Introduction

Fernando Diaz, Google

Artificial Intelligence in Decision Support

Consumer domains: web search, music recommendation, etc

Specialized domains: legal discovery, clinical decision support, etc.

Accuracy is Not Enough

reducing $\frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$

does not necessarily improve **user satisfaction**

user-focused metrics

objective: emphasize development of tools to help individuals with their goals.

metric definition problem: how to develop a metric that accurately reflects how much an individual is being helped

why does the right metric matter?

Evaluation of recommender has long been divided between accuracy metrics (e.g., precision/recall) and error metrics (notably, RMSE and MAE). The mathematical convenience and fitness with formal optimization methods, have made error metrics like RMSE more popular, and they are indeed dominating the literature. However, it is well recognized that accuracy measures may be a more natural yardstick, as they directly assess the quality of top-N recommendations.

This work shows, through an extensive empirical study, that the convenient assumption that an error metric such as RMSE can serve as good proxy for top-N accuracy is questionable at best. There is no monotonic relation between error metrics and accuracy metrics. This may call for a re-evaluation of optimization goals for top-N systems. On the bright side we have presented simple and efficient variants of known algorithms, which are useless in RMSE terms, and yet deliver superior results when pursuing top-N accuracy.

task: recommendation

metric: minimize error of predicted ratings

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. RecSys, 2010.

why does the right metric matter?

In this paper, we have made a first attempt to systematically investigate the correlation of automatic ROUGE scores with human evaluation for meeting summarization. Adaptations on ROUGE setting based on meeting characteristics are proposed and evaluated using Spearman's rank coefficient. Our experimental results show that in general the correlation between ROUGE scores and human evaluation is low, with ROUGE SU4 score showing better correlation than ROUGE-1 score. There is significant improvement in correlation when disfluencies are removed and speaker information is leveraged, especially for evaluating system-generated summaries. In addition, we observe that the correlation is affected differently by those factors for human summaries and system-generated summaries.

task: summarization

metric: target summary word recall

Feifan Liu and Yang Liu. Correlation between ROUGE and human evaluation of extractive meeting summaries. ACL, 2008

why does the right metric matter?

task: recommendation

metric: user clicks

> Most recommendation engines today are based on predicting user engagement, e.g. predicting whether a user will click an item or not. However, there is potentially a large gap between engagement signals and a desired notion of *value* that is worth optimizing for [Ekstrand and Willemsen, 2016]. Just because a user engages with an item doesn't mean they value it. A user might reply to an item because they are angry about it, or click an item in order to gain more information about it [Wen et al., 2019], or watch addictive videos out of temptation.

Smitha Milli, Luca Belli, and Moritz Hardt. From optimizing engagement to measuring value. 2020.

intrinsic vs. extrinsic evaluation

intrinsic evaluation: measuring the performance of a *component* of a system, independent of how it contributes to the end task (and often independent of the user)

extrinsic evaluation: measuring the performance of a *component* of a system, in the context of how it contributes to the end task and user.

Karen Spärck Jones and Julia R. Galliers. Evaluating Natural Language Processing Systems: An Analysis and Review,, 1995.

situated versus simulated environments

simulated evaluation: some aspects of evaluation are constructed in controlled settings (e.g. user models, labels, log data); most “offline evaluation”.

situated evaluation: almost all aspects of the evaluation are observed only as the system is used (e.g. production tests); most “online evaluation”.

task-oriented domains

summarization

extraction

retrieval

recommendation

what we will cover

offline metrics

online/behavioral metrics

multiple metrics

open problems

what we will not cover

optimization of behavior toward a metric

aggregation population level metrics

experimentation