5-8: Unary Data Evaluation

Introduction

- We've talked a lot about ratings data
- Some metrics are applicable for unary
  - P/R and friends
- This time: unary data
  - Often studied under ‘implicit feedback’
  - Unary data is positive-only (purchase, like)

Implicit Feedback Data

- Many recommender contexts have no ratings or other explicit data
- Often data owners have non-unary data
  - Like vs. saw but didn't like
  - Clicked vs. saw but skipped
- W/ negative examples, can just do standard eval
- Use more data if you have it

Unary Data

- We often don't have negatives
- Anyone using a data dump
  - song plays (don't know didn't play)
  - click logs
- Intrinsic to certain domains/tasks
  - research papers
  - physical store purchases

Problems

- No negative examples
- How do we know if the recommender is wrong?
  - Or if the user just didn't know about the item?
- Put differently: how do we avoid punishing the recommender for doing its job?

Metrics

- Precision/Recall/MAP
  - but is ‘bad’ really bad?
- MRR (Mean Reciprocal Rank)
  - still gets pushed down
- % At or Before Rank
  - histogram of raw data for MRR
- nDCG
  - first item may still be misjudged

Mitigation strategies

- Synthesize unary data to get negatives
  - e.g. ratings, >= 3.5 stars is ‘like’
  - only recommend from rated data
- Limit domain of recommendation
  - recommend from good + N unknown
  - limits likelihood of good-but-unknown
    - these items are probably excluded

Best we can do

- These evaluations are the best we have
- So use them
  - But be aware of limitations when reporting
- Look for alternatives
  - User testing
  - Try to get negative data
- Corroborate with additional evidence

Promising Directions

- One-sided classification
- New metrics and protocols (e.g. clarity)

Conclusion

- Evaluating recommenders is hard
- Offline evaluation doubly so
- We don't have great methods right now
- Be aware of problems when making claims
  - Both in research and industry

5-8: Unary Data Evaluation