תרגיל מס' 3

מציאת תת-קבוצה של תכונות חשובות ביותר

לספרית sklearn יש תת-ספריה של data. אנחנו ניקח משם את הדאטה של digits.

(ראה למשל: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html)

על דאטה הזה (עשר מחלקות ו-64 תכונות) עליכם למצוא את 5 התכונות החשובות ביותר על סמך 2 אלגוריתמים.

1) האלגוריתם החמדן – נתחיל מקבוצה ריקה וכל פעם מוסיף את התכונה שמובילה לשגיאה הקטנה ביותר
2) MI – שיטה שנלמדה בתרגיל. שימו לב שהערכים של כל תכונה הם בין 0 ל-16. אתם יכולים לחלק את הערכים לשלוש קבוצות אפשריות 0-4, 5-10 ו- 11-16 (ואז יהיו רק שלושה ערכים לכל תכונה – במקום 17 ערכים – ויש בזה גם הגיון לבן, אפור ושחור). כמובן של y יש עשר אפשרויות 0-9.

הפלט של ריצת התוכנית יהיה רשימה של 5 התכונות של האלגוריתם שהשיג את התוצאה הטובה ביותר. השתמשו ב-Logistic Regression עם מימוש ל-k מחלקות! מה שכן עליכם להיעזר הפעם ברגולריזציה! (שדה של penalty ופרמטר C שיקבעו בהתחלה על כל התכונות – יש לשחק עם ערכי C ולמצוא את האופטימאלי וב penalty לבחור את l2).

נא תארו את מסקנותיכם מהתרגיל ומהתכונות!