

# ביולוגיה חישובית – מטלה 3 – רשתות נוירונים – דו"ח תרגיל

שלומי בן שושן, ת"ז 311408264

## שאלה 1

איך חיבתם את המרחק בין ישוב בקלט לבין התא בגריד שמייצג אותו?

**פתרון:**

באמצעות נוסחת  $RMSD$ . בהינתן וקטור  $V = (V_1, \dots, V_d)$  שמתאר את ההצבעות בישוב בקלט, ובהינתן וקטור  $N = (N_1, \dots, N_d)$  שמתאר את הנוירון של התא שמייצג את  $V$ , האלגוריתם מחשב:

$$distance(V, N) = RMSD(V, N) = \sqrt{\frac{1}{d} \sum_{i=1}^d (V_i - N_i)^2}$$

## שאלה 2

באיזה אופן קירבתם את התא אליו מופה קלט מסוים לאותו קלט ובאיזה אופן שונו שכניו?

**פתרון:**

באמצעות כלל העדכון שהוצע בהנחיות לתרגיל. בהינתן וקטור קלט  $V_s$  שמיוצג ע"י הנוירון  $N_k$ , תהי קבוצה  $L_1(N_k)$  שמכילה את השכנים המידיים של  $N_k$ , ותהי קבוצה  $L_2(N_k)$  שמכילה את השכנים המידיים של כל תא ב- $L_1(N_k)$ , אך לא מכילה את  $N_k$  ולא אף תא מ- $L_1(N_k)$ .

כלומר:

$$L_1(N_k) := \{N_i \mid N_i \in adj(N_k)\}$$

$$L_2(N_k) := \left\{ N_i \mid N_i \in \left( \left( \bigcup_{N_j \in L_1(N_k)} L_1(N_j) \right) - \{N_k\} - \{N_g \mid N_g \in L_1(N_k)\} \right) \right\}$$

ואפשר גם לסמן:

$$L_0(N_k) := \{N_k\}$$

כעת, האלגוריתם מחשב את הנוירונים  $N_i \in L_0(N_k) \cup L_1(N_k) \cup L_2(N_k)$  עבור איפוק  $t + 1$  ע"י:

$$N_i^{t+1} = \begin{cases} 0.3 \cdot V_s + 0.7 \cdot N_i^t, & N_i \in L_0(N_k) \\ 0.2 \cdot V_s + 0.8 \cdot N_i^t, & N_i \in L_1(N_k) \\ 0.1 \cdot V_s + 0.7 \cdot N_i^t, & N_i \in L_2(N_k) \end{cases}$$

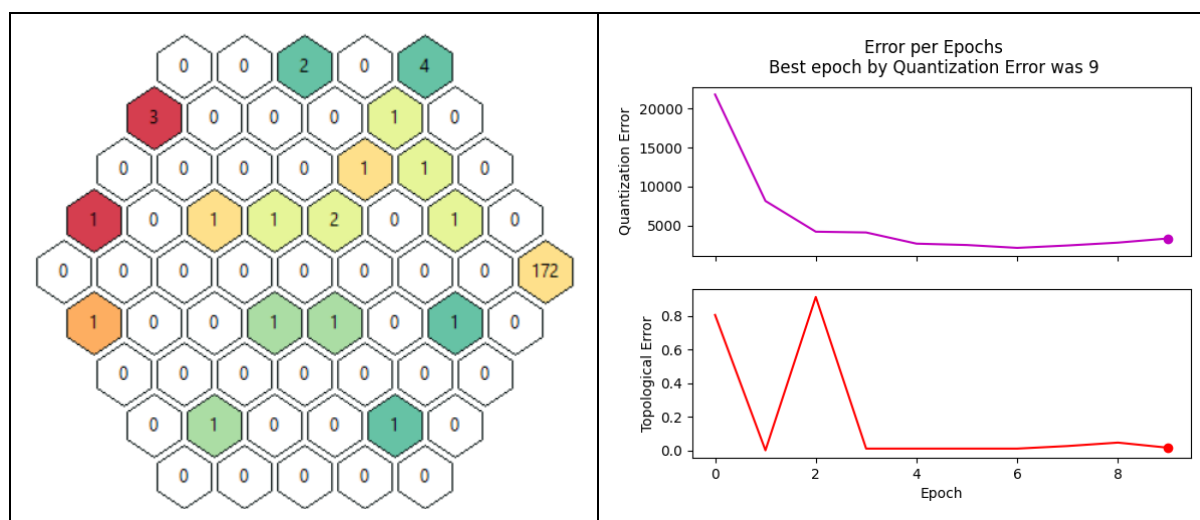
### שאלה 3

האם זה משנה באיזה סדר מציגים את הישובים למערכת?

