Sequence-Aware Recommender Systems

Tutorial at TheWebConf 2019, San Francisco

Paolo Cremonesi | Politecnico di Milano, Italy
Massimo Quadrana | Pandora, USA
Dietmar Jannach | University of Klagenfurt, Austria

Evaluation

Agenda

- 09:00 – 09:45 Introduction & Problem Definition
- 09:45 – 10:30 Algorithms I
- 10:30 – 11:00 Coffee break
- 11:00 – 11:30 Algorithms II
- 11:30 – 12:00 Evaluation
- 12:00 – 12:20 Hands-on
- 12:20 – 12:30 Conclusion / Questions

Evaluation approaches

- Some common strategies
  - Field test (A/B test): Run two or more algorithms in parallel in a real-world application. Optimize for suitable (business) metric.
  - Laboratory study (user study): Let users interact with two or more versions of an application. Compare observed behavior and answers to questionnaires.
  - “Offline” analysis: Learn prediction models on historical data. Evaluate on held-out data.
- Other:
  - Simulations, quasi-experimental designs, exploratory studies.

Offline evaluation

- Usually the approach with the least effort
- Allows for high level of reproducibility
  - In theory at least
- Established evaluation procedures and metrics exist
- But comes with a number of limitations
  - Rather “post-diction” than “prediction”
  - Prediction accuracy measures not necessarily indicative of value for user or provider
  - Computational metrics for other quality factors (novelty, diversity, serendipity) mostly not validated
  - Datasets can be biased

Accuracy evaluation

- Academic research often abstracts
  - from the specifics of the domain and
  - from the purpose of a recommender
- Abstract accuracy measures like RMSE, precision, recall etc. are used
  - Remember the matrix completion problem
- Similar hide-and-predict evaluation schemes can be applied for certain sequence-aware recommenders
  - Allows for the usage of common information retrieval measures

