**הרצאה 6**

Gaussian Mixture Models(GMM)

בנוי מכמה התפלגויות נורמליות

K means הוא מקרה פרטי של שיטה זו.

Soft clustering- כל נקודה שייכת לכל הקלסטרים בהסתברויות שונות וניתן להגיד לכל קלאסטר מה ההסתברות שהנקודה תהיה בו. לכן עבור נקודות אשר נמצאות באמצע בין מספר קלאסים- נוכל להגיד שאנחנו לא יודעים לאיזה קלאסטר הנקודה שייכת, אלא הסתברות של השייכות לקלאסטרים.

בK means אנחנו מחשבים את מרחק הנקודה ממרכז הקלאסטר, אך קלאסטרים הם לא בהכרח עגולים ולכן בכיוון מסוים, נדרש לחשב את המרחק לא בהכרח בצורה כדורית, כמו בK means שבה מניחים שהclassהוא כדורי.

כאשר מחשבים את מרכז הקלאסטר מחדש, מחשבים את המרחק של כל הנקודות ממרכז הקלאסטר, עבור כל קלאסטר על מנת להבין מה ההסתברות שהנקודה שייכת לקלאסטר.

מחשבים את הממוצע על ידי הערך של כל תכונה חלקי מספר התכונות.

מחשבים את סטיית התקן.

לכן קיבלנו 2 שלבים: לחשב את הקלאסטרים לאחר מכן את הפרמטרים של הקלאסטרים.

משתמשים בGMM Expectation Maximization

**בK means**:

השונות היא זהה עבור כל הקלסטרים

Expectation- מחשבים את ההסתברות שכל נקודה שייכת לכל קלאסטר

בוחרים את הקלאסטר עם ההסתברות הגבוהה ביותר שהנקודה שייכת אליו

כיוון שהשונות זהה לכל הקלאסטרים אז מחשבים את המרחק של הנקודה מכל קלאסטר . הקלאסטר שאליו המרחק הכי קצר- אליו משייכים את הנקודה

Maximization- לחשב את המרכז מחדש של כל קלאסטר, מרכז הקלאסטר-ממוצע הערכים בקלאסטר.

מבצעים את תהליך זה שוב ושוב על מנת לשפר את הסיווג לקלאסטרים של הנקודות.

**Soft K means**

מניחים עדיין שהשונות היא אותו דבר עבור כל קלאסטר

נותנים לכל נקודה את ההסתברות שהיא שייכת לכל קלאסטר

תהליך:

מחשבים את ההסתברות של כל נקודה להיות שייכת לכל קלאסטר

מחשבים את ההסתברות של נקודהל היות שייכת לקלאסטר חלקי סכום כל ההסתברויות וכך נקבל וקטור באורך K של משקלים עבור כל נקודה- נקבל ערכים בין 0 ל1 שסכום הערכים בקלאסטר הוא 1.

לאחר מכן נחשב את המרכז של כל קלאסטר- סכום המשקלים של כל הנקודות בקלאסטר זה חלקי מספר הנקודות בקלאסטר

עבור Full GMM

הממוצע שהוא מרכז הקלאסטר הוא הממוצע המשוקלל של הנקודות: סכום הנקודות כפול המשקל של כל נקודה, לחלק לסכום המשקלים של הנקודות.

על מנת לייצר נקודה מקלאסטר ממוצע מבצעים:

*ממוצע הקלאסטר ועוד פונקציית random מהתפלגות גאוסיאנית+סטיית התקן של הקלאסטר*

**Agglomerative Clustering**

*שיטת קלאסטרינג בה לוקחים קלאסטרים קטנים אשר קרובים אחד לשני ומחברים אותם לקלאסטרים גדולים יותר*

*נקודת ההתחלה היא של כל נקודה היא קלאסטר*

*ניתן לעצור כאשר נשארו K קלאסטרים*

*או כאשר יש מרחק מסוים אשר מגדירים מראש לכל קלאסטר*

איך מגדירים מרחק בין קבוצות?

*קיימות מספר גירסאות:*

- closet clusters by centroid: *מחשבים את המרחק בין מרכז של כל קלאסטר*

- closest clusters single link: *מחשבים את המרחק בין הנקודות הקרובות ביותר בכל קלאסטר*

- :closet cluster by complete linkage*מחשבים את המרחק בין הנקודות הכי רחוקות בכל קלאסטר*

Complete *ו*centroid *הם שיטות טובות למציאת המרחק על מנת לאחד בין קלאסטרים*

Divisive Clustering

*מבוסס על תאוריית הגרפים*

*נתייחס לכל נקודה כצומת בגרף*

*כל קשת נותנת את רמת הקירבה בין 2 נקודות, ככל שהן יותר דמות, רמת הקירבה תהיה גדולה יותר*

Aff- *רמת קירבה*

*ניתן להגדיר את רמת הקירבה באמצעות מרחק או באמצעות רמת דמיון ביניהם*

***מטריצת קרבה:***

*מכניסים את כל רמות הקרבה למטריצה A*

*מסדרים את המטריצה כמטריצה סימטרית שהאלכסון- הקרבה בין נקודה לעצמה שווה 1, שאר הערכים בין 0 ל1.*

*מחפשים במטריצה את C שהוא המקסימום סכום ערכי הקרבות על מנת להבין איזה קלאסטר לחלק.*

*במקרה זה, ללא ענישה על קלאסטר גדול האלגוריתם ייקח את כל הנקודות במטריצה.*

*לכן במקום מחלקים את הסכום במספר האלמטנים במטריצה בריבוע- הבעיה היא שהפתרון הכי טוב יהיה לקחת נקודה 1.*

*לכן הפתרון לכך הוא לחלק את סכום ערכי הקרבה בקלאסטר חלקי גודל הקלאסטר*

*כך ככל שנגדיל את הקלאסטר נקבל ערך קרבה גדול יותר(מונה) אך נקבל ענישה על הגדלת הקלאסטר באמצעות חלוקה במסר הנקודות בקלאסטר.*

*נגדיר אלפא שווה 1 ולכן נקבל:*

**

*כאשר הוקטור W הוא וקטור יחידה*

*על מנת לעבור למשתנים סדרתיים ולא דיסקרטיים נגדיר כהערכין בw יכולים להיות כל מספר*

*לכן נקבל:*

*Text

Description automatically generated*

*וקטור עצמי- וקטור שעבורו* הטרנספורמציה או המטריצה מתנהגים כמו סקלר. אינטואיטיבית, השפעת הטרנספורמציה היא "כיווץ" או "מתיחה" של הווקטור, מבלי ש"תזיז" או "תעקם" אותו.

קיבלנו וקטור עם כל מיני מספרים ממשיים איך מזה מגיעים למי נמצא בקלאסטר ומי לא?

*נגדיר סף מסוים שממנו כל הערכים הופכים ל1 וכל הערכים מתחת הופכים ל0*

*כך נחזור למצב של ערכים דיסקרטיים*

*ניתן להזיז את הסף וכל פעם נקודה אחת תעבור מעל או מתחת לסף*

*לכן ניתן לקבל n פתרונות שונים על פי ערך הסף שקובעים.*

*נבדוק איזה סף נותן את הערך המקסימלי של:*

**

*ונבחר אותו.*