

5-1: Introduction to Evaluation of Recommender Systems

Goals for Today

- To understand ways of evaluating the “goodness” of a recommendation, and of a recommender algorithm or system
 - Accuracy metrics
 - Error metrics
 - Decision-support metrics
 - User and Usage-centered metrics
- To understand how predictions and recommendations (including top-n) are evaluated
- To understand retrospective and live approaches to evaluation

Why a whole module on evaluation?

- Zillions of algorithms, but which to pick?
- Lessons from commercial experience
- Lessons from the Netflix Challenge
- Lessons from (and for) the research community

A Historical Look

- The early days
 - Accuracy and error measures:
 - MAE, RMSE, MSE
 - Decision-support metrics:
 - ROC AUC, Breese score, later precision/recall
 - Error meets decision-support/user experience:
 - “Reversals”
 - User-centered metrics:
 - Coverage, user retention, recommendation uptake, satisfaction

A Commercial Look

- Nobody cared about accuracy ...
 - The supermarket recommender
- Lift, cross-sales, up-sales, conversions
- Led to thinking about different measures anchored not only to user experience, but recommender goals

Moving Forward ...

- Lots of new metrics developed as researchers looked at tuning for specific purposes:
 - More sophisticated top-n / rank metrics
 - Serendipity
 - Diversity
- More systematic evaluation of the recommender as a whole (not just the recommendations)

Theme 1: Prediction vs. Top-N

- Key distinction:
 - Prediction is mostly about accuracy, possibly decision support; focused locally
 - Top-N is mostly about ranking, decision support; focused comparatively

Theme 2: More than Just Metrics

- Even simple evaluations are hard ...
 - How to calculate Mean Absolute Error
 - Easy to compute error of a single prediction
 - Average across predictions or across users?
 - How to handle lack of coverage?
- Comparative evaluation is even harder ...
 - Proper baseline
 - Different coverage, etc.

Theme 3: Unary Data

- Many of the metrics we present are designed specifically for evaluating data with a multi-point rating scale (e.g., 1-5).
- Some measures don't work well at all for unary data (e.g., purchase data)
- Special coverage of unary evaluation ...

Theme 4: Dead vs. Live Recs?

- Retrospective (dead data) evaluation looks at how recommender would have predicted or recommended for items already consumed/rated.
- Prospective (live experiment) evaluation looks at how recommendations are actually received.
- Fundamental differences ...

Looking forward ...

- This module includes:
 - Lectures on the major types of evaluation, and on how to conduct a rigorous evaluation
 - A “rant” on when evaluations may be meaningless
 - Assignments focused on conducting evaluation, both by hand and on a large scale with LensKit
- Going forward, evaluation should be part of your toolkit ...

5-1: Introduction to Evaluation of Recommender Systems