

אם מאתחלים בנקודות B,C,D אזי הנקודות ב-A ישויכו למרכז שבקבוצה B, והמרכז יזוז למרכז הכובד שלהן. בשלב הבא הנקודות ב-B ישויכו למרכז C או למרכז הכובד שלהן, תלוי מי יותר קרוב. אם ישויכו למרכז C נגיע לפתרונות לעיל, אם לא, לא יהיה שינוי. כל החישובים נעשים על הציר  . נדון עתה רק בציר ה-X. מרכז הכובד של הנקודות ב-A ו-B הינו . נקודת המעבר שבה הנקודות ב-B ישויכו למרכז שבקבוצה C החדש מקיימת 

1. מן הניתוח לעיל עולה כי יש שני פתרונות אפשרייםA,(B,C),D או (A,B),C,D. ההבדל בערך פונקציית המטרה הוא  , כאשר  הוא ערך קואורדינטת ה-x של המרכז כאשר הקבוצות A,B נמצאות יחד. ראשית נמצא אותו על ידי גזירה של פו' המטרה למקרה של חלוקה ל (A,B),C,D (אפשר גם ע"י חישוב ממוצע משוקלל של הנקודות ב A,B), מתקיים 

נציב בהפרש פונקציית המטרה



הביטוי מתאפס כאשר  . עבור ערכים גדולים יותר הפתרון הינו (A,B),C,D, וקטנים יותר הפתרוןA,(B,C),D .

כלומר, במידה והמרכזים ההתחלתיים הן B,C,D:

עבור  האלגוריתם ישדך את C ו-B וזהו הפתרון האופטימלי גלובלית, עבור  האלגוריתם ישדך את B ו-A, וזהו הפתרון האופטימלי גלובלית. כאשר  האלגוריתם ישדך את A ו-B אולם זהו אינו הפתרון האופטימלי גלובלית.

עבור בחירה אחרת של המרכזים ההתחלתיים נקבל את האשכולות A,(B,C),D , וכאמור, זהו הפתרון הגלובלי עבור .

1. אפשר לאתחל שני מרכזים להיות מאוד רחוקים, ואז כל הנקודות ישויכו למרכז השלישי.

## תרגיל (קריאה עצמית)

נתבונן בבעיית "האשכול" החד-מימדית הבאה:

clustering0

כאשר הנקודות  ממוקמות באופן אחיד באינטרוול  ומספרן  (וכמובן ).

הראו כי האלגוריתם  עם  מתכנס למינימום הגלובלי של השגיאה הריבועית **מכל תנאי התחלה** **סביר** (כלומר, המרכזים ההתחלתיים ממוקמים באינטרוול ).

## פתרון

נסמן ב את נקודת ההחלטה באיטרציה *n* וב- את המרכזים באיטרציה *n*. מהנתון, בקירוב הרצף:



עבור  כלשהו.

באיטרציה הראשונה, נקבל



ובאופן כללי,



נפתור את הרקורסיה:

