

Top-N Protocols and Unary Data

Introduction

- We've seen
 - Prediction metrics
 - Top-N metrics
 - Cross-validation protocols
- Now: some particular protocol considerations for top-N evaluation,
 - Particularly for unary data
- Unfortunately, some deep unsolved problems

Missing Data

- Most of our data is missing
 - Unrated items
 - Unpurchased items
- What do we know about items with no record?
 - User *probably* doesn't like item...
 - ... but if they do, it might be a *really good* recommendation



Irrelevant item – user does not like apples

Relevant item – user would like apples
And may have never heard of them!

We can't tell the difference.

Popularity Bias

- Popular items have the most ratings/purchases
- So 'Popular' recommender can automatically do very well
 - Popular is a good baseline recommender anyway
 - But if we want the recommender to find less-known items the user might like...
- Recommenders that effectively find relevant but less-popular items perform poorly on metrics

Studied by Alejandro Bellogin

Problem Summary

- Penalizing good recommendations due to missing data
- Data favors ‘most popular’
 - Correlation between bias and performance unknown

Solution 1: Rank Effectiveness

If we have ratings, or other negative feedback:

1. Ask recommender to rank test items, rather than recommend from entire universe.
2. Measure if rank is consistent with user ratings
 - MAP
 - nDCG
 - NDPM

Requires ratings, so no good on unary data.

Proposed Solution 2: Limit Domain



test
items

- Limit *candidate set* for recommendation
 - Recommend from test + N randomly-selected unknown items
 - Introduced by Koren (2008)
- Good-but-unknown items are *probably not* in candidate set
 - Assuming sufficiently low prevalence of good items
- However: seems to exacerbate popularity bias
 - Randomly-selected items probably aren't popular
 - Found by Mahant (2016)

Best Known Practices

- These are the best we have
- 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

Top-N Protocols and Unary Data