Preface for Friday

Say we were working for the Census Borough and had some data in the form of age-salary pairs (data points are (age, salary)). We want to be able to answer questions about the database that cut out rectangular sections of the data (when graphed). For example, "How many people have a salary between 10 and 20 that have ages between 20 and 30?" We want to be able to obtain a sample of the full dataset that is representative enough for us to accurately answer such questions.

This topic will be discussed in more detail on Friday, 5/5/06.

Review of Collaborative Filtering

Recall our topic from last class.

How do we measure the goodness of a collaborative filtering algorithm?

One possibility is to measure the Utility of the algorithm. The Utility of an algorithm is the sum over all users of the probability that a user would purchase the item recommended. Note: the probability of a user buying an item is independent of our recommendation.

How do we compare algorithms?

We can examine the worst case over all settings (probability matrices), which is the ratio:

$\min \frac{\pi(ALG)}{\pi(OPT)}$ $\pi = Utility$ $ALG =$ any algorithm $OPT =$ the optimal algorithm

Theorem: Assume sample size s, number of items each user buys in a transaction, is equal to 2, all items in a category are equally likely, and ALG is any algorithm.

1) $\min \frac{\pi(ALG)}{\pi(OPT)} \leq \frac{2}{\sqrt{k}+1}$ $k =$ number of categories

2) $\min \frac{\pi(VRC)}{\pi(OPT)} = \frac{2}{\sqrt{k}+1}$

Sketch of the proof for 2):

In the worst case, VRC does as well as any algorithm (but on other data, some other algorithm could do better).

At some point in our proof (recall lecture on 5/1/06), we had a ratio:

$$\min \frac{a_1 + a_2 + \ldots + a_n}{b_1 + b_2 + \ldots + b_n} \quad \text{occurs if we minimize } \frac{a_i}{b_i}$$

However, examine the following:

$$\frac{1+1}{2+3} = \frac{2}{5} \qquad \frac{2+1}{4+3} = \frac{3}{7}$$

The ratio of one element changing does change the ratio of the entire sum. Instead, we fix every ratio to be the minimum ratio of elements in the sum, which is what we claimed in the proof. For the rest of the proof, pick a certain "bad" distribution to show equality.

Another Algorithm for Collaborative Filtering

Suppose we have matrices $A = PW$
Suppose we also know W, the probabilities of an item given a category.
Let u be a row of A (the probabilities of buying each item for a specific user)
Let $\tilde{u}$ be an estimate for u based on s purchases by that user.
Clearly the row of A is in a space spanned by W, but it is not likely that the estimate is in the space spanned by W.

How do we project $\tilde{u}$ into the space spanned by W?

Orthogonally (Spectral method)? This method would minimize the squared error between u and $\tilde{u}$.

However, we're mainly interested in the entries in u with maximum probability. We only care about $\tilde{u}$'s error among the top elements (most probable elements) of u, since these are the ones we would recommend. We want each component of $\tilde{u}$ to be close in value to the corresponding component of u (which does not necessarily minimize total error).

<u>Pseudo Inverse</u>

For the inverse of W, it is true that $WW^{-1}x = x$ for any x.
For the pseudo inverse, we only require $WW^{-1}x = x$ for x in the space of W.

What is the value of $WW^{-1}(u - \tilde{u}) = u - WW^{-1}\tilde{u}$? How close are u and $WW^{-1}\tilde{u}$?

We would like each component of $WW^{-1}\tilde{u}$ to be within ε of u with high probability.
What is the bound for $v^T(u - \tilde{u})$ if the maximum component of v is bounded by some constant B? ($B =$ max element of WW^{-1})

$\tilde{u} = \frac{1}{s}\sum_{i=1}^{s} \bar{u}_i$ $\bar{u}_i$ is an indicator variable for the i th sample

$v^T\tilde{u} = \frac{1}{s}\sum_{i=1}^{s} v^T\bar{u}_i$ sum of independent random variables that are bounded by the elements of v^T

Using Chebyshev's Inequality:

$$P\left(\left|v^T\tilde{u} - v^T u\right| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2} = \frac{B^2}{s\varepsilon^2}$$

We want to pick W^{-1} such that the largest element in W^{-1} is small ($\rightarrow B$ is small) to keep this bound tight. We can use linear programming to optimize W^{-1}.

If we had used the Spectral method from earlier: $\frac{1}{\lambda_{\min}}$ is the largest element in W^{-1}.

Instead we find that the maximum element in W^{-1} can be bounded by the following:

$R\frac{W^+}{\min_{\|x\|_2=1}\|Wx\|_2}$ W^+ is the largest element in W

This algorithm differs much from our previous approach (where we simply picked some rules to follow based on the maximum probable category for a user). Instead, we are trying to generate an estimate of the user's probability of buying each item based upon the existing purchase data we have for all users, which is represented in W.