No.: Date:

James Gan

Week 11 - Recommendation System.

11.1 - Clustering

Goal: Partition observations into distinct groups, so that the observations within each group are similar to each other, while observations in different groups are quite different from each other

Common Clustering methods

- Hierarchical Clustering:
  We don't know in advance how many clusters we want.
  (Bottom-up approach / agglomerative)

- K-means clustering.
  Partition the observations into a pre-specified no of clusters.
  (top-down).

Hierarchical Clustering

Algo:

1. Begin w/ n observations & a measure of all $\frac{n(n-1)}{2}$ pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, \ldots, 2$
   - 2a. Examine all the pairwise inter-cluster dissimilarities among the i clusters & identify the pair of clusters that are least dissimilar.
     (Fuse these 2 clusters)
   - 2b. Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

How to determine dissimilarity?

Common types of linkages:

- Complete
- Average
- Single
- Centroid

No.: Date:

K-means Clustering

Idea:

A good clustering is where the within-cluster variation is as small as possible.

$$\min_{C_1,\dots,C_K} \frac{1}{|C_k|} \sum_{k=1}^{K} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Algo:

1. Randomly assign a number, from 1 to K, to each observations.
2. Iterate until the cluster assignment does not change.
   2a. Assignment: For each of the K clusters, compute the cluster centroid
   2b. Update: Assign each observation to the cluster whose centroid is closest.

11.2 – Recommender Systems

The 2 major paradigms of recommender systems:

① Collaborative methods
   - 1a) user-user
   - 1b) item-item
   - 1c) Matrix Factorization

② Content-based methods

① Collaborative Filtering Methods

Based solely on the past interaction recorded between users & items in order to produce new recommendations.

(The interactions are stored as "user-item interaction matrix")

1a) Memory based (user-user & item-item)

Directly works in values of recorded interactions, assuming no model, & are essentially based on nearest neighbours search. (Low bias, high variance)

1b) Model based (Matrix factorization)

Assume an underlying "generative" model that explains the user-item interactions, and try to discover it in order to make new predictions (higher bias, lower variance)

Advantages:
- Require no info about users/items
- More user interest in items, the more new recommendation become accurate.

Disadvantage:
- "Cold start" problem: impossible to recommend anything to new users.

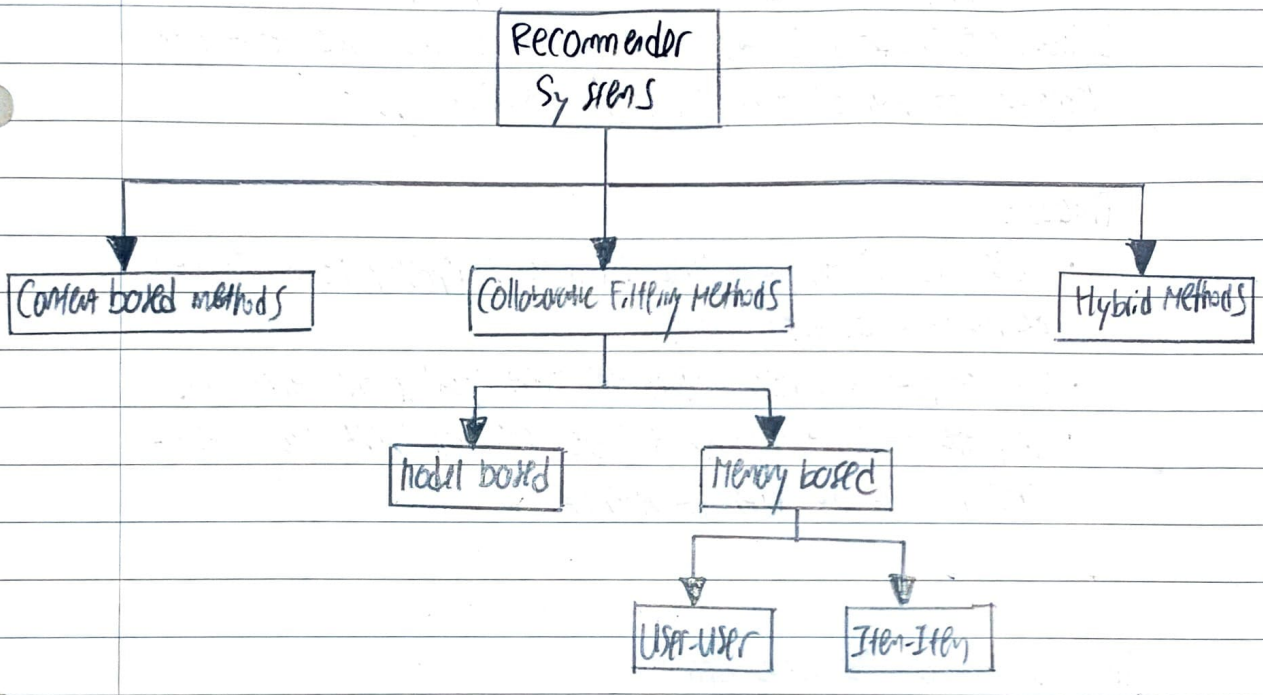
No.: Date:

② Content based methods.

Approach uses additional information about users. Build a model based on available features that explain the observed user-item observations. (Highest bias but lowest variance)

Advantages: Suffer for less from cold start problem than collaborative approaches.

Disadvantage: Unseen new features will result in drawback.



1a) Memory based collaborative approach

The main characteristics of user-user & item-item approach is that they only use information from the user-item interaction matrix & they assume no model to produce new recommendations.

1a1) User-user approach

1a2) item-item approach.

1a1) User-user approach

To make a new recommendation, the user-user approach roughly tries to identify users w the most similar "interaction profile" (nearest neighbor), in order to suggest items that are popular among these neighbors.

No.: Date:

Process:

- Calculate "Similarity" between our user of interest & every other user.
- Once Similarities to every user is computed, we can keep the k-nearest neighbour to our user.
- Suggest the most popular items among them.

1(a 2) item-item approach

Find items similar to the ones the user already "positively" interacted w.
2 items are considered to be similar if most of the users that have interacted w both of them, did it in a similar way.

Process:

- Consider the item this user liked the most & represent it by its vector of interaction w every user.
- We compute similarities between the "best item" & all other items.
- Keep the k-nearest neighbour to the selected "best item" that are new to our user of interest.

1b) Model based Collaborative Approach

Rely only on user-item interactions information & assume a latent model supposed to explain these interactions.