Spring 2009 **CSC 466:** Knowledge Discovery from Data Alexander Dekhtyar

Collaborative Filtering and Recommender Systems

Evaluation

In [2], evaluation measures for recommender systems are separated into three categories:

- **Predictive Accuracy Measures.** These measures evaluate how close the recommender system came to predicting actual rating/utility values.
- **Classification Accuracy Measures.** These measures evaluate the frequency with which a recommender system makes correct/incorrect decisions regarding items.
- **Rank Accuracy Measures.** These measures evaluate the correctness of the ordering of items performed by the recommendation system.

Predictive Accuracy Measures

Mean Absolute Error (MAE). Mean Absolute Error measures the average deviation (error) in the predicted rating vs. the true rating. Let $u(c,s)$ be the true ratings, and $u^p(c,s)$ be the ratings predicted by a recommender system. Let $W = \{(c,s)\}$ be a set of *user-item* pairs for which the recommender system made predictions. Then, the mean absolute error, denoted $|\bar{E}|$, is defined as follows:

$$|\bar{E}| = \frac{\sum_{(c,s)\in W} |u^p(c,s) - u(c,s)|}{|W|}$$

Variations include:

Mean Squared Error. Mean squared error punishes *big mistakes* more severely.

$$|\bar{E^2}| = \frac{\sum_{(c,s)\in W} (u^p(c,s) - u(c,s))^2}{|W|}$$

Root Mean Squared Error. A variant of mean squared error.

$$|\sqrt{\bar{E^2}}| = \sqrt{|\bar{E^2}|} = \sqrt{\frac{\sum_{(c,s)\in W}(u^p(c,s) - u(c,s))^2}{|W|}}$$

Normalized Mean Absolute Error (NMAE). This measure normalizes MAE by the range of available rating values. Let $r_{\min}$ be the smallest possible rating and $r_{\max}$ be the largest possible rating. Then, NMAE is defined as follows:

$$\text{NMAE} = \frac{|\bar{E}|}{r_{\max} - r_{\min}} = \frac{1}{|W|}\frac{\sum_{(c,s)\in W}|u^p(c,s) - u(c,s)|}{r_{\max} - r_{\min}}$$

Mean Absolute Error on the extermes. Consider the range $[r_{\min}, r_{\max}]$ of all possible values of the ratings. Select a notion of *extreme positive* and *extreme negative* ratings: pick two more numbers $r_{neg} < r_{pos}$, such that:

$u(c,s) \in [r_{\min}, r_{neg}]$ are your *extreme negative ratings*;
$u(c,s) \in (r_{neg}, r_{pos})$ are your *relatively neutral ratings*;
$u(c,s) \in [r_{pos}, r_{\max}]$ are your *extreme positive ratings*.

Compute MAE for extreme (positive and negative) ratings only.

Properties of Predictive Accuracy Measures.

Advantages. Predictive accuracy measures have a number of important benefits.

- **Measure actual predictions.** Predictive accuracy measures assess the accuracy of the actual predictions.
- **Induce order.** Using predictive accuracy measures one can order all predictions.
- **Easy to compute.** All predictive accuracy measures can be computed efficiently.
- **Known Statistical Properties.** MAE, and MAE-based error estimates have *well-known* statistical properties that allow for straightforward significance testing of differences in accuracy of different recommender systems.

Disadvantages.

- **Too specific.** Recommender systems that output results to users usually output ranked results, or simply a set of recommendations. Accuracy (or inaccuracy) of actual ratings may be the wrong way to measure the success of recommendations.
- **Too sensitive.** Ratings systems with low-granularity rating scales may not require absolutely correct predictions.

Classification Accuracy Measures

Classification Accuracy measures apply to evaluations of recommender systems which make granular decisions about user-item pairs: e.g., *Recommend/ Do not recommend*. The measures evaluate the frequency of the system making correct/incorrect decisions.

Precision and Recall. To use these metrics, recommender system must convert its ratings scale into a binary $\{\text{Do not recommend}, \text{Recommend}\}$ scale. Items for which the prediction is to *recommend* are shown to the user, other items — are not shown. The transition mechanism is up to recommender systems.

