Download the Movielens dataset 1 (the 100K dataset) from http://www.grouplens.org/node/73

In this assignment you will solve the cold start problem. The 5 folds defined in the dataset is not meant for this. So you have to create your own five folds by hiding users.

Say there are 100 users, you create one fold by hiding first 20 users and use the rest 80 for training. During testing, you introduce the hidden 20 users as ‘new’ users.

You will solve this as a clustering problem.

You have to use k-means clustering and fuzzy c-means clustering. In each case, see how results vary as you change the number of clusters. Take 3 cases: 5 clusters, 10 clusters and 20 clusters.

When the new user is assigned to a cluster, how you predict the rating, is also upon you. You need to explain in a brief-write-up what logic you are following.

As in the previous assignment, you will use MAE as the measure for rating prediction.

Table 1. MAE values User based

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fold # | K-means clustering | | | Fuzzy C-means clusterin g | | |
| 1 | K=5 | K=10 | K=20 | C=5 | C=10 | C=20 |
| 2 |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |
| Average |  |  |  |  |  |  |