5-3: Basic Decision Support Metrics

Goals for Today

- To understand the concept of “decision support” metrics
- To learn a set of decision-support metrics, including:
  - Error rate and Reversals
  - Precision, Recall, MAP
  - Receiver operating characteristic
- To understand the usefulness and limitations of these metrics

What is “Decision Support”

- Measure how well a recommender helps users *make good decisions*
  - Good decisions are about choosing “good” items and avoiding “bad” ones
- For predictions: 4* vs. 2.5* worse than 2.5* vs. 1*
- For recommendations, top of list is what matters most.

Errors and Reversals

- What is an “error?”
  - Ad hoc measure of wrong predictions
  - E.g., determine that 3.5-5* = good, 1-2.5* = bad
    - Error is when a good movie (for a user) gets a bad prediction (or vice versa)
  - Can also be used for top-n – every time a bad movie appears in the top-n, it is an error.
  - Usually reported as total number (compared between algorithms), average error rate per user, etc.
  - Not widely used in research

- Reversals are large mistakes – e.g., off by 3 points on a 5-point scale
  - Intuition is that these are likely really bad – lead to loss of confidence
  - Again, reported as total or average rate

Precision and Recall

- Information Retrieval Metrics
  - Precision is the percentage of selected items that are “relevant”
    - $P = \frac{N_{rs}}{N_s}$
  - Recall is the percentage of relevant items that are selected
    - $R = \frac{N_{rs}}{N_r}$

Precision and Recall (2)

- Different Goals
  - Precision is about returning mostly useful stuff
    - Not wasting user time
    - Assumption is that there is more useful stuff than you want
  - Recall is about not missing useful stuff
    - Not making a bad oversight
    - Assumption is that you have time to filter through results to find the key result you need
  - When these two goals are in balance, F-metrics
    - $F_1 = \frac{2PR}{P+R}$

Precision and Recall (3)

- Problem #1 with precision/recall
  - Need ground truth for all items
    - But if we had ground truth, why bother with a recommender
  - Ways this is addressed
    - Fake precision/recall by limiting to rated items
      - Common – results in interesting biases
    - Human-rating experiments that compute precision/recall over some random subset

Precision and Recall (4)

- Problem #2 with precision/recall
  - Covers entire data set – not targeted on top-recommended items
    - precision/recall inherently about "full query"
  - Addressed through P@n, R@n
    - Precision@n is the percentage of the top-n items that are "good": $P@n = \frac{N_{r@n}}{n}$
    - Some have proposed computing this as an average over a set of experiments with 1 "hit" and a large number of presumed misses
    - Recall@n is effectively the same

Mean Average Precision (MAP)

- In IR, MAP averages over both multiple queries and over position in top-n retrieval
  - $MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$ where $AveP = \frac{\sum_{k=1}^{n} P(k) * rel(k)}{\#\ relevant\ docs}$
  - More intuitively, this is computing an estimate of the area under the precision-recall curve, and then averaging across queries
- In recommender systems, this has been adapted in several ways, most commonly across users. Unfortunately, not standardized.
  - Useful comparing algorithm variants; for other comparisons, need to be really careful …

Receiver Operating Characteristic

- The ROC curve is a plot of the performance of a classifier or filter at different thresholds. It plots true-positives against false positives:
  - http://en.wikipedia.org/wiki/Receiver_operating_characteristic
- In recommender systems, the curve reflects trade-offs as you vary the prediction cut-off for recommending (vs. not).
- Area under the curve is often used as a measure of recommender effectiveness

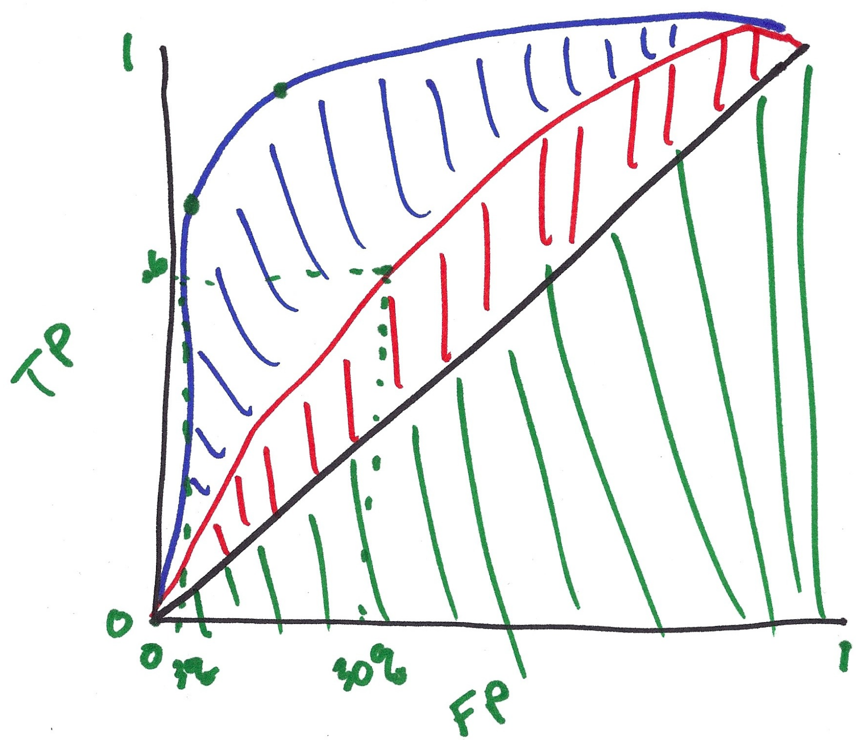
Reflections …

- Once again, all of these metrics tend to correlate highly with each other (good replacements for each other)
- Precision@n and overall precision are perhaps the most widely used (and easily understood)
- ROC provides insight if the goal is to tune the recommender's use as a filter, or identify "sweet spots" in its performance
- None of these metrics overcome the problem of being based on rated items only (and the inherent variation that comes from this limitation)

Looking forward …

- Next, we look at rank metrics, then a bit of a rant on hidden-data evaluation, and then a broader set of metrics to look at business relevance …

5-3: Basic Decision Support Metrics



1
TP
0
0
1%
30%
FP
1