

## Quiz 7 Thursday - Rubrics

1. [1 point] What are the characteristics of points belonging to a cluster?  
**Members of a cluster are close/similar to each other.**  
**Members of different clusters are far/dissimilar from each.**
2. [3 points] Hierarchical Clustering: Imagine we are clustering the number of items bought on Amazon for a given user each day. We wish to perform a hierarchical clustering of the number of items bought among all users: 1, 3, 8, 16, and 30. Show what happens at each step **until there are two clusters**, and give these two clusters. Assume clusters are represented by their centroid (average), and at each step choose to merge two clusters whose resulting cluster has the **smallest diameter**.

Your answer should be in steps where each step shows the members of the new cluster formed, and its centroid. More specifically, if you are merging a cluster  $C1 = \{x, y, z\}$  of centroid  $c1$  with a cluster  $C2 = \{p, q\}$  of centroid  $c2$ , you should report  $\{x, y, z, p, q\}$  in the table, as well as the centroid obtained with these 5 points).

Initial clusters :  $\{1\}, \{3\}, \{8\}, \{16\}, \{30\}$

Centroids: 1, 3, 8, 16, 30

First step :  $\{1, 3\}, \{8\}, \{16\}, \{30\}$  [0.5 points]

Centroids: 2, 8, 16, 30 [0.5 points]

Second step :  $\{1, 3, 8\}, \{16\}, \{30\}$  [0.5 points]

Centroids: 4, 16, 30 [0.5 points]

Third step :  $\{1, 3, 8\}, \{16, 30\}$  [0.5 points]

Centroids: 4, 23 [0.5 points]

3. [1 point] What are the advantages of Item-based Collaborative filtering to User-based methods? Give with at least 2 advantages.

-Match user's rated items to similar items, Rather than matching similar users

-In practice, often leads to faster online systems and better recommendations

-Similarities between pairs of items  $i$  and  $j$  are computed off-line

-Predict rating of user a on item i with a simple weighted average.

- Item based doesn't suffer from cold start problems as much as user based does.

- User based is more prone to shilling attacks than Item based Collaborative filtering.

Rubrics: 0.5 points for each correct advantage.

4. [1 point] Select ALL of the statements that are TRUE about Inverse User Frequency.
  - A. Universally liked items are as useful in capturing similarity as less common items.
  - B. Inverse frequency  $f_j = \log(n_j/n)$ ,  $n_j$  is number of users who have rated item  $j$ ,  $n$  is total number of users.
  - C. If everyone has rated item  $j$ , then  $f_j$  is zero.

D. When transform ratings by multiplying the original rating by  $f_j$ , less popular items will have greater effect on prediction

Ans.) C, D

5. [1 point] For 2 points  $A(x_1, y_1)$  and  $B(x_2, y_2)$  and the distance between them being  $\text{dist} = ((x_1 - x_2)^2 + (y_1 - y_2)^2)^{0.5}$ . **dist** could be which of the following distance measures?
- A. Manhattan Distance
  - B. Euclidean Distance
  - C. Minkowski Distance
  - D. Edit Distance

Answer. B,C

6. [2 points] What are the Characteristics and Challenges of Collaborative Filtering with Cold start problems? Name two approaches on how to deal with this problem?

•Cold start problem

- When a new user or item has just entered the system

- Hard to find similarities: not enough information to make good recommendations

- New item problem: can't be recommended until some users rate it

- Also applies to obscure items

- Also called "first-rater problem"

- New users:

- not given good recommendations because of lack of rating or purchase history

•Approaches:

- Content-based systems do not rely on ratings from other users

- Hybrid CF (content-boosted CF): external content information can be used to produce predictions for new users or new items

- Research on effectively selecting items to be rated by a user to rapidly improve recommendation performance

Rubrics: 1 point for challenges and 0.5 points for each approaches

7. [1 point] Which of the following Hybrid Recommendation Systems involves the concept of breaking ties using one of its components?
- A. Feature Augmentation
  - B. Cascade Hybrid
  - C. Meta-Level Hybrid
  - D. Switching Hybrid

Answer. B