# תרגול 13: אישכול

## אלגוריתם K-Means

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

<u>סימונים</u>: $\boldsymbol{\mu}_i$ הוא הסנטרואיד של קבוצה $G_i$ ($i = 1,...,K$).

<u>האלגוריתם</u>:

- אתחול: בחירת $K$ נקודות מרכזיות $\left\{\boldsymbol{\mu}_i^{(0)}\right\}_{i=1}^{K}$. $t = 0$.
- התהליך:

1. סיווג הנקודות הקיימות באמצעות אלגוריתם 1-NN ביחס לסנטרואידים. כלומר, נקודה $\mathbf{x}$ שייכת לקבוצה $G_i^{(t)}$ אם
$$i = \underset{j=1,...,C}{\arg\min}\left\{\left\|\mathbf{x} - \boldsymbol{\mu}_j^{(t)}\right\|\right\}.$$
נניח כי במקרה של שוויון תמיד נבחר את הקבוצה עם האינדקס הקטן.
2. מציאת הסנטרואידים החדשים: $\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{\left|G_i^{(t)}\right|}\sum_{\mathbf{x}\in G_i^{(t)}} \mathbf{x}$,
כאשר $\left|G_i^{(t)}\right|$ הוא מספר האיברים בקבוצה ה-$i$.
אם $\left|G_i^{(t)}\right| = 0$, $\boldsymbol{\mu}_i^{(t+1)} = \boldsymbol{\mu}_i^{(t)}$.
3. $t \leftarrow t+1$ וחזרה לשלב 1 עד להתכנסות ($\boldsymbol{\mu}_i^{(t+1)} \approx \boldsymbol{\mu}_i^{(t)}$ לכל $i$).

האלגוריתם שואף למזער את סכום השגיאות הריבועיות: $\sum_{i=1}^{K}\sum_{\mathbf{x}\in G_i}\left\|\mathbf{x} - \boldsymbol{\mu}_i\right\|^2$.

לאלגוריתם זה מובטחת התכנסות ל<u>מינימום מקומי</u>. בניסיון מתקבל שתהליך זה עמיד ($robust$) לתנאי ההתחלה, אך מומלץ לבחור את תנאי ההתחלה על פי מחשבה ושימוש במידע מקדים במידת האפשר.

### קביעת מספר הקבוצות $K$ ע"פ המידע עצמו

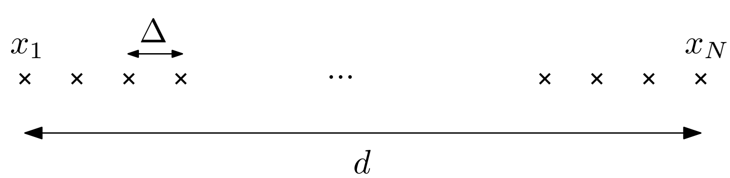
לעתים קרובות, לא יודעים מראש את ערכו של $K$ ומעוניינים לקבוע $K$ "טבעי", תוך כדי ריצת האלגוריתם.

נגדיר את שגיאת האִשכול $E(K) = \sqrt{\sum_{i=1}^{K} \sum_{\mathbf{x} \in G_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2}$ .

ככל שמגדילים את $K$, 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 $K$, וחישוב את $E(K)$ בסיום כל שלב. עוצרים את התהליך ב-$K$ שנותן (למשל) $1 - \frac{E(K)}{E(K-1)} < \varepsilon$, כאשר $\varepsilon$ הוא סף מסוים.

### תרגיל

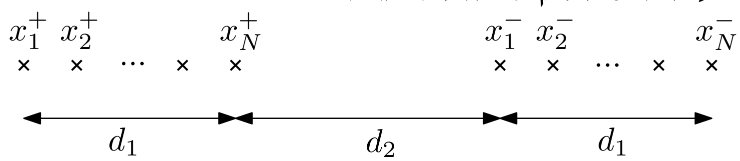
נתבונן בבעיית "האשכול" החד-מימדית הבאה:




כאשר הנקודות $\left\{x_j\right\}_{j=1}^{N}$ ממוקמות באופן אחיד באינטרוול $[0,d]$ ומספרן $N \to \infty$ (וכמובן $\Delta \to 0$).

א. הראו כי האלגוריתם K-Means עם $K = 2$ מתכנס למינימום הגלובלי של השגיאה הריבועית **מכל תנאי התחלה סביר** (כלומר, המרכזים ההתחלתיים ממוקמים באינטרוול $[0,d]$).

נתבונן כעת בבעיית אשכול קצת יותר אמתית:




כלומר נתונים שני אשכולות $\left\{x_j^+\right\}_{j=1}^N$ , $\left\{x_j^-\right\}_{j=1}^N$, כאשר הנקודות בכל אשכול ממוקמות באופן אחיד, ושוב נניח כי $N \to \infty$.

ב. היעזרו בסעיף א׳ על מנת להראות כי גם במקרה זה האלגוריתם K-Means עם $K = 2$ מתכנס למינימום הגלובלי של השגיאה הריבועית מכל תנאי התחלה **סביר**.

ג. האם התוצאות של סעיפים א׳ וב׳ עדיין תקפות **לכל תנאי התחלה** של המרכזים ב- $\mathbb{R}$ ?