Recommender Systems

- Which items should a system (Amazon, Netflix, etc.) recommend to you?
  - Based on other items you've bought
  - Based on other items you've looked at
  - etc.
- Two kinds
  - Collaborative filtering - based on ratings (e.g., how you've rated earlier items you've bought)
  - Content-based recommender systems - based on content (i.e., based on what you've bought)
- Many variants and intermediate approaches

Collaborative Filtering

- Originated at Xerox PARC
  - PARC = Palo Alto Research Center
  - David Goldberg, 1992

- Two-step algorithm
  - Find users like yourself
  - See what they liked
- “Like yourself” = many similarity metrics

Finding People “Like Yourself”

- What does “like yourself” mean?
  - People whose ratings were similar to yours on other items
  - Many ways to decide exactly what is “similar”
  - Also many ways to combine people’s ratings
- One way to use other people’s ratings
  - Look at every person in the data base
  - See who has similar taste to you
  - Rank them
  - Take the top 5 (If you just take the top 1, that person might have idiosyncratic taste)

Create Recommendations

- For each item (e.g., movie)
  - Calculate total ratings: For every person who saw the movie, multiply their similarity to you (as a percent) by their rating. Add them up.
  - Calculate total similarities: For every person who saw the movie, add up their similarity percents.
  - Divide total ratings by total similarities.
- Why divide by total similarities? (i.e., why not just use total ratings?)
  - If a lot of people like you saw a movie but they hated it, that's not what we want to recommend.

Characteristics of Similarity Metrics

- $0 <= sim(A, B) <= 1$
- $sim(A, A) = 1$
- $sim(A, B) = sim(B, A)$

Three Similarity Metrics

- Euclidean
- Pearson
- Jaccard
- Many others and variants

Euclidean Similiarity Metric

- See handout for formula
- Why invert distance?
  - So greater similarity will have higher value
- Why add 1 to distance?
  - So denominator is never 0

Pearson Similiarity Metric

- See handout for formula
- Normalizes ratings = converts to units from the mean instead of absolute ratings
  - Unit = standard deviation
- I.e., uses “good” or “bad” with respect to that reviewer’s usage of the term instead of absolute meaning

Jaccard Similiarity Metric

- Jaccard = sum for all reviewers
  - # of times both people gave the movie the same rating
    /
    # of times both people saw the movie

Finding People “Like Yourself”

- Calculate similarity of every other person to yourself
- If you want to know whom to follow
  - Rank by similarity
  - Take top 1 or top 5
- If you want to use this data to find items (e.g., movies) you might like, see next slide

Create Recommendations

- For each item (e.g., movie)
  - SUM (for each reviewer who saw the movie, their similarity to you * their rating)
    /
    SUM (for each reviewer that saw the movie, their similarity to you)
  - Purpose of denominator = we do not want movies that more people saw to automatically get better ratings

Create Recommendations

- This is the same calculation as GPA
  - Calculate total quality points: For each course you've taken, multiply grade by credit hours. Add them up.
  - Calculate total credit hours.
  - Divide total quality points by total credit hours.
- If we just looked at total quality points, seniors could rank higher just because they've taken more courses, even if they got bad grades in them.