IMPLEMENTATION OF ADJUSTED EUCLIDEAN DISTANCE

Presentation by
Name: Dhan Raj Rai
ID: s4063739

REASONS FOR THE AED TO WORK

The Adjusted Euclidean Distance gives priority to the length of vectors:

- The length of the vectors helps to retain the scalar values of the vectors which are the rating values thus adding the extra meaning to the vectors in the ground of their rating value information which otherwise would have caused the loss of actual rating information.
-

Table 1. A fragment of a rating matrix for a recommender system

	Item1	Itme2	Item3	Item4	Item5
User1	1	1	1	1	1
User2	2	2	2	2	2
User3	5	5	5	5	5
User4	1	2	4	3	3
User5	1	2	3	4	5
User6	2	2	3	4	∅
User7	3	∅	∅	∅	∅

Quantitative analysis as follows: whether we use Eq. (8) or Eq. (9), we can discover two groups of arithmetic expression

Group 1

sim(user2, user1) = sim(user2, user3) ,
sim(user4, user1) = sim(user4, user3) ,
sim(user5, user1) = sim(user5, user3) .

REASONS FOR THE AED TO WORK

AED ensures that for different vectors their own dimensional vector spaces are taken into account:

- It helps the vector to maintain its own vector-specific dimensions
- The approach greatly helps to reduce the dimensional bias, suppression, and cohesion that most probably persist due to common dimensional space vectors.
-

Table 1. A fragment of a rating matrix for a recommender system

	Item1	Itme2	Item3	Item4	Item5
User1	1	1	1	1	1
User2	2	2	2	2	2
User3	5	5	5	5	5
User4	1	2	4	3	3
User5	1	2	3	4	5
User6	2	2	3	4	∅
User7	3	∅	∅	∅	∅

Quantitative analysis as follows: whether we use Eq. (8) or Eq. (9), we can discover two groups of arithmetic expression

Group 1

sim(user2, user1) = sim(user2, user3) ,

sim(user4, user1) = sim(user4, user3) ,

sim(user5, user1) = sim(user5, user3) .

Group 2

sim(user6, user5) < sim(user6, user7).

REASONS FOR THE AED TO WORK

AED emphasizes on the co-rated items:

- The emphasis prevents the AED from losing the dimensional effect thereby retaining the different dimension numbers of different dimensional spaces.
- For which, AED introduces case-based approach whereby different case is handled using different techniques.
-

$$sim(u,v) = 1 - \frac{dist(\vec{u},\vec{v})}{dist_{\max}} = 1 - \frac{\sqrt{\sum_{s \in S_{uv}} (r_{u,s} - r_{v,s})^2}}{\sqrt{\sum_{k=1}^{m} (V_{\max} - V_{\min})^2}} \quad ,(10)$$

where S_{uv} denotes the set of all items co-rated by both users u and v, i.e., $S_{uv} = \{s \in S \mid r_{u,s} \neq \varnothing \;\&\; r_{v,s} \neq \varnothing\}$ (S is the set of all

Accordingly, we suggest the aggregate function for predicting unknown rating value $r_{c,s}$ as follow:

$$r_{c,s} = \begin{cases} k \sum_{c' \in \hat{C}} sim(c,c') \times r_{c',s} \quad , & \hat{C}_0 = \varnothing \\ k_0 \sum_{c' \in \hat{C}_0} m_{c'} \times r_{c',s} \quad , & \hat{C}_0 \neq \varnothing \end{cases} , \quad (11)$$

REASONS FOR THE AED TO WORK

AED mitigates the sparsity issues:

- The AED improves the accuracy of recommendations as the AED adjustment ensures that only genuinely similar users influence the recommendations.
- Thus achieving more personalized and relevant recommendations.
-

$$sim(u,v) = 1 - \frac{dist(\vec{u},\vec{v})}{dist_{\max}} = 1 - \frac{\sqrt{\sum_{s \in S_{uv}} (r_{u,s} - r_{v,s})^2}}{\sqrt{\sum_{k=1}^{m} (V_{\max} - V_{\min})^2}} \quad ,(10)$$

where S_{uv} denotes the set of all items co-rated by both users u and v, i.e., $S_{uv} = \{s \in S \mid r_{u,s} \neq \varnothing \,\&\, r_{v,s} \neq \varnothing\}$ (S is the set of all

Accordingly, we suggest the aggregate function for predicting unknown rating value $r_{c,s}$ as follow:

$$r_{c,s} = \begin{cases} k \sum_{c' \in \hat{C}} sim(c,c') \times r_{c',s} \quad , & \hat{C}_0 = \varnothing \\ k_0 \sum_{c' \in \hat{C}_0} m_{c'} \times r_{c',s} \quad , & \hat{C}_0 \neq \varnothing \end{cases} , \quad (11)$$

WORKING MECHANISM OF THE AED

VECTOR LENGTH	AED considers the user vectors' lengths to take the user preferences into account. The adjustments in normalizing the ratings of thus ensuring the differences in rating scales between users are accounted for.
VECTOR DIMENSIONS	AED ensures that different dimensional spaces with varying numbers of dimensions are considered. It helps the AED to provide the user vector specific dimensional spaces.