Each item can be either *relevant* or *irrelevant* to the user. We get, therefore, the following matrix:

	Recommended	Not Recommended	Total
Relevant	RR	RN	$R = RR + RN$
Not Relevant	FP	NN	$IR = FP + NN$
Total	$REC = RR + FP$	$NREC = RN + NN$	$N = R + IR = REC + NREC$

Precision is the fraction of all recommended items that are relevant.

$$precision = \frac{RR}{RR + FP} = \frac{RR}{REC}$$

Recall is the fraction of all relevant items that were recommended.

$$recall = \frac{RR}{RR + RN} = \frac{RR}{R}$$

F-measure. Recall and precision measure different facets of the accuracy of the recommender system. They can be combined in a single quantity, the **F-measure**:

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

ROC curves. ROC (*relative operating characteristic* or *reciever operating characteristic*) curves measure the ability of theinformation filtering system to tell signal (relevant user-item pairs) from noise (items that are irrelevant for users).

Idea: Assume that there is a probability distribution associated with the predicted level of relevance for relevant and irrelevant items. The better the system, the *more different* the two probability distributions are.

ROC curves are constructed as follows.

- Rank all recommendations by the rating score.
- For each rating cut-off point (a.k.a., for each position in the ranked list):
  - Compute recall;
  - Compute **fallout**:

$$Fallout = \frac{FP}{REC} = \frac{FP}{RR + FP}$$

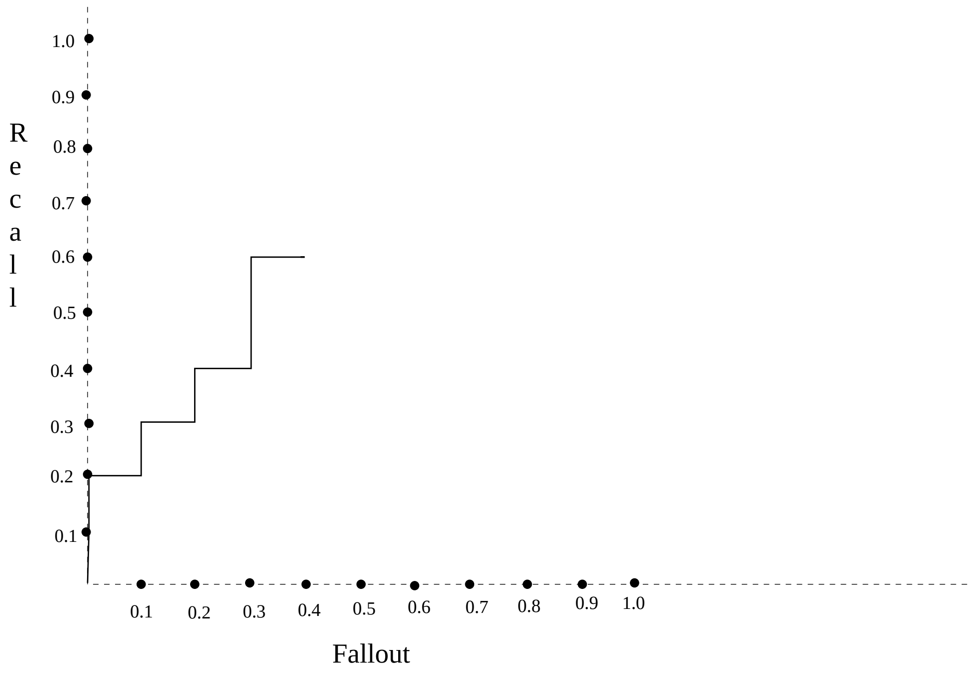

Figure 1: Recall vs. Fallout ROC curve for the example.

