Offline Evaluation and Metrics Quiz

TOTAL POINTS 11

1. Why would you use a different metric for evaluating prediction vs. top-N recommendation?

1 point

○ Because prediction metrics are usually on a 1 to 5 scale, and you need a larger scale for top-N metrics.

○ Because you need different algorithms to compute predictions vs. top-N recommendation.

● Because predictions are mostly about accuracy and error within a particular item, while top-N is mostly about ranking and comparisons between items.

○ Because predictions are a harder problem; recommendations are just suggestions and can never be wrong.

2. Which of the following is an advantage of nDCG *compared with Spearman rank correlation*?

1 point

○ nDCG doesn’t care about the range of ratings a user uses.

● Ranking accuracy at the top of the list is weighted more heavily than accuracy further down the list.

○ nDCG only considers the order of items, not the numeric scores the recommender gives them.

○ In nDCG Large moves (e.g., off by 10 positions) are penalized more than small moves.

3. Which of the following statements about diversity metrics is **not true**?

1 point

● A recommendation list with high diversity will have a mix of highly-scored and lower-scored items near the top.

○ The goal of measuring and tuning diversity is to prevent over-specialization into a narrow portion of the product or item space.

○ Diversity measures how much the top items in a recommendation list vary.

○ Diversity metrics generally use a separate measure of similarity, either pairwise or for a list.

4. In some top-n evaluations, instead of considering all items, the recommender recommends from the items the user has rated/consumed/purchased plus a random subset of all items. Why is this useful?

1 point

● The evaluation can only judge whether the returned items are among the rated/consumed/purchased ones; having too many other items just increases the number of desirable but not-yet-consumed items, making it harder to tell whether the recommendations are good.

○ To make the results a well-formed random sample for statistical analysis purposes.

○ It is useful to evaluate both accuracy and decision-support in a recommender. If all of the items are available to recommend, that makes the user's decision-making much harder.

○ Most recommender algorithms are more accurate over a smaller set of candidate items, so this reduction makes it easiest to obtain a desired level of accuracy.

5. Which of the following is a true statement about why someone might prefer to use RMSE (Root Mean Squared Error) instead of MSE (Mean Squared Error) or MAE (Mean Absolute Error)?

1 point

○ RMSE penalizes all errors the same, regardless of size, while MAE penalizes large errors more than small ones.

○ RMSE can be negative or positive, while both MSE and MAE are always positive.

○ RMSE is expressed in the same units as the ratings, unlike MSE.

● RMSE is expressed in the same units as the ratings, unlike MAE.

6. When computing serendipity, we depend upon a prior "primitive" estimate of obviousness and a determination of whether a recommended item is actually relevant. Why do we need these measures?

1 point

● Because serendipity is measuring the degree to which an algorithm is giving recommendations for non-obvious, but still relevant, products or items.

○ Because serendipity scores are measuring how broad a set of items can be recommended -- for instance, recommending books from different genres and authors.

○ Because serendipity is trying to measure the degree to which an algorithm recommends things the user doesn't want, but that the system is trying to push or sell to the user.

○ Because serendipity scores measure the degree to which a user has tastes that differ from the average overall user taste--i.e., to which the user prefers non-popular items.

7. What is the **major** problem of offline evaluation with unary data? 1 point

● If the recommender picks something the user didn't purchase, we do not know if they didn't like it (bad recommendation) or didn't know about it (potentially great recommendation).

○ Users don't really express unary preferences; just because somebody bought two things doesn't mean she liked them equally well.

○ It is too hard to compute metrics with unary data.

○ It is usually too hard to obtain unary data because users don't understand the concept.

8. When holding out ratings from a user's profile for evaluation, what is the benefit of holding out the last ratings rather than holding out random ratings? 1 point

● It more accurately simulates the recommender's knowledge when the held-out ratings were given.

○ It prevents us from evaluating performance on the user's earliest ratings, which usually aren't very meaningful anyway.

○ It makes the evaluation more deterministic. Non-deterministic evaluation is inherently less useful.

○ The most recent ratings have the least information in them, so we don't lose as much accuracy as we would if we held out earlier ratings.

9. 1 point

You've learned about many techniques for evaluation. We also pointed out that most evaluation techniques do not address the question of whether the items recommended are actually useful recommendations. Instead, those evaluations focus on whether the recommender is successful at retrieving “covered up” old ratings. Which of the following evaluation metrics successfully focuses on whether the recommender can produce recommendations for new items that haven’t already been experienced by the user?

○ Accuracy metrics such as RMSE

○ Decision-support metrics such as top-N precision

○ Rank metrics such as nCDG

● None of the above

10. What is the purpose of decision-support metrics such as reversals, precision, or ROC?

1 point

○ They’re designed to measure the total amount of error in the predictions given by a recommender.

○ They’re designed to measure whether the top-recommended items are indeed the best items available.

● They’re designed to measure how effectively a recommender can be used to distinguish between desirable and undesirable items.

○ They’re designed to measure what percentage of items in the product set users actually like.

11. Which of these statements best explains how we perform an n-fold cross validation for getting a more accurate measure of the accuracy experienced by users in a recommender system?

1 point

○ Randomly withhold n ratings from the dataset. Predict each rating from all other data, and average the results.

○ Pick a random set of test ratings. Divide the remaining ratings into n batches of test data. Train the recommender separately on each batch, and predict the test ratings with each trained recommender. Average the results.

● Divide the data set into n partitions of **users**; hold one partition out as test data; train the recommender on the other partitions. Divide the withheld user data into “query” data used for training and “test” data. Measure the accuracy of predicting the test data from the query data for each user and average.

○ Divide the data set into n partitions of **items**; hold one partition out as test data and train the recommender on the other partitions. Now measure the accuracy of prediction for the withheld items for each user and average them.

□ I, **BAL KRISHNA NYAUPANE**, understand that submitting work that isn’t my own may result in permanent failure of this course or deactivation of my Coursera account.

Learn more about Coursera’s Honor Code