Raeesa Youssef
Jakob Erben
Johannes Sindlinger

Problem Set 2
Mining Massive Datasets, 22/23

2022-11-05

1 Exercise

Feature	A	B	C
Processor Speed	3.06	2.68	2.92
Disk Size	500	320	640
Main-Memory-Size	6	4	6

(a) If $\alpha = \beta = 1$, all features keep the same values when scaling. So the cosinus similarity is given by

$$\begin{aligned} cos(A, B) &= \frac{a \cdot b}{||a|| \cdot ||b||} \\ &= \frac{3.06 \cdot 2.68 + 500 \cdot 320 + 6 \cdot 4}{\sqrt{3.06^2 + 500^2 + 6^2} \cdot \sqrt{2.68^2 + 320^2 + 4^2}} \\ &= 0.9999973 \end{aligned}$$

For A, C and B, C follows using the same calculation:

$$\begin{aligned} cos(A, C) &= 0.9999953 \\ cos(B, C) &= 0.9999878 \end{aligned}$$

(b) For $\alpha = 0.01$ and $\beta = 0.5$ the adapted features yield:

Feature	A	B	C
Processor Speed	3.06	2.68	2.92
Disk Size	5	3.2	6.4
Main-Memory-Size	3	2	3

Using the same formula as above in part a), we obtain the following results:

$$\begin{aligned} cos(A, B) &= 0.9908815 \\ cos(A, C) &= 0.9915547 \\ cos(B, C) &= 0.9691779 \end{aligned}$$

(c) The averages of A, B, C are given as

$$\begin{aligned} avg(A) &= 2.887 \\ avg(B) &= 486.667 \\ avg(C) &= 5.333 \end{aligned}$$

So $\alpha = \frac{1}{486.667} = 0,0021$ and $\beta = \frac{1}{5.333} = 0.1875$. The value for scaling feature A would be $\frac{1}{avg(A)} = \frac{1}{2.887} = 0,3464$. So the feature table changes to

Feature	A	B	C
Processor Speed	1.060	0.9284	1.0115
Disk Size	1.027	0.6575	1.3151
Main-Memory-Size	1.125	0.75	1.125

And the cosine values then result as

$$\begin{aligned} cos(A, B) &= 0.9898552 \\ cos(A, C) &= 0.9915270 \\ cos(B, C) &= 0.9692788 \end{aligned}$$