#### פתרון:

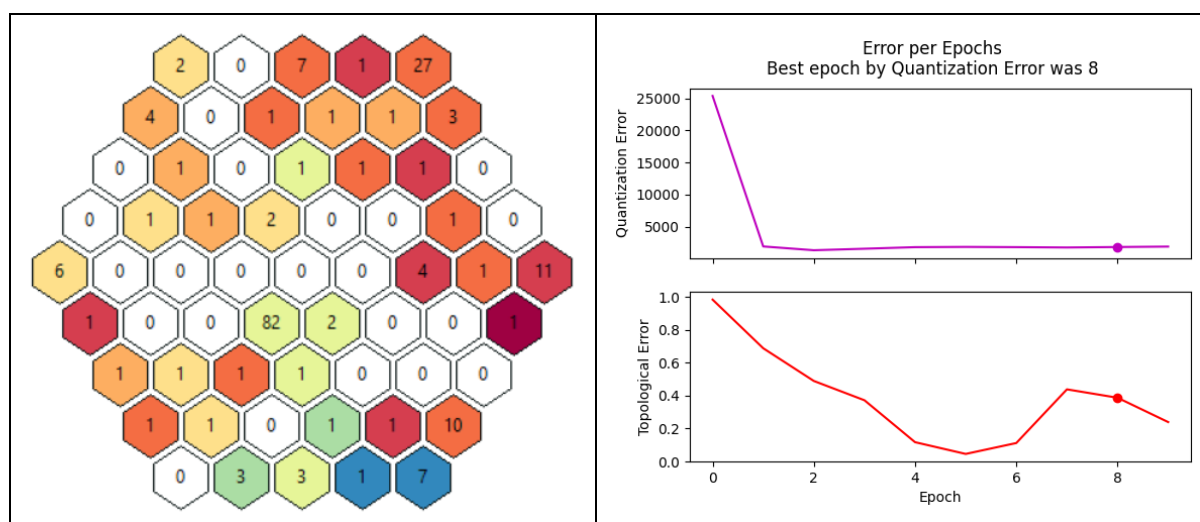
התשובה היא כן. לצורך המענה על השאלה נערך ניסוי בו האלגוריתם הורץ פעמיים, כאשר לפני כל ריצה וקטורי הקלט מוינו בסדר שונה. האלגוריתם הורץ במשך עשרה איפוקים, והפתרון שנבחר היה הפתרון עבורו חושבה ה-Quantization Error הנמוכה ביותר. עבור הפתרון הנבחר חושבה גם Topological Error והופקו תרשימים לטובת הערכת ההשפעה של סדר הקלטים.

ריצה 1: מיון וקטורי הקלט בסדר עולה על-פי סכום ערכיהם (שמתאר את מספר ההצבעות הכולל בישוב).



ערכי השגיאות שחושבו הם:  $QE = 3327.6833$ ,  $TE = 0.0153$ .

ריצה 2: מיון וקטורי הקלט בסדר יורד על-פי סכום ערכיהם (שמתאר את מספר ההצבעות הכולל בישוב).



ערכי השגיאות שחושבו הם:  $QE = 1809.5090$ ,  $TE = 0.3878$ .

#### הערות על ההצגה:

1. המספרים בתאי הרשת ההקסגונית מציינים את מספר היישובים שמיוצגים ע"י אותו תא.
2. כל ישוב בקלט משויך ב-economic cluster שמתואר ע"י מספר בין 1 ל-10. לכל economic cluster התוכנית משייכת צבע על-פי הסקאלה שמופיעה בעמוד הבא. צבע התא ברשת ההקסגונית נקבע ע"י חישוב ה-economic cluster הממוצע המעוגל של כל הישובים המיוצגים ע"י אותו תא.

### Economic Clusters

n/a	1	2	3	4	5	6	7	8	9	10
-----	---	---	---	---	---	---	---	---	---	----

**ניתוח ממצאים:** מתן קדימות ליישובים בהם סך ההצבעות קטן יותר, כלומר לווקטורים שסכום ערכיהם קטן יותר (ריצה 1), גורם לרשת להיות דלילה יותר, ולקיבוץ של יישובים רבים לכדי ייצוג ע"י אותו תא. ניתן לראות שיש ברשת תא שמייצג 172 יישובים! לעומת זאת, מתן קדימות ליישובים בהם סך ההצבעות גדול יותר (ריצה 2), עשוי לגרום לפיזור גבוה יותר ברשת, אך עדיין ניכרת הטיה בתוצאה, שכן שני תאים מייצגים את רוב היישובים (בגדלים 82, ו-27). בהלימה למוצג ברשת, ניתן לראות שה-Quantization Error בריצה השנייה נמוכה יותר. הסיבה לתופעות אלה היא שקירוב הנוירון לווקטור נעשה באחוזים, כך שווקטור עשוי למשוך אליו נוירון יותר מהצפוי במידה ווקטור דומה לו משך את הנוירון אליו רגע לפני כן.