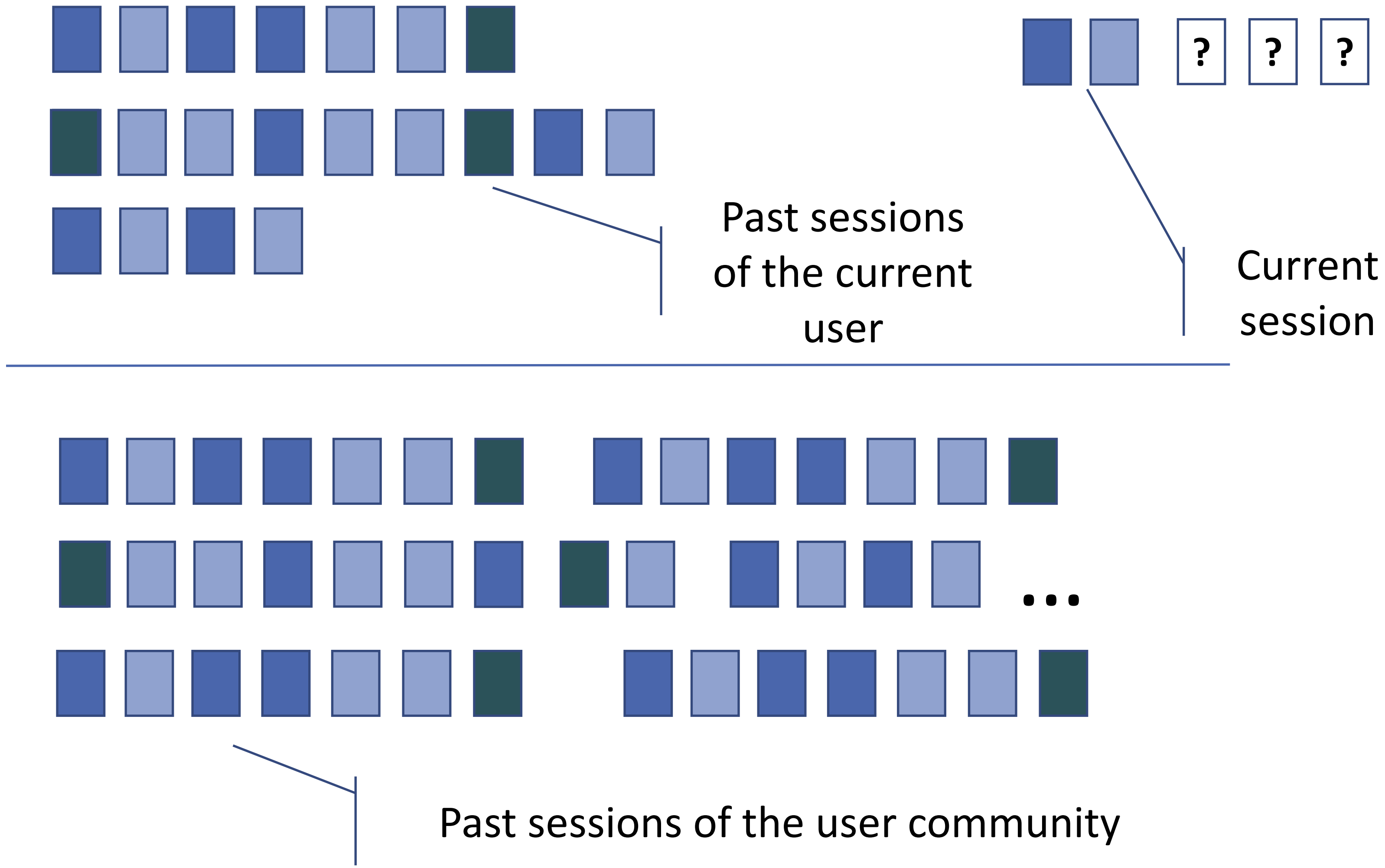
Evaluation of other quality factors

- Diversity, novelty etc. can be assessed in similar ways as in traditional evaluation setups
  - including the usual trade-offs
- For other aspects, no common procedures and measures are established yet
- Example: Reminders and repeated purchases
  - Reminding can be very effective in terms of recall
  - Not clear, however, how much reminding is enough
  - Reminding might also have limited business value

Accuracy: Problem abstraction

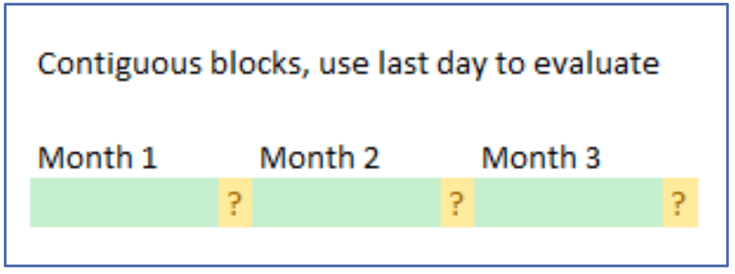
- We consider session-based and session-aware recommendation scenarios in the following
- Reduce the problem to predict subsequent items in a session
  - Makes it irrelevant if the recommender should recommend accessories or alternatives

Problem Abstraction



Evaluation protocols: partitioning

- Creating multiple training and test splits
  - Typical cross-validation cannot be applied due to importance of sequences
  - Alternatives, e.g.,:
    - Sliding window over the data
    - Evaluate on contiguous blocks of data
    - Repeated random subsampling