- Plot recall vs. fallout

Example. Consider the following list of ten recommendations:

Position	Recommendation	Rellevant?
1	s_1	Yes
2	s_2	Yes
3	s_3	No
4	s_4	Yes
5	s_5	No
6	s_6	Yes
7	s_7	No
8	s_8	Yes
9	s_9	Yes
10	s_{10}	No

Assume also, that the are a total of 10 relevant recommendations that can be given. The ROC curve for this dataset is shown in Figure 1.

The ROC area (**Swet's measure**) is defined as the **area under the ROC curve**.

Features

Advantages

- **Appropriate for practical systems/empirical system evaluation.** These measures can be used to evaluate the actual performance of a recommender system w.r.t. a specific user/set of users.
- **Well-established measures**. Precision, recall, F-measure, ROC are all well-established measures with known properties.

- **Single number.** (for ROC). ROC is a robust single-number measure.

Disadvantages

- **Ground truth.** Accuracy prediction measures require knowledge of actual rating values. However, classification prediction measures require knowing whether a specific recommendation was found to be relevant by and end-user. This may be difficult to obtain.
- **Need for large data set.** These measures may require evaluationon large sets of date to really provide good intuition.
- **Insensitivity to ordering.** While ROC curves allows one to *observe* the effects of the ranking order of recommendation, swaps in ranking often preserve the measures.

Rank Accuracy Measures

A third view of the task of a recommender system is that it *ranks* all items w.r.t. a user (or ranks all user-item pairs), such that higher-ranked recommendations are more likely to be relevant to users. Individual rating predictions may be incorrect, but as long as the order is caught correctly, rank accuracy measures will evaluate the system as having high accuracy.

Prediction-Rating Correlation. If a variance of one variable can be explained by the variance in another, the two variables are said to correlate.

Let $s_1, \ldots s_n$ be items and let $u_1, \ldots, u_n \in \{1, \ldots, n\}$ be their *true* order rank. Let recommender system predict the ranks $u_1^p, \ldots, u_n^p$ for these items (i.e., u_i is the true rank of the item and u_i^p is the predicted rank). Let $\bar{u}$ be the mean of $u_1, \ldots, u_n$, and $\bar{u^p}$ be the mean of $u_1^p, \ldots, u_n^p$ The **Spearman ρ correlation** is defined as follows:

$$\rho = \frac{\sum_{i=1} n(u_i - \bar{u})(u_i^p - \bar{u^p})}{n \cdot stdev(u) \cdot stdev(u^p)}.$$

Kendall's Tau. Consider the rankings $u_1, \ldots, u_n$ and $u_1^p, \ldots, u_n^p$ defined above. Let C be the number of *concondant pairs*, i.e., correctly predicted pairs of rankings. Let D be the number of *discordant pairs* - pairs, whose rankings were predicted incorrectly. Let TR be the number of pairs of items in the true ordering that have tied ranks and TP be the number of pairs of items in the predicted ordering that have tied ranks. **Kendall's Tau** measure is defined as:

$$Tau = \frac{C - D}{\sqrt{(C + D + TR)(C + D + TP)}}.$$

Half-life utility measure. The half-life utility measure assumes that the user is presented with a long list of recommendations, but will only observe the top few of them. The measure is defined as the difference between the user's rating of an item and the *default* rating of an item, which is usually chosen to be neutral or slightly negative. However, the likliehood that the user will observe a specific item on the ordered list is estimated using a *exponential decay* function, parameterized by a *half-life decay parameter*.

Let $u(c, s_j)$ be user c's rating of item s_j: jth item on the recommendation list. Let α be the half-life decay parameter. Let d be our default rating. The expected utility of item s_j is computed as follows:

$$R_c = \sum_j \frac{\max(u(c, s_j) - d, 0)}{2^{\frac{j-1}{\alpha-1}}}$$

The half-life is the rank of the item on the list, such that there is a 50% chance that the user will view the item.

References

[1] G. Adomavicius, A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, No. 6, June 2005.

[2] J. Herlocker, J. Konstan, L Terveen, J. Reidl. Evaluating Collaborative Filtering Recommender Systems, *ACM Transaction on Information Systems*, Vol 22, No. 1, January 2004, pp. 5–53.