● Exercise 4: Application of k-means and FCM to collaborative filtering and comparison with GroupLens results

・Recommendation based on k-Means and FCM clustering results

The two phases of collaborative filtering have close relation to k-Means and FCM clustering, where mutually familiar objects (users) are merged into a cluster and their average scores are summarized into the cluster centers (mean vectors).

Then, once we extract user clusters with their cluster centers (mean score vectors), we can perform collaborative filtering based on the clustering results such that:

1. Neighborhood user search ---> Find the cluster of maximum membership
2. Averaging of scores by neighborhood users ---> cluster center = average score

・Handling missing values in k-Means and FCM clustering results

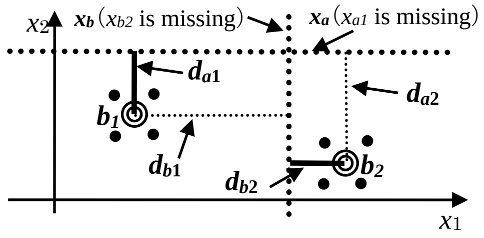
Assume that an $n \times m$ data matrix $X = \{x_{ij}\}$ includes some missing values. We introduce a bivariate variable $h_{ij} \in \{0, 1\}$ such that $h_{ij} = 1$ for observed x_{ij} while $h_{ij} = 0$ for missing x_{ij}. Then, the partial distance d_{ci}^2 among $\boldsymbol{x}_i$ and $\boldsymbol{b}_c$ is given by

$$d_{ci}^2 = \sum_{j=1}^{m} h_{ij} \cdot \left(x_{ij} - b_{cj}\right)^2 \qquad (☆)$$

and the objective function is as follows:

$$L_{fcm} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} \left\{ \sum_{j=1}^{m} h_{ij} \cdot \left(x_{ij} - b_{cj}\right)^2 \right\}$$

In the following figure, partial distances for incomplete objects $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are calculated by considering only their observed variables.



Then, u_{ci} is updated by

$$u_{ci} = \left[\sum_{l=1}^{C} \left(\frac{\| \boldsymbol{x}_i - \boldsymbol{b}_c \|^2}{\| \boldsymbol{x}_i - \boldsymbol{b}_l \|^2} \right)^{\frac{1}{\theta-1}} \right]^{-1}$$

by replacing $\|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2$ with d_{ci}^2 of Eq. (☆). The updating rule for $\boldsymbol{b}_c$ is as:

$$b_{cj} = \frac{\sum_{i=1}^{n} u_{ci}^{\theta} h_{ij} x_{ij}}{\sum_{i=1}^{n} u_{ci}^{\theta} h_{ij}}$$

・Application of k-means and FCM to MovieLens 100K Dataset

Apply k-means and FCM to MovieLens 100K dataset and perform collaborative filtering by using the clustering results with nearest cluster assignment and cluster center-based prediction. Because we have no a priori knowledge on the number of user clusters, try several cluster numbers and check their recommendation ability.
Then, compare their recommendation ability with that of GroupLens.

Be noted that the performance of the clustering-based collaborative filtering can be inferior to that of GroupLens because the clustering process often causes information loss in extracting cluster centers. That is, in the clustering-based collaborative filtering, the information of whole rating data are summarized into some cluster centers and the original rating data are ignored in the prediction process.
This data summarization can contribute to reduction of information overload and the prediction process can be achieved with a smaller memory requirement, i.e., we must store cluster centers only!

In this comparative study, you should consider the influence of clustering-based data summarization through comparison with the result of GroupLens.

“Deadline”
You should finish this exercise in 4 weeks. (by 31, August 2022)