12: Clustering Algorithms

הגדרות

האיברים של $[n]$ מייצגים יחידות (או נקודות) מידע. זה לא חייב להיות במרחב אוקלידי – צריך רק שיהיה מושג של מרחק בין 2 נקודות, ושיתקיים:

$D := (d_{ij}) \in M_{n\times n}$ היא מטריצת המרחקים. d_{ij} זה המרחק בין נקודה i לנקודה j.

לכל $i,j,k \in [n]$ צריך להתקיים: $d_{ij} \geq 0$, וגם $d_{ij} + d_{jk} \geq d_{ik}$.

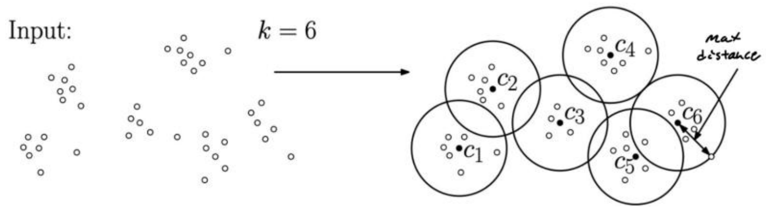
אפשר לראות את D בתור גרף מלא, שקודקודיו הם $[n]$, והמשקל של כל צלע ij הוא d_{ij}.

עבור $i \in [n], S \subseteq [n]$, נגדיר $d(i,S)$ את המרחק של הנקודה i מהקבוצה S: $\min_{j\in S} d_{ij}$. אם $i \in S$, אז $d(i,S) = 0$.

The K-centers Problem

בהינתן מספר $k \in \mathbb{N}$, קבוצה $[n]$, ומטריצת מרחקים $D := (d_{ij})$ שמקיימת את אי"ש המשולש,

נרצה למצוא $S \subseteq [n], |S| = k$ שממזערת את $\max_{i\in[n]} d(i,S)$.



נהוג לקרוא ל-$\max_{i\in[n]} d(i,S)$ **הרדיוס** של S. הנקודות של S הן *centroids* שמסביבן נרצה לקבץ את שאר הנקודות.

בעיית *k-center* היא *NPH*. להלן פסודו-קוד לאלגוריתם חמדן:

1. נתחיל עם S שמכילה קודקוד שרירותי.
2. כל עוד $|S| < k$, נמצא את הקודקוד שהכי רחוק מ-S ונוסיף אותו ל-S.

טענה: האלגוריתם הזה הוא 2-מקרב עבר בעיית *k-center*.

אלגוריתם *f*-מקרב

תהי בעיית מינימיזציה Π כלשהי, ויהי A אלגוריתם עבור הבעיה. האלגוריתם ייקרא *f*-קירוב (*f-approximation*) עבור Π אם:

$$A(I) \leq f \cdot OPT(I)$$

לכל מופע I של Π. כלומר, אם הערך ש-A מחזיר הוא לכל היותר f פעמים הערך האופטימלי.

למה הקירוב כפלי (כפול f) ולא חיבורי (ועוד f)? כי אין כמעט בעיות שעומדות בתנאי הזה, אז זה לא שמיש.

אותה הגדרה עובדת עבור בעיית מקסימום, אבל f יהיה שבר כלשהו (קטן מ-1) והסימן יהיה הפוך: $A(I) \geq f \cdot OPT(I)$.

אז עבור בעיית *k-center*, נאמר שהאלגוריתם הוא 2-מקרב. כלומר, אם:

$$OPT(I) := \min_{S\subseteq[n],|S|=k} \left(\max_{i\in[n]}\big(d(i,S)\big)\right)$$

אנחנו טוענים שאם $S \subseteq [n], |S| = k$ היא הפלט של האלגוריתם החמדן, אז:

$$\max_{i\in[n]}\big(d(i,S)\big) \leq 2 \cdot OPT(I)$$

כלומר המרחק המקסימום מהקבוצה לנקודה כלשהי, הוא לכל היותר פעמיים המרחק המקסימום שמתקבל באלגוריתם האופטימלי.

