UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Recommender Systems

Evaluation Metrics

Rodrygo Santos
rodrygo@dcc.ufmg.br

Why evaluate?

UFMG COMPUTER SCIENCE

- Gazillions of algorithms
  - Collaborative, content-based, hybrid…
  - ***Which one to choose?***
- Evaluation enables an informed choice
  - Rigor of science
  - Efficiency of practice

Recommender evaluation

UFMG COMPUTER SCIENCE

- Lessons from academia
  - Evaluation methodologies
  - User behavioral models
  - Evaluation metrics
- Lessons from industry
  - ***What works in practice?***

A historical look

UFMG COMPUTER SCIENCE

- Accuracy and error metrics
  - MAE, MSE, RMSE
- Decision-support metrics
  - ROC, AUC, precision, recall
- Ranking accuracy
  - Reversals, early performance
- User-centered metrics
  - Coverage, user retention, satisfaction

A commercial look

UFMG COMPUTER SCIENCE

- Nobody cared about accuracy…
  - The supermarket recommender
  - Lift, cross-sales, up-sales, conversions
- Not only user experience
  - Recommender goals also matter

Moving forward

UFMG COMPUTER SCIENCE

- Metrics tuned for specific purposes
  - Sophisticated rank-based metrics
  - Diversity and novelty
  - Serendipity
- Holistic evaluations
  - Beyond just the recommendations
  - Whole-page relevance

Which method?

UFMG COMPUTER SCIENCE

Retrospective evaluation

- Offline experiments
  - How well can we predict (hidden) ***past preferences***?
- Highly reproducible
  - Multiple evaluations share the ***same data***
- Cheap, but incomplete
  - How to handle ***missing*** user preferences?

Prospective evaluation

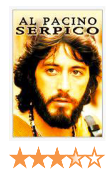
- Online experiments
  - How well can we predict ***future preferences***?
- Poorly reproducible
  - Multiple evaluations use ***different data***
- Costly, but realistic
  - Users are actually ***exposed*** to the recommendations

Which output?

UFMG COMPUTER SCIENCE

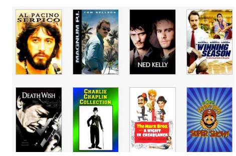
Prediction

- Mostly about accuracy
  - Possibly decision support
- Focused locally



Recommendation

- Mostly about ranking
  - Definitely decision support
- Focused comparatively



Which ground-truth?

UFMG COMPUTER SCIENCE

Explicit feedback

- Traditional (e.g., 1-5 stars) but ***expensive***
- Potentially noisy
  - What users ***like*** vs. what they ***say*** they like

Implicit feedback

- ***Abundant:*** views, clicks, dwells, purchases, etc.
- Troublesome
  - How to factor in ***negative feedback***?

How to quantify recommendation effectiveness?

A note on terminology

UFMG COMPUTER SCIENCE

Effectiveness

Effectiveness is about doing the right thing. In recommendation, it's about recommending items that the user will find interesting.

Efficiency

Efficiency is about doing something (good or bad) in an optimal way. In recommendation, it's about doing things faster or with fewer resources.

Evaluation metrics

UFMG COMPUTER SCIENCE

- Prediction accuracy
  - *How well does it estimate absolute preferences?*
- Decision support
  - *How well does it return "good" things?*
- Ranking accuracy
  - *How well does it estimate relative preferences?*

Accuracy metrics

UFMG COMPUTER SCIENCE

- Accuracy of a prediction
  - Closeness to the actual preference
- Actual preference unknown from system
  - ***Hidden*** in an offline evaluation
  - ***Truly unknown*** in an online evaluation
- Typically measured by ***error metrics***
  - Distance to the actual preference
    - e.g., predicted = 3, actual = 2.5, error = 0.5

MAE: mean absolute error

UFMG COMPUTER SCIENCE

- What is error?
  - Difference from the actual preference

    $\hat{r}_i - r_i$
- Absolute error removes direction
  - Two wrongs don't make a right!

    $\left|\hat{r}_i - r_i\right|$
- MAE

  $\frac{1}{n}\sum_{i=1}^{n}\left|\hat{r}_i - r_i\right|$ for n ratings considered

MSE: mean squared error

UFMG COMPUTER SCIENCE

- Why ***squared*** error?
  - Removes sign (avoids need for absolute value)
  - Penalizes large errors more than small
- MSE

  $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_i - r_i\right)^2$ for n ratings considered
- Disadvantage
  - Not an intuitive scale

RMSE: root mean squared error

UFMG COMPUTER SCIENCE

- RMSE

  $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_i - r_i\right)^2}$ for n ratings considered
- Advantage
  - Same scale as the ratings

WAIT A SECOND!

UFMG COMPUTER SCIENCE

- What could go wrong with average errors?
  - We averaged over all ratings
- What if a user has 10k ratings and another 10?
  - The evaluation will be biased!
- Alternative?
  - Average over user averages
  - In practice, look at both

Reflections

UFMG COMPUTER SCIENCE

- In general, all discussed error metrics move together (good replacements for each other)
  - Squared may matter for large scales with algorithms that have occasional huge errors
  - Benefit: lots of published MAE results
- A few drawbacks
  - Different rating scales are not comparable
  - Errors can be dominated by irrelevant parts of the ratings space (popular users or items)

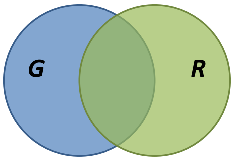
Decision support metrics

UFMG COMPUTER SCIENCE

- Decision support
  - How well a recommender helps the user make "good" decisions and avoid "bad" ones
- What is "good" and "bad"?
  - Depends on the application
- In general
  - Predictions: 4* vs. 2.5* is worse than 2.5* vs. 1*
  - Recommendations: early positions matter most

Precision and recall

UFMG COMPUTER SCIENCE



G: relevant
R: retrieved

Precision

- Percentage of returned items that are relevant

$$Prec = \frac{|G \cap R|}{|R|}$$

Recall

- Percentage of relevant items that are returned

$$Rec = \frac{|G \cap R|}{|G|}$$

Precision and recall

UFMG COMPUTER SCIENCE

Precision

- About having mostly useful stuff in a recommendation
  - Not wasting the user's time
- Key assumption
  - There is more useful stuff than you want to examine

Recall

- About not missing useful stuff in a recommendation
  - Not making a bad oversight
- Key assumption
  - You have time to filter through recommendations

We can also combine both

$$F1 = \frac{2\,Prec\,Rec}{Prec + Rec}$$

Precision and recall

UFMG COMPUTER SCIENCE

Problem #1

- Cover entire dataset
  - Inherently "full query"

Solution

- Ranking cutoffs
  - *Prec@n*
  - *Rec@n*

Problem #2

- Need full ground-truth
  - Only way to exactly compute recall after all
- If we had full ground-truth
  - Wouldn't need a recommender!

Solution

- Limit to rated items
  - Most common approach
  - ***What to do with missing judgements?***

MAP

UFMG COMPUTER SCIENCE

- AP: average precision

$$\frac{\sum_{i=1}^{n} Prec@k \times rel(i)}{|G|}$$

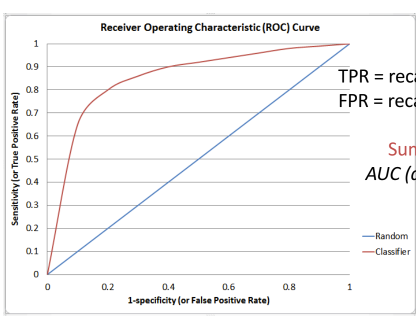
Summary statistic
(estimated area under precision-recall curve)

- MAP (mean AP)

$$\frac{1}{m}\sum_{i=1}^{m} AP(q_i)$$ for *m* "queries" considered

ROC / AUC

UFMG COMPUTER SCIENCE



TPR = recall of relevants
FPR = recall of non-relevants

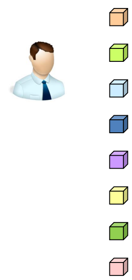
Summary statistic
AUC (area under curve)

Reflections

UFMG COMPUTER SCIENCE

- Once again, all of these metrics tend to correlate highly with each other
  - *Prec@n* and overall precision are perhaps the most widely used (and easily understood)
  - ROC/AUC provides insight if the goal is to tune the recommender's use as a filter
- None of these metrics overcome the problem of being based on rated items only
  - And the inherent noise that this brings

Ranking metrics



- Why ranking?
  - Place items in order of preference
- Key assumption
  - Users will inspect recommended items from top to bottom

MRR: mean reciprocal rank

- RR
  $\frac{1}{i}$ i is the position of the first relevant
- Similar to precision and recall
  - *Prec/Rec* measures goodness at being relevant (precision) and finding things (recall)
  - *RR* measures how deep you need to dig in
- MRR
  $\frac{1}{m}\sum_{i=1}^{m} RR(q_i)$ for m “queries” considered

Correlation coefficients

- Measure how well we got the order right
- Spearman’s ρ
  $$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$
  n: number of items
  d: rank difference
- Kendall’s τ
  $$\tau = 2\frac{n_c - n_d}{n(n-1)}$$
  n: number of pairs
  nc: number of concordant pairs
  nd: number of discordant pairs

Problem: errors at high positions as important as those at low positions

DCG: discounted cumulative gain

- Measure *utility* of item at each position
  - Discount by log of position i
  $$DCG = \sum_{i=1}^{n} \frac{2^{rel(i)} - 1}{log_2(i+1)}$$
- In practice, normalized by ideal DCG and averaged across all “queries”
  $$nDCG = \frac{1}{m}\sum_{i=1}^{m} \frac{DCG(q_i)}{iDCG(q_i)}$$ for m “queries” considered

Reflections

- Several metrics to measure a recommender’s ability to order the recommended items
  - Mostly borrowed from search evaluation
- nDCG increasingly common
  - MRR also used

Business metrics

- We are interested in satisfying the user
  - Accuracy metrics
  - Decision support metrics
  - Rank metrics
- But also the recommendation provider
  - Coverage
  - Diversity
  - Serendipity

Coverage

UFMG COMPUTER SCIENCE

- Measures the percentage of products for which a recommender can make a prediction
  - Or a prediction that's personalized
  - Or a prediction above a confidence threshold
    - e.g., how many 5-stars movies will I be recommended?
- Business interest
  - Reach the entire catalog (aka the long tail)

Diversity

UFMG COMPUTER SCIENCE

- Measures of how different the recommendations are
  - Applied to a top-*n* list
- Examples
  - Intra-list similarity is the average pairwise similarity; lower score means higher diversity
  - Metrics borrowed from search measure diversity with respect to different user aspects
    - Interest for different item features

Serendipity

UFMG COMPUTER SCIENCE

- Measures "the occurrence of events by chance in a happy or beneficial way"
  - In RS: surprising, delightful unexpectedness
- Several ways to operationalize
  - Typically, based on rarity

Summary

UFMG COMPUTER SCIENCE

- Several metrics for different purposes
  - No one-size-fits-all solution
  - Different metrics, different quality estimates
- Metrics may not well correlate with practice
  - Must look outside the box

Summary

UFMG COMPUTER SCIENCE

"In industry, we care about keeping our users and making them happy; not improving accuracy of recommendations by 1%"

Tao Ye, Senior Scientist at Pandora
RecSys 2015, Industry Panel