Homework Assignment 4

IT531, Summer 2018

**Question 1**

Consider the rating data set in Table 1:

|  |  |  |
| --- | --- | --- |
| User | Item | Rating |
| u1 | m1 | 2 |
| u1 | m2 | 3 |
| u2 | m1 | 5 |
| u2 | m3 | 2 |
| u3 | m1 | 3 |
| u3 | m2 | 3 |
| u3 | m3 | 1 |
| u4 | m2 | 2 |
| u4 | m3 | 2 |

Table 1

Predict user u1’s rating for item m3 using item based CF. Use cosine similarity.

1. User item matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | M1 | M2 | M3 |
| U1 | 2 | 3 | ? |
| U2 | 5 | ? | 2 |
| U3 | 3 | 3 | 1 |
| U4 | ? | 2 | 2 |

1. calculate the Item-to-item similarity matrix

Sim(m1,m2)

V1 = 2\*U1+ 3\*U3

V3 = 3\*U1+ 3\*U3

cos(v1,v3) = (2\*3 + 3\*3)/sqrt((4 + 9)\*(9+9)) = 0.98

sim(m1,m3)

V2 = 5\*U2 + 3\*U3

V3 = 2\*U2 + 1\*U3

Cos(v2,v3) = (5\*2 + 3\*1)/sqrt((25 + 9)\*(4+1)) = 0.99

Sim(m2,m3)

V3 = 3\*U3 + 2\*U4

V4 = 1\*U3 + 2\*U4

Cos(v3,v4) = (3\*1 + 2\*2)/sqrt((9 + 4)\*(1+4)) = 0.86

1. item-to-item similarity matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | M1 | M2 | M3 |
| M1 | ? | 0.98 | 0.99 |
| M2 | 0.98 | ? | 0.86 |
| M3 | 0.99 | 0.86 | ? |

1. user1 already rate M1(2) and M2(3)

U1 in M3 rating = (2 \* 0.99 + 3 \* 0.86)/(0.99+0.86) = 2.46

**Question 2**

Consider household mower ownership data in Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Household** | **Income (K)** | **Lot Size (K sq ft)** | **Class (Owns a Mower?)** |
| 1 | 55 | 18 | No |
| 2 | 85 | 20 | Yes |
| 3 | 48 | 16 | No |
| 4 | 61 | 21 | No |
| 5 | 87 | 19 | No |
| 6 | 110 | 23 | Yes |
| 7 | 108 | 22 | Yes |
| 8 | 83 | 24 | Yes |
| 9 | 69 | 20 | No |
| 10 | 93 | 18 | Yes |

Table 2

Use k-NN classifier to classify a new household with income 60K and lost size 20K sq ft. Use K=1 and K=3. Use min-max normalization and Euclidean distance.

(1) Normalization

|  |  |
| --- | --- |
| 0.112903226 | 0.25 |
| 0.596774194 | 0.5 |
| 0 | 0 |
| 0.209677419 | 0.625 |
| 0.629032258 | 0.375 |
| 1 | 0.875 |
| 0.967741935 | 0.75 |
| 0.564516129 | 1 |
| 0.338709677 | 0.5 |
| 0.725806452 | 0.25 |

(2) Normalization {60K,20K sq ft} = (0.19,0.5)

(3) Calculate Euclidean distance

|  |  |  |
| --- | --- | --- |
| **Household** | **Euclidean distance** | **Class (Owns a Mower?)** |
| 1 | 0.143145161 | No |
| 2 | 0.403225806 | Yes |
| 3 | 0.443548387 | No |
| 4 | 0.031754032 | No |
| 5 | 0.451108871 | No |
| 6 | 0.947076613 | Yes |
| 7 | 0.836693548 | Yes |
| 8 | 0.620967742 | Yes |
| 9 | 0.14516129 | No |
| 10 | 0.594758065 | Yes |

(4)

K=4, he/she will be the class of household=1, and not owns a mower

K=3, he/she will be the class of household=1/4/9, and not owns a mower

**Question 3**

Consider public utility data set in Table 3 consisting of the scores of two variables on each of seven individuals:

|  |  |  |
| --- | --- | --- |
| **Company** | **Sales** | **Fuel Cost** |
| Arizona | 9,077 | 0.628 |
| Boston | 5088 | 1.555 |
| Central | 9212 | 0.8 |
| Commonwealth | 6423 | 0.7 |
| Edison | 3300 | 2.044 |

Table 3

Use K-Means clustering to group this dataset into two clusters.

Answer:

1. Normalize attributes

|  |  |  |
| --- | --- | --- |
| **Company** | **Sales** | **Fuel Cost** |
| Arizona | 0.977165088 | 0 |
| Boston | 0.302435724 | 0.654661017 |
| Central | 1 | 0.121468927 |
| Commonwealth | 0.528247632 | 0.050847458 |
| Edison | 0 | 1 |

1. First round

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **distance to cluster 1 (0.977165088,0)** | **distance to cluster 2(0.528247632,0.050847458)** | **assigned cluster** |
| Arizona | 0 | 0.451787944 | cluster1 |
| Boston | 0.940128056 | 0.644656367 | cluster2 |
| Central | 0.123596656 | 0.477009107 | cluster2 |
| Commonwealth | 0.451787944 | 0 | cluster2 |
| Edison | 1.39816008 | 1.08624864 | cluster1 |

1. Second round

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **distance to cluster 1 (0.4885,0.5)** | **distance to cluster 2(0.61,0.276)** | **assigned cluster** |
| Arizona | 0.69908004 | 0.458945468 | cluster2 |
| Boston | 0.242013778 | 0.488240085 | cluster1 |
| Central | 0.636265344 | 0.419162259 | cluster2 |
| Commonwealth | 0.450900571 | 0.239292781 | cluster2 |
| Edison | 0.69908004 | 0.94712599 | cluster1 |

1. Third round

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **distance to cluster 1 (0.4885,0.5)** | **distance to cluster 2(0.61,0.276)** | **assigned cluster** |
| Arizona | 1.169044306 | 0.153202579 | cluster2 |
| Boston | 0.229524716 | 0.800278478 | cluster1 |
| Central | 1.10393464 | 0.176860051 | cluster2 |
| Commonwealth | 0.863178646 | 0.306960717 | cluster2 |
| Edison | 0.229524716 | 1.259315843 | cluster1 |

1. Since there are no change, we stop here