WORKING MECHANISM OF THE AED

$$sim(u,v) = 1 - \frac{dist(\vec{u},\vec{v})}{dist_{\max}} = 1 - \frac{\sqrt{\sum_{s \in S_{uv}} (r_{u,s} - r_{v,s})^2}}{\sqrt{\sum_{k=1}^{m} (V_{\max} - V_{\min})^2}} \quad ,(10)$$

$S_{uv} = \{s \in S \mid r_{u,s} \neq \varnothing \,\&\, r_{v,s} \neq \varnothing\}$ (S is the set of all items) and $m = |S_{uv}|$ denotes the number of members in S_{uv}. In Eq. (10), $sim(u,v) \in [0,1]$, which represents

rus = rating of user u on item s
rvs = rating of user v on item s
sim(u,v) = similarity between user u and v

Where,
Vmax = highest rating value
Vmin = lowest rating value
u = rating vector of user u
v = rating vector of user v
dist(u, v) = Euclidean distance between u and v
distmax = maximal Euclidean distance
Suv = set of all items co-rated by both users u and v
m = number of members in Suv

WORKING MECHANISM OF THE AED

$$r_{c,s} = \begin{cases} k \sum_{c' \in \hat{C}} sim(c,c') \times r_{c',s}, & \hat{C}_0 = \varnothing \\ k_0 \sum_{c' \in \hat{C}_0} m_{c'} \times r_{c',s}, & \hat{C}_0 \neq \varnothing \end{cases}, \quad (11)$$

Euclidean distance between arbitrary member c' and user c equal zero), $k = 1/\sum_{c' \in \hat{C}} |sim(c,c')|$, $sim(c,c')$

is computed by Eq. (10), and $m_{c'}$ is the number of co-rated items for user c and c', $k_0 = 1/\sum_{c' \in \hat{C}_0} m_{c'}$.

Where,
rcs = predicting unknown rating value
C_cap = the set of N users that are most similar to c
C_cap_0 = subset of C_cap whose arbitrary member c_cap satisfies dist(c, c_cap) = 0
m_c_cap = the number of co-rated items for user c and c_cap

COMPARISON BETWEEN PEARSON CORRELATION COEFFICIENT AND ADJUSTED EUCLIDEAN DISTANCE

EVALUATION OF PCC

The evaluation is based on two metrics (MAE and RMSE). Given a data set, the values of MAE and RMSE vary depending on the value of K. In this task, I set the value of k to 10. Then I changed the value with the step increment of 5. I performed this increment up to 101.

There was no case wherein the MAE and RMSE computed for predictions through PPC method overtook the AED method.

```
np_predictions = np.zeros((n_users, n_items))

K = 95
EPSILON = 1e-9

for (i, j), rating in np.ndenumerate(test_ds):
```

```
MAE_PCC = evaluation_score_pcc[0] # 0 is an intial value, you need to up
RMSE_PCC = evaluation_score_pcc[1] # 0 is an intial value, you need to u
```

```
# Please don't change this cell

print("==================== The MAE and RMSE of Your Implementation ===
print("MAE: {}, RMSE: {}" .format(MAE_PCC, RMSE_PCC))
```

```
==================== The MAE and RMSE of Your Implementation =========
============
MAE: 0.8495926778484268, RMSE: 1.088806103271781
```

EVALUATION OF AED

The AED showed a similar nature in terms of MAE and RMSE metrics. The difference between the centered cosine similarity and the adjusted Euclidean distance methods is in terms of the comparative values of MAE and RMSE.

For the data set, irrespective of the value of K, the values of MAE and RMSE for the predictions made through the AED method are always less than that of the PCC.

Significant weighting is not counted.

```
np_predictions_ead = np.zeros((n_users, n_items))

K = 95
EPSILON = 1e-9

# Precompute sums and counts for all users to avoid repeated computation
train_ds_sum = np.sum(train_ds, axis=1)
```

```
MAE = evaluation_score[0] # 0 is an intial value, you need to update thi
RMSE = evaluation_score[1] # 0 is an intial value, you need to update th
```

```
# Please don't change this cell

print("==================== The MAE and RMSE of Your Implementation ===
print("MAE: {}, RMSE: {}" .format(MAE, RMSE))
```

```
==================== The MAE and RMSE of Your Implementation =========
============
MAE: 0.1713535700910984, RMSE: 0.4687313902086791
```

CONCLUSION

- The AED outperforms the Centered Cosine Similarity or Pearson Correlation Coefficient in terms of both MAE and RMSE.
- It takes the vector length of each user into account thereby giving the users’ preferences weightage.
- The AED considers the different number of dimensions for different dimensional vector spaces to emphasize the co-rated items between the two user vectors.
- The AED has better sparsity handling capacity that usually previals in the data set.

REFERENCES

REFERENCES 1

Canva - Practical Data Science with Python course materials

REFERENCES 2

Sun, H., Peng, Y., Chen, J., Liu, C., & Sun, Y. (2011). A New Similarity Measure Based on Adjusted Euclidean Distance for Memory-based Collaborative Filtering. Journal of Software, 6(6). https://doi.org/10.4304/jsw.6.6.993-1000.

THANK YOU

Presentation by Dhan Raj Rai

My presentation is based on my own understanding and comprehension despite taking screenshots from the report paper