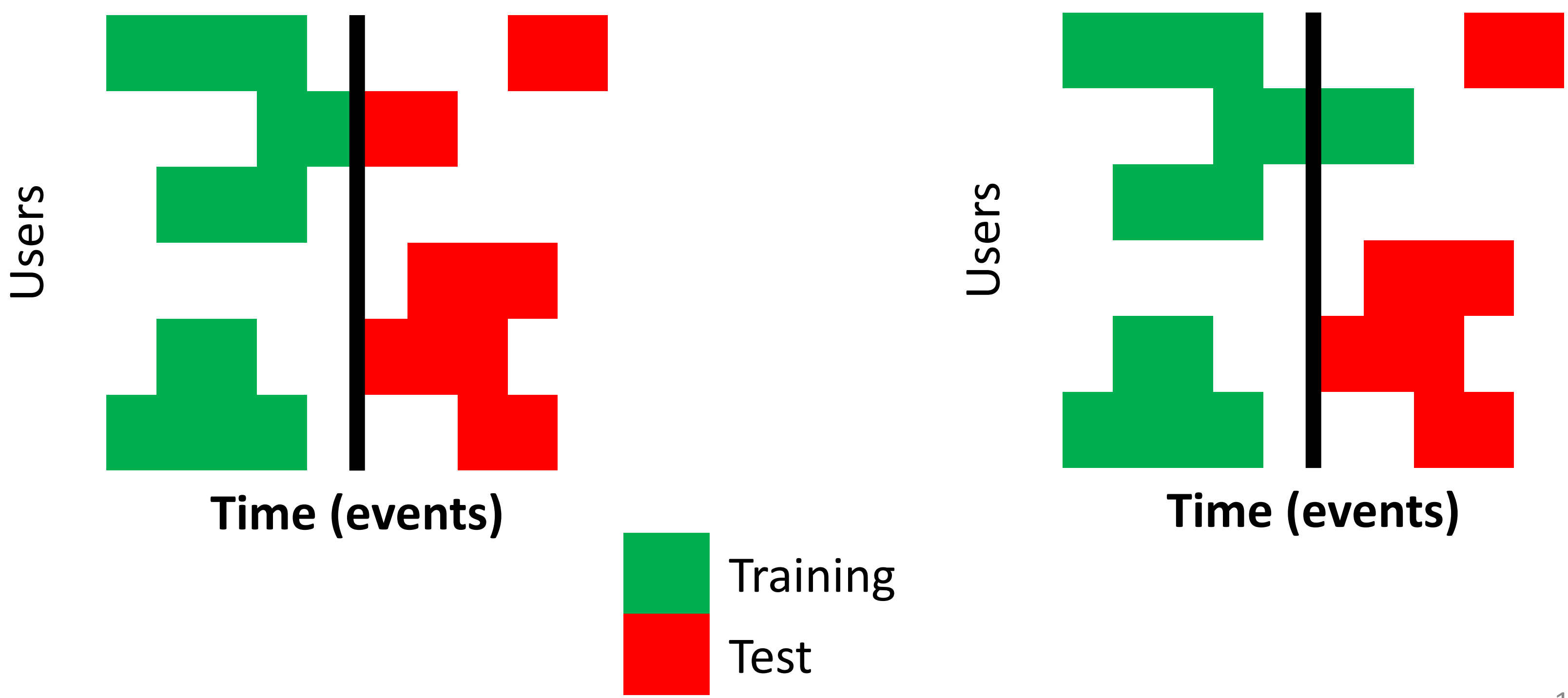


  - Several recent works use a single training-test split

Evaluation protocols: partitioning

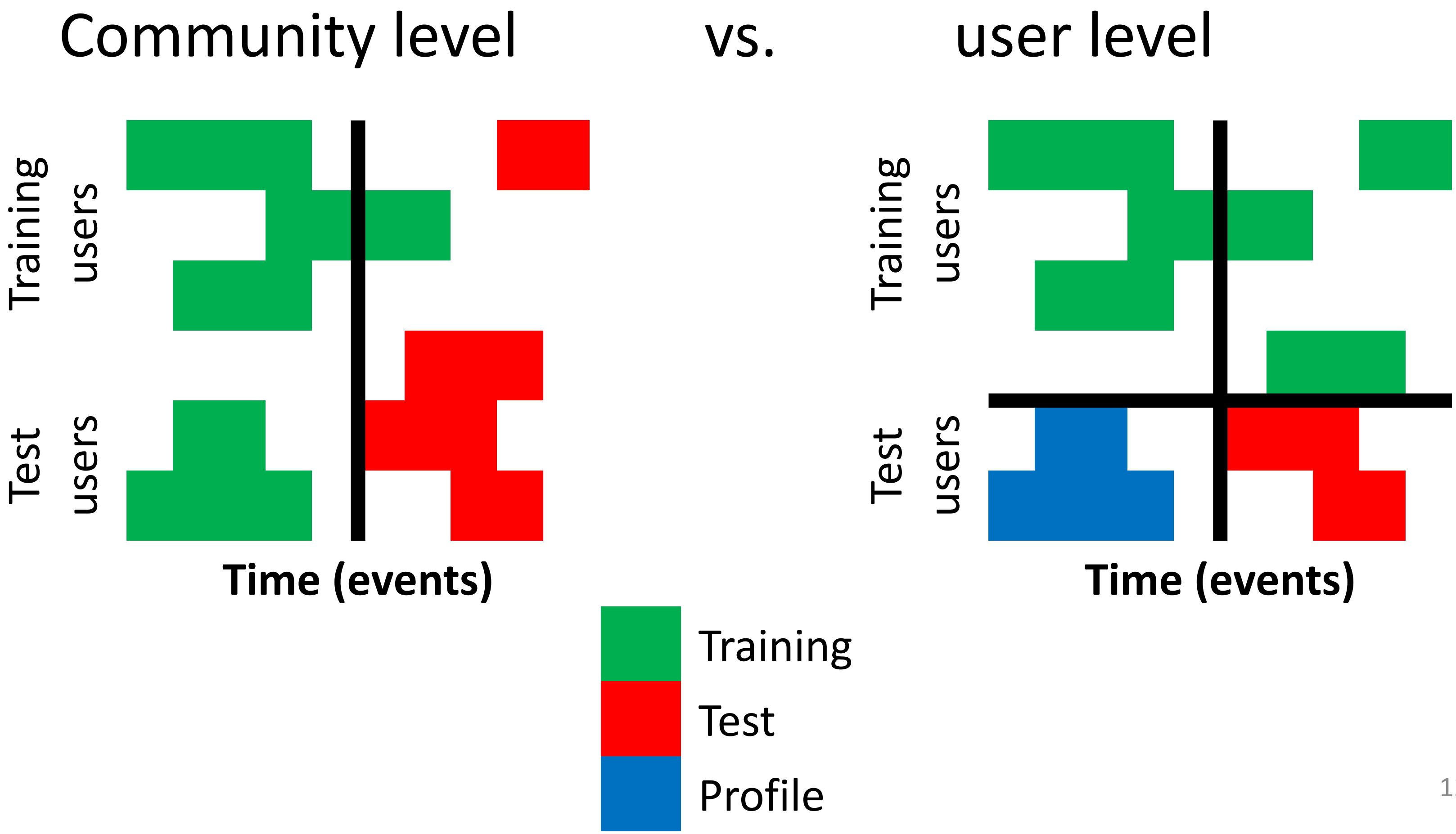
- Splitting criteria

Event level vs. session level



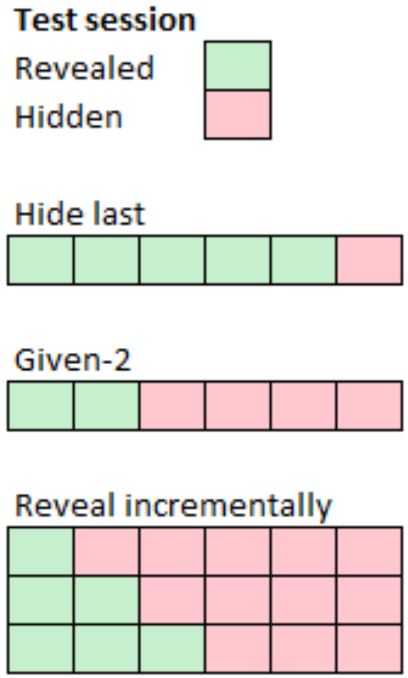
Evaluation protocols: partitioning

- Splitting criteria



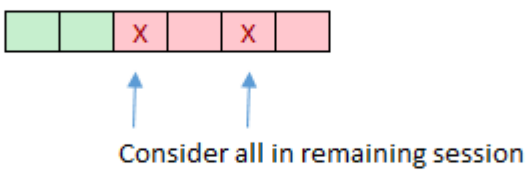
Evaluation protocols: revealing

- Hiding and revealing data in a test session
- Variants from the literature
  - Hide and predict last item in session
  - Given-N evaluation: Reveal the first N items in the session and predict the rest
  - Reveal items incrementally and evaluate after each revealed item
  - Predict only certain types of actions, e.g., purchases but not item views

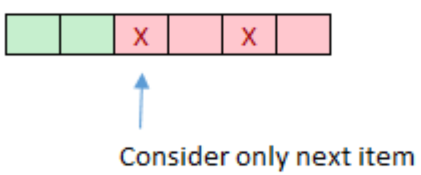


Evaluation protocols: measuring

- Ground truth for comparison
  - Compare top-n recommendations with all hidden items in the current session
    - Apply, e.g., precision, recall, MAP etc.



  - Compare top-n recommendations only with immediate next item in the current session
    - Apply, e.g., hit rate, mean reciprocal rank etc.



A performance comparison

User Modeling and User-Adapted Interaction (2018) 28:331–390
https://doi.org/10.1007/s11257-018-9209-6

CrossMark

Evaluation of session-based recommendation algorithms

Malte Ludewig[1] · Dietmar Jannach[2]

