Lecturer: Prof. John Hopcroft
Scribe: Jang Ho Kim
May 1, 2006

Lecture 39

How do you evaluate an algorithm for collaborative filter?

- Utility – sum over all users of the probability that user will buy recommended item
- Optimum algorithm, OPT, would recommend the highest probability item for each user.
- Utility of OPT, $\prod(\mathrm{OPT}) = \sum_{user} \max_{item} P_{item}$
  (look at each row, take the highest probability and take the sum)
- Simple case:
  Consider two categories, c_1 and c_2, each user buys two items. All items in a category are equally likely. Take the sum over all users. Category c_1 is preferred.
  Graph: items are vertices and users are edges
  User is c_1 edge → recommend any item in c_1
  (since all items in c_1 are equally likely)
  User is cross edge → if $c_1 > c_2$ then recommend c_1
  Otherwise, either
  User is c_2 edge → depends on structure of matrix
- To consider which algorithm is better consider the following in worst case.
$$\min_{data} \frac{\prod(\mathrm{ALG})}{\prod(\mathrm{OPT})}$$
- Sample size: s
  Number of users: m
  Number of items: n
  Number of categories: k
- No algorithm can do better than $\frac{2}{\sqrt{k}+1}$
- Theorem: For sample size s=2, for any ALG,
  i) $\min_{data} \frac{\prod(\mathrm{ALG})}{\prod(\mathrm{OPT})} \le \frac{2}{\sqrt{k}+1}$
  ii) $\min_{data} \frac{\prod(\mathrm{VRC})}{\prod(\mathrm{OPT})} \le \frac{2}{\sqrt{k}+1}$ (VRC = vote randomly)

  Proof: $\prod(\mathrm{OPT}) = \sum_{u} \max_{i} P_i(u)$

  Utility for VRC:
  - c_i edge occurs with probability $P_i^2(u)$: utility $P_i(u)$

- cross edge occurs with probability $P_i(u)P_j(u)$: utility $\frac{P_i(u)+P_j(u)}{2}$

$$\sum_u \left[\sum_i P_i^{\,3}(u) + \sum_{i \neq j} P_i(u)P_j(u)\frac{P_i(u)+P_j(u)}{2} \right]$$

$$= \sum_u \left[\sum_i P_i^{\,3}(u) + \sum_{i \neq j} \frac{P_i^{\,2}(u)P_j(u) + P_i(u)P_j^{\,2}(u)}{2} \right]$$

$$= \sum_u \left[\sum_i P_i^{\,3}(u) + \sum_{i \neq j} \frac{P_i^{\,2}(u)P_j(u)}{2} + \sum_{i \neq j} \frac{P_i(u)P_j^{\,2}(u)}{2} \right]$$

$$= \sum_u \left[\sum_i P_i^{\,3}(u) + \sum_{\substack{i,j \\ i \neq j}} P_i^{\,2}(u)P_j(u) \right]$$

$$= \sum_u \left[\sum_i P_i^{\,3}(u) + \sum_i P_i^{\,2}(u) \sum_{\substack{j \\ i \neq j}} P_j(u) \right]$$

$$= \sum_u \left[\sum_i P_i^{\,3}(u) + \sum_i P_i^{\,2}(u)\left(1 - P_i(u)\right) \right]$$

$$= \sum_u \sum_i P_i^{\,2}(u)$$

$$\frac{\prod(\text{VRC})}{\prod(\text{OPT})} = \frac{\sum_u \sum_i P_i^{\,2}(u)}{\sum_u \max_i P_i(u)}$$ Want to minimize

Let $P_m(u)$ be maximum for user u.

Maximum occurs for vector, $\left(P_m, \frac{1-P_m}{k-1}, \frac{1-P_m}{k-1}, \cdots, \frac{1-P_m}{k-1} \right)$

For each user u, let m be the value of i for which $P_i(u)$ is maximized.

Then $\frac{\prod(\text{VRC})}{\prod(\text{OPT})}$ is minimized for each u if $\sum_i P_i^{\,2}(u)$ is minimized subject to keep $P_m(u)$ fixed. This occurs if $P_i(u) = P_j(u)$ for all i and j not equal to m.

Now, $\frac{\prod(\text{VRC})}{\prod(\text{OPT})} = \frac{\sum_u \left[P_m^{\,2}(u) + \left(\frac{1-P_m(u)}{k-1} \right)^2 (k-1) \right]}{\sum_u P_m(u)}$