אלגוריתם K־MEANS

אלגוריתם להצברה (CLUSTERING) של נקודות ב־$\mathbb{R}^d$ ל־ K צברים. מטרתו היא למצוא את מרכזי הצברים כך שסכום המרחקים הריבועיים בין כל הנקודה ומרכז הצבר אליו שוייכה יהיה הקטן ביותר.

שלבי האלגוריתם:

1. בוחרים באקראי K מרכזים $\mu_1, ..., \mu_k$ השייכים ל־ $\mathbb{R}^d$.
2. משייכים כל נקודה למרכז הקרוב אליה ביותר:

$$c(x_i) = argmin_j \left(||x_i - \mu_j||\right)$$

כאשר $c(x_i)$ משמעותו לאיזה צבר נשייך את x_i (מקבל ערכים מ־1 עד K).

3. מעדכנים כל מרכז להיות הממוצע (מרכז כובד) של כל הנקודות ששויכו אליו:

$$\mu_j = \frac{\sum\limits_{i:c(x_i)=j} x_i}{\sum\limits_{i:c(x_i)=j} 1}$$

המכנה: $\sum\limits_{i:c(x_i)=j} 1$ הוא בעצם מספר הנקודות ששויכו ל־ μ_j.

4. חוזרים על צעדים 3־2 עד שאין שינוי באף אחד מהמרכזים.

קשר בין K־MEANS ל־EM של תערובת גאוסיינים:

אם ב־EM נחליף את $P(h_i = m \mid x_i, \hat{\theta}_t)$ בפונקציה שמקבלת 1 עבור ה־m שמביא את ההסתברות הנ"ל למקסימום ו־0 לשאר נקבל כלל שיוך קשה (CLUSTERING HARD). אם בנוסף נניח שהפריורים (c_j) שווים ושהווריאנס ידוע ושווה למטריצת היחידה $\Sigma = \mathrm{I}$, נקבל את אלגוריתם K־MEANS. במצב כזה אנחנו בעצם ממקסמים את הפונקציה הבאה:

$$log\left[P\left(x_i \mid h_{i,}, \hat{\theta}_t\right)\right] \propto -(x_i - \hat{\mu}_t)^T \underset{\mathrm{I}}{\underset{\|}{\Sigma}}^{-1} (x_i - \hat{\mu}_t) = -\left\|x_i - \hat{\mu}_t\right\|^2$$