Week 3: Lab 3 Machine Learning Part II

Solutions

1 Part A: Understanding K-Means Clustering Algorithm

In this exercise, you will be manually running k-means algorithm on a toy dataset. Consider that you are given a 8 data samples as: R1 (185, 72), R2 (170,56), R3 (168,60), R4 (179,68), R5 (182,72), R6 (188,77), R7 (180,71), and R8 (180,70) and two initial clusters as C1 (185,72) and C2 (170,56).

Q 1: Assign each of data sample (e.g., data rows) to one of the two clusters.
Q 2: Estimate the cluster centroids after each row assignment.

Answer: The data samples are given. Take Euclidean distance between each centroid and observed data samples since the first two data samples (R1 and R2) are assigned as a initial clusters therefore we will start from third data sample (R3) as:

$$\begin{aligned} \text{Distance between R3 \& C1} &= \sqrt{(168-185)^2+(60-72)^2} = 20.80 \\ \text{Distance between R3 \& C2} &= \sqrt{(168-170)^2+(60-56)^2} = 4.47 \end{aligned} \tag{1}$$

So, data sample R3 (168,60) is assigned to cluster C2 (170,56) and cluster C2 is updated as $(\frac{168+170}{2}, \frac{60+56}{2}) = C2(169, 58)$. Similarly, we compute a pairwise distance between the remaining data samples with C1 and updated C2 as:

$$\begin{aligned} \text{Distance between R4 \& C1} &= \sqrt{(179-185)^2+(68-72)^2} = 7.21 \\ \text{Distance between R4 \& C2} &= \sqrt{(179-169)^2+(68-58)^2} = 14.86 \end{aligned} \tag{2}$$

R4 (179,68) is assigned to C1 (185,72) and C1 is updated accordingly as $(\frac{185+179}{2}, \frac{72+68}{2}) = C1(132, 70)$.

$$\begin{aligned} \text{Distance between R5 \& C1} &= \sqrt{(182-132)^2+(72-70)^2} = 50.04 \\ \text{Distance between R5 \& C2} &= \sqrt{(182-169)^2+(72-58)^2} = 19.10 \end{aligned} \tag{3}$$

R5 (182,72) is assigned to C2 (169,58) and C2 is updated accordingly as $(\frac{182+169}{2}, \frac{72+58}{2}) = C2(175.50, 65.0)$. C1 is C1(132,70) and C2 is C2(175.50,65.0)

$$\begin{aligned} \text{Distance between R6 \& C1} &= \sqrt{(188-132)^2+(77-70)^2} = 56.43 \\ \text{Distance between R6 \& C2} &= \sqrt{(188-175.50)^2+(77-65.0)^2} = 17.32 \end{aligned} \tag{4}$$

R6 (188,77) is assigned to C2 (175.50,65) and C2 is updated accordingly as $(\frac{188+175.50}{2}, \frac{77+65}{2}) = C2(181.75, 71.0)$. C1 is C1(132,70) and C2 is C2(181.75,71.0)

$$\begin{aligned} \text{Distance between R7 \& C1} &= \sqrt{(180-132)^2+(71-70)^2} = 48.01 \\ \text{Distance between R7 \& C2} &= \sqrt{(180-181.75)^2+(71-71.0)^2} = 1.75 \end{aligned} \tag{5}$$

R7 (180,71) is assigned to C2 (181.75,71.0) and C2 is updated accordingly as $(\frac{180+181.75}{2}, \frac{71+71}{2}) = C2(180.87, 71.0)$. C1 is C1(132,70) and C2 is C2(180.87,71.0)

$$\begin{aligned} \text{Distance between R8 \& C1} &= \sqrt{(180-132)^2+(70-70)^2} = 48.0 \\ \text{Distance between R8 \& C2} &= \sqrt{(180-180.87)^2+(70-71.0)^2} = 1.32 \end{aligned} \tag{6}$$

R8 (180,70) is assigned to C2 (180.87,71.0) and C2 is updated accordingly as $(\frac{180+180.87}{2}, \frac{70+71}{2}) = C2(180.43, 70.50)$. C1 is C1(132,70) and C2 is C2(180.87,71.0)

So C1={ R1,R4} and C2={ R2,R3,R5,R6,R7,R8} and final clusters are C1={ 132,70 } and C2={ 180.81,71.0 }.