Received: 28 March 2018 / Accepted in revised form: 11 September 2018 / Published online: 1 October 2018

Abstract
Recommender systems help users find relevant items of interest, for example on e-commerce or media streaming sites. Most academic research is concerned with approaches that personalize the recommendations according to long-term user profiles. In many real-world applications, however, such long-term profiles often do not exist and recommendations therefore have to be made solely based on the observed behavior of a user during an ongoing session. Given the high practical relevance of the problem, an increased interest in this problem can be observed in recent years, leading to a number of proposals for *session-based recommendation algorithms* that typically aim to predict the user's immediate next actions. In this work, we present the results of an in-depth performance comparison of a number of such algorithms, using a variety of datasets and evaluation measures. Our comparison includes the most recent approaches based on recurrent neural networks like GRU4REC, factorized Markov model approaches such as FISM or FOSSIL, as well as simpler methods based, e.g., on nearest neighbor schemes. Our experiments reveal that algorithms of this latter class, despite their sometimes almost trivial nature, often perform equally well or significantly better than today's more complex approaches based on deep neural networks. Our results therefore suggest that there is substantial room for improvement regarding the development of more sophisticated session-based recommendation algorithms.

Performance comparison

- Background
  - Some recent neural methods do not outperform very old baselines (if properly tuned) in IR, SIGIR Forum 2018

OPINION

The Neural Hype and Comparisons Against Weak Baselines

Jimmy Lin
David R. Cheriton School of Computer Science, University of Waterloo

Performance comparison

- Background
  - And a similar phenomenon was already observed in 2009

Improvements That Don't Add Up:
Ad-Hoc Retrieval Results Since 1998

Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel

Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

{tgar,alistair,wew,jz}@csse.unimelb.edu.au

Performance comparison

- Background
  - The machine learning hype reveals and emphasizes some existing problems (2018)

Troubling Trends in Machine Learning Scholarship

Zachary C. Lipton* & Jacob Steinhardt*
Carnegie Mellon University, Stanford University
zlipton@cmu.edu, jsteinhardt@cs.stanford.edu

July 27, 2018

Performance comparison

- Background
  - And some of the problems are not new (2012)

Machine Learning that Matters

Kiri L. Wagstaff KIRI.L.WAGSTAFF@JPL.NASA.GOV
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA

Evaluation goal

- Assess the effectiveness of the landmark method "GRU4REC"
  - One of the first neural approaches for session-based recommendation (based on Recurrent Neural Networks)
  - Used in many subsequent studies as a baseline
  - Largely improved since first version (CIKM '18)
- Compare with conceptually much simpler and longer-known methods

Evaluation baselines (selection)

- Association rules (AR) of size two
  - To implement “Customers who bought …”
- Sequential rules (SR) of size two
  - Same as AR, but takes sequence of items into account
- Session-based k-nearest neighbors (SKNN)
  - Look for *k* most similar past sessions
  - Recommend items that appeared in these past sessions
  - Variants with different similarity functions
    - V-SKNN: Puts more weight on later items in the sessions
- Others, including algorithms for *sequential* recommendation

Baselines, scalability

- The AR and SR methods are trivial and rules can be learned by scanning the training data once.
- SKNN methods would not scale in naïve implementation
- Approach for kNN methods:
  - Sampling, e.g., consider only the last few thousand sessions
  - Use data structures that allow us to quickly determine possible neighbors for a given target session
  - Prediction time per recommendation below 30ms

Jannach, D. and Ludewig, M.: "**When Recurrent Neural Networks meet the Neighborhood for Session-Based Recommendation**". In: Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017). Como, Italy, 2017

Datasets

- Different datasets from the e-commerce domain are publicly available today
  - Yoochoose (ACM RecSys ‘15 challenge), Retail Rocket, Diginetica, TMALL
- Media datasets
  - News: CLEF NewsReel Challenge
  - Listening logs: 30Music, Nowplaying
  - Playlists: Art-of-the-mix, last.fm
- Social media
  - XING (ACM RecSys Challenge ‘16/’17), with user IDs
- Non-public datasets
  - E-commerce (Zalando), Music (8Tracks)

Main outcomes (I)

- In almost all configurations and measurements, even the latest version GRU4REC was outperformed by one of the simple methods
- For example, when using precision and recall

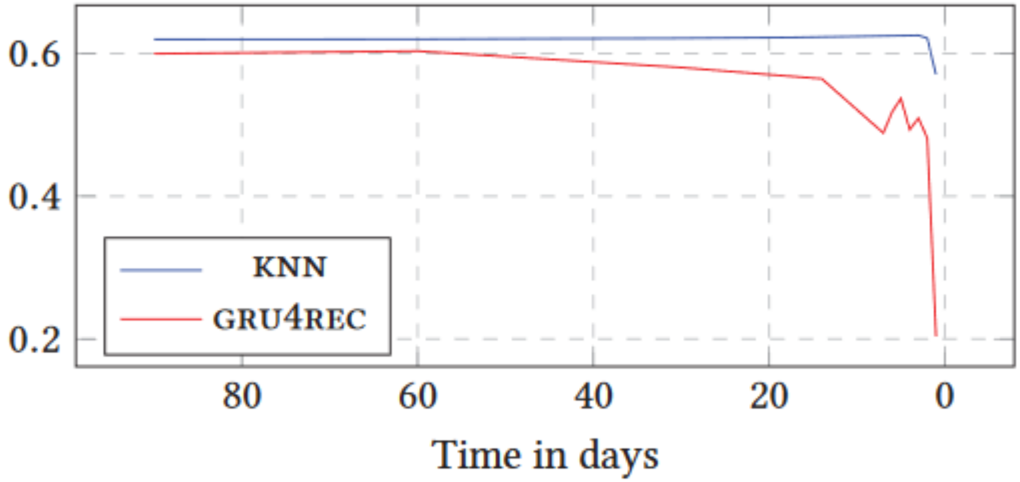
Dataset	RSC15		TMALL		ROCKET		ZALANDO	
Metric	P@20	R@20	P@20	R@20	P@20	R@20	P@20	R@20
SKNN	0.086	0.464	**0.095**	**0.312**	**0.056**	**0.478**	0.074	0.202
V-SKNN	**0.092**	0.494	0.088	0.291	0.055	0.462	**0.076**	**0.207**
SMF	0.092