הוכחה: תהי $S^* := \{j_1, j_2, \ldots, j_k\} \subseteq [n]$ שהיא פתרון אופטימלי. נסמן $r^* := \max_{i\in[n]}\big(d(i,S^*)\big)$ את הרדיוס של הפיתרון.

הקבוצה S^* מגדירה חלוקה על הדאטא: $V_1 \cup V_2 \cup \ldots \cup V_k = [n]$ כך:

נשים נקודה $\ell \in [n]$ ב- V_i אם מתוך $\{j_1, j_2, \ldots, j_k\}$, היא הכי קרובה ל- j_i. (אם זה תיקו נבחר שרירותית ביניהם).

12: Clustering Algorithms

אבחנה: נביט במקבץ (*cluster*) ה-i עבור $i \in [k]$, ויהיו $x, y \in V_i$. אזי, $d_{xy} \le 2 \cdot r^*$. למה? כי:

$$d_{xy} \le d_{xj_i} + d_{j_i y} \le 2r^*$$

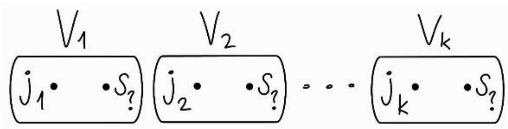
המרחק בין x ל-y הוא לכל היותר סכום המרחקים בין x, y למרכז של המקבץ. שהוא לכל היותר פעמיים הרדיוס.

אז המרחק בין 2 נקודות בתוך מקבץ, הוא לכל היותר פעמיים הרדיוס.

יהי $S := \{s_1, s_2, \dots, s_k\} \subseteq [n]$ הפתרון החמדן, ונסמן $r := \max_{i \in [n]}\big(d(i, S)\big)$ את הרדיוס של הפתרון החמדן. אנחנו רוצים להוכיח: $r \le 2 \cdot r^*$.

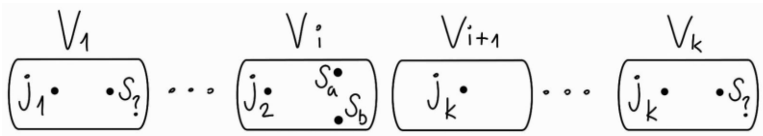
נחלק לשני מקרים:

במקרה הראשון: $|S \cap V_i| = 1$ לכל $i \in [k]$. כלומר, בכל קבוצה לפי החלוקה של האופטימלי, יש נקודה אחת מהמרכזים של החמדן:



במקרה הזה, לכל נקודה שנבחר, היא נמצאת במקבץ עם נקודה מ-S. אז לפי האבחנה לעיל, הטענה מתקיימת.

במקרה השני, קיים $i \in [k]$ כך ש $|S \cap V_i| \ge 2$. כלומר יש קבוצה אחת בחלוקה של האופטימלי, שיש בה לפחות 2 מרכזים של החמדן:



יהיו $s, s' \in S$ בתוך V_i.

בה"כ, החמדן בחר קודם את s ואז את s'.

בגלל שהן באותו מקבץ, לפי האבחנה לעיל, $d_{ss'} \le 2 \cdot r^*$.

כאשר s' נבחר ע"י החמדן, הוא היה הנקודה הרחוקה ביותר מ-S הנוכחית. ו-s כבר הייתה ב-S.

כלומר, כל שאר הנקודות הן במרחק לכל היותר $2r^*$ מ-S, כנדרש.

The k-suppliers Problem

בהינתן $A \cup B := [n]$, כך ש $|A| \ge k$ עבור $k \in \mathbb{N}$ כלשהו. A מייצגת שירותים או ספק כלשהם, ו-B היא הלקוחות.

נרצה למצוא $S \subseteq A, |S| = k$ שממזערת את $\max_{b \in B}\big(d(b, S)\big)$.

נציע את האלגוריתם הבא:

עבור לקוח $b \in B$, נסמן $a_b \in A$ את הספק הכי קרוב.

נפעיל את *k-center* (אלגוריתם 2-מקרב) על B. תהי S' הקבוצה המוחזרת.

נחזיר את $S := (a_b : b \in S')$.

אם $|S| < k$, נוסיף ל-S נקודות באקראי.

כלומר, נקבל k צרכנים שהם פיזור טוב בין שאר הצרכנים, וניקח את הספק שהכי קרוב לכל אחד מהם.

