Assignment - 9
Computing Proximities

HARSH SIDDHAPURA
1230169813

02/25/2024

Similarities Between Binary Data Points

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5
User 1	1	0	1	0	1
User 2	1	0	0	1	0
User 3	0	0	1	0	1

1. **Compute the similarity between User 1 & User 2 using SMC (Simple Matching Coefficient)**
   - Total number of attributes = 5
   - Number of Matching attributes = Dot Product = 1 + 1 + 0 + 0 + 0 = 2
   - Similarity between User 1 & User 2 using SMC = **2/5 = 0.4**

2. **Compute the similarity between User 1 & User 3 using SMC (Simple Matching Coefficient)**
   - Total number of attributes = 5
   - Number of Matching attributes = Dot Product = 0 + 1 + 1 + 1 + 1 = 4
   - Similarity between User 1 & User 2 using SMC = **4/5 = 0.8**

3. **Based on the similarity numbers obtained in 1 & 2, which user is more similar to User 1?**
   Based on the Simple Matching Coefficient (SMC) calculations, User 3 is more similar to User 1. This is because the similarity between User 1 and User 3 is 0.8, which is higher than the similarity between User 1 and User 2, which is 0.4. The higher the SMC, the more similar the users are. Therefore, User 3 is more similar to User 1.

4. **Repeat steps 1, 2 & 3 using the Jaccard Coefficient.**

    1. **Compute the similarity between User 1 & User 2**
       - Intersection of Sets = {1} = 1
       - Union of Sets = {1, 3, 4, 5}= 4
       - Similarity between User 1 & User 2:
         - Intersection/Union = **1/4 = 0.25**

2. **Compute the similarity between User 1 & User 3**
    - Intersection of Sets = {3, 5} = 2
    - Union of Sets = {1, 3, 5}= 3
    - Similarity between User 1 & User 3:
        - Intersection/Union = **2/3 = 0.66**

3. **Based on the similarity numbers obtained in 1 & 2, which user is more similar to User 1?**
   Based on the Jaccard Coefficient calculations, User 3 is more similar to User 1. This is because the similarity between User 1 and User 3 is 0.66, which is higher than the similarity between User 1 and User 2, which is 0.25. The higher the Jaccard Coefficient, the more similar the users are. Therefore, User 3 is more similar to User 1.

5. **Which method do you think is more suitable for computing user similarities between two users in this context?**

Given the context of a user-movie dataset where there are tens of thousands of movies and each user typically watches only a few hundred, the Jaccard Coefficient might be more suitable for computing user similarities.

The reason is that the Jaccard Coefficient only considers the set of movies watched by either user and disregards movies that were never watched by any of them. This makes it more robust to the problem of sparsity in large datasets, where the majority of the user-movie pairs have not been watched, leading to many 0's in the data.

On the other hand, the Simple Matching Coefficient (SMC) considers all movies, regardless of whether any of the two users watched them or not. This means it could be heavily influenced by the large number of 0's (unwatched movies), which might not be as informative for determining user similarity in this context.

Therefore, in this specific scenario, the Jaccard Coefficient might provide a more meaningful measure of user similarity. However, the choice of similarity measure should always be guided by the specific characteristics and requirements of your data and task.

Proximities Between Continuous Data Points

	Attribute 1	Attribute 2	Attribute 3
Point 1	3	1	0
Point 2	2	0	1
Point 3	0	1	2

1. **Compute the Cosine Similarity between Point 1 & Point 2**
   - P1.P2 = (3*2) + (1*0) + (0*1) = 6
   - ||P1|| = sqrt(3^2 + 1^2 + 0^2) = sqrt(10)
   - ||P2|| = sqrt(2^2 + 0^2 + 1^2) = sqrt(5)

   - Cos(P1 , P2)
     = P1.P2 / ||P1||.||P2||
     = 6 / sqrt(10) * sqrt(5)
     = **0.848**

2. **Compute the Cosine Similarity between Point 1 & Point 3**
   - P1.P3 = (3*0) + (1*1) + (0*2) = 1
   - ||P1|| = sqrt(3^2 + 1^2 + 0^2) = sqrt(10)
   - ||P3|| = sqrt(0^2 + 1^2 + 2^2) = sqrt(5)

   - Cos(P1 , P3)
     = P1.P3 / ||P1||.||P3||
     = 1 / sqrt(10) * sqrt(5)
     = **0.1414**

3. **Based on the numbers you got in 1 & 2, which point is more similar to Point 1?**
   Based on the Cosine Similarity calculations, Point 2 is more similar to Point 1. This is because the cosine similarity between Point 1 and Point 2 is 0.848, which is higher than the cosine similarity between Point 1 and Point 3, which is 0.1414. The higher the cosine similarity, the more similar the points are. Therefore, Point 2 is more similar to Point 1.

4. **Compute the Euclidean distance between Point 1 & Point 2.**
    - ED(P1,P2) = sqrt((x2-x1)^2 + (y2-y1)^2 + (z2-z1)^2)
    - P1 = (3,1,0)
    - P2 = (2,0,1)

    - ED(P1,P2)
        = sqrt((3-2)^2 + (1-0)^2 + (0-1)^2)
        = sqrt(1+1+1)
        = **sqrt(3)**
        = **1.73**

5. **Compute the Euclidean distance between Point 1 & Point 3.**
    - ED(P1,P3) = sqrt((x3-x1)^2 + (y3-y1)^2 + (z3-z1)^2)
    - P1 = (3,1,0)
    - P3 = (0,1,2)

    - ED(P1,P3)
        = sqrt((3-0)^2 + (1-1)^2 + (0-2)^2)
        = sqrt(9+0+4)
        = **sqrt(13)**
        = **3.605**

6. **Based on the numbers you got in 4 & 5, which point is farther from Point 1?**
    Based on the Euclidean distance calculations, Point 3 is farther from Point 1. This is because the Euclidean distance between Point 1 and Point 3 is 3.605, which is greater than the Euclidean distance between Point 1 and Point 2, which is 1.73. The greater the Euclidean distance, the farther the points are. Therefore, Point 3 is farther from Point 1.