**הרצאה 8**

Nonparametric Clustering of High Dimensional Data

לא פרמטרי- הנתונים לא מגיעים מהתפלגות ברורה כלשהי (לא בהתפלגות גאוסיינית לדוגמה).

נניח כי יש לנו מכונה שמייצרת נקודות על גרף תלת ממדי, לדוגמה מכונה שמבצעת מדדיות מסוימות. הנקודות הן הערך האמיתי, אך יש רעש כיוון שהחיישן של המכונה לא מדויק לגמרי, לכן נקבל גושים של נקודות.

המערכת עוברת בין המצבים השונים- כל גוש זה מצב.

אנו רוצים לדעת באיזה מצב נמצאת המערכת באמצעות בדיקה באיזה קלאסטר היא נמצאת כרגע.

Chart, scatter chart

Description automatically generated

במציאות הנתונים נראים כך, כי במדידות לפעמים החיישנים נותנים ערכים מוזרים (קורה במציאות).

לכן קשה להשתמש בK means, כי אם נשתמש בK=3 אז הנקודות החריגות שרחוקות ממה שאמור להיות מרכז הקלאסטר יגרמו למרכז הקלאסטר לזוז למקום אחר. כיוון שמשתמשים בממוצע- נקודות מאוד חריגות יזיזו את מרכז הקלאסטר הרבה.

יכול להיות אף מצב בעייתי יותר שהנתונים נראים כך:

Chart, scatter chart

Description automatically generated

לדוגמה פיסקלים צבעוניים כאשר צבע בכל פיקסל נקבע על פי 3 הערכים שהוא מקבל.

LUV- שיטה להגדרת צבעים על פי ערכים, כמו RGB.

נגדיר דמיון בין פיקסלים על פי המרחק בין הנקודות בגרף אשר משקפות ערכים ב LUV.