טענה: האלגוריתם המוצע הוא 3-מקרב.

הוכחה: נסמן $r^* := \min_{S^* \subseteq A, |S^*| = k}\left(\max_{b \in B}\big(d(b, S^*)\big)\right)$ את הרדיוס האופטימלי עבור *k-suppliers*.

בפרט, אנחנו מניחים ש**קיימת** קבוצה $Y \subseteq A, |Y| = k$ כך שכל נקודה ב-B קרובה אליה עד כדי r^* (זה הפיתרון האופטימלי).

ניקח $b \in B$ כלשהו. צ"ל: $d(b, S) \le 3r^*$.

<u>מקרה ראשון</u>: קיים $b' \in S'$ כך ש: $d_{bb'} \leq 2r^*$. (כלומר לכל b שנבחר, יש b' ב-S' שקרוב אליו עד כדי $2r^*$).

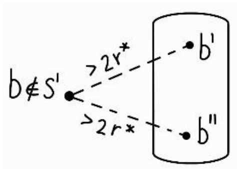
אזי לפי אי"ש המשולש, מתקיים:

$$d_{ba_{b'}} \leq d_{bb'} + d_{b'a_{b'}} \leq 3r^*$$

כי $d_{bb'} \leq 2r^*$ לפי ההנחה. ו- $d_{b'a_{b'}} \leq r^*$ כי $a_{b'}$ היא הנקודה הכי קרובה ל- b' מתוך כל A. אז $d(b,S) \leq 3r^*$, כנדרש.

<u>מקרה שני</u>: $d_{bb'} > 2r^*$ לכל $b' \in S'$. אזי, בפרט $b \notin S'$. ניזכר שהחמדן תמיד לקח את הנקודה הכי רחוקה מ-S' הנוכחית.

אם $b \notin S'$ וגם $d(b,S') > 2r^*$:



אז זה אומר שהמרחק בין כל 2 נקודות ב-S' הוא יותר מ- $2r^*$. למה?

כי כשבחרנו את b'', היא הייתה הנקודה הרחוקה ביותר מכל שאר הנקודות ב-S'.

אז בפרט היא הייתה רחוקה יותר מ- b' מאשר b.

נשים לב שמתקיים: $|S' \cup \{b\}| = k+1$. וגם, בקבוצה הזו, כל שתי נקודות הן במרחק לפחות $2r^*$.

כי המרחק בין כל 2 נקודות ב-S' הוא יותר מ- $2r^*$, וגם b רחוקה לפחות $2r^*$ מכל נקודה ב- S'.

כלומר, קיימת קבוצה $X \subseteq B, |X| = k+1$ כך ש: $d_{uv} > 2r^*$ לכל $u, v \in X$.

(קבוצה ב-B שכל האיברים שלה רחוקים לפחות $2r^*$ אחד מהשני).

אזי, אין $Y \subseteq A, |Y| = k$ שיכולה לתת איבר a שקרוב עד כדי r^* לכל $b \in X$. למה?

נב"ש שיש קבוצה Y כזו. מכיוון ש $|X| = k+1$, אז לפחות שני איברים של X, נגיד u, v מקיימים:

$$d(u,Y) = d_{ua}, \qquad d(v,Y) = d_{va}$$

כלומר המרחק שלהם מ-Y נקבע לפי אותה נקודה ב-Y, כי $|Y| = k$ (שובך היונים).

וגם, $d_{ua}, d_{va} \leq r^*$ כי הנחנו ש-Y יכולה לתת שירות ל-X. ובסה"כ, נקבל:

$$d_{uv} \leq d_{ua} + d_{av} \leq 2r^*$$

סתירה.

כלומר אין קבוצה ב-A שנותנת נקודה שקרובה עד כדי r^* לכל נקודה ב-B. סתירה להנחה מההתחלה – בעצם הראינו שהפתרון האופטימלי לא אפשרי.

שזו סתירה, כי התחלנו בהנחה שהפתרון האופטימלי נותן רדיוס r^*.

כלומר המקרה השני לא אפשרי, אז רק המקרה הראשון מתקיים. מש"ל.