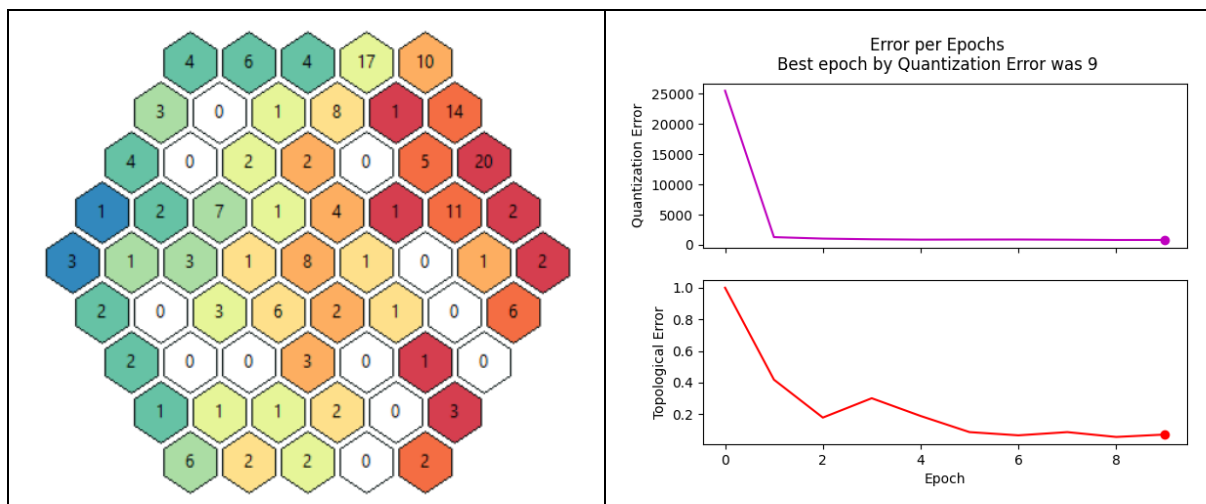
### מסקנה: סידור וקטורי הקלט אכן יוצר הטיה (bias) שמשפיעה על תוצאות האלגוריתם.

#### הערות:

1. בקובץ הקלט הנתון מספר ההצבעות ביישוב לא בהכרח היה זהה לסכום ההצבעות מפלגות באותו ישוב. לכן, האלגוריתם מחשב את הפרש הקולות ומקצה אותו למפלגה פיקטיבית בשם "Others" אשר מייצגת את ההצבעות העודפות (למפלגות שלא מיוצגות בקלט).
2. נעשו חזרות רבות על הניסוי – גם עם מדידת Topological Error, וגם עם שיטות מיון שונות כמו מיון לפי ה-RMS של הווקטורים, לפי השונות (Variance) שלהם, מגדול לקטן ומקטן לגדול. בדו"ח זה הוצג ניסוי מייצג בלבד על-מנת שלא להאריך בעמודים.

**טיפול:** כדי להתמודד עם ה-bias, האלגוריתם הסופי מבצע shuffle על סידור וקטורי הקלט לפני כל איפוק. התוצאות של האלגוריתם המתוקן טובות משמעותית בהשוואה לתוצאות בניסוי לעיל. עניין זה שמתבטא הן ב-Quantization Error והן ב-Topological Error.

#### להלן דוגמת הרצה:



ערכי השגיאות שחושבו הם:  $QE = 741.0064$ ,  $TE = 0.0714$ .

## שאלה 4

כיצד בחרתם את הפתרון להגשה מבין הפתרונות השונים?

### פתרון:

במהלך הניסויים האלגוריתם רץ במשך עשרה איפוקים, כאשר בכל איפוק מתקבל פתרון. הפתרון שנבחר להצגה הוא הפתרון עבורה חושבה ה-Quantization Error הנמוכה ביותר. ניתן לראות למשל שבריצה 2 לעיל הפתרון הנבחר לא היה הפלט של האיפוק האחרון, אלא הפתרון שהתקבל באיפוק מספר 8.

מבחינת האפליקציה המצורפת כחלק מהפתרון, המשתמש יכול לבחור את מדידת השגיאות המועדפת עליו, וגם לשנות את מספר האיפוקים. הפתרון שמוצג יהיה האופטימלי מבין הפתרונות שחושבו על-פי מדידת השגיאות שנבחרה ע"י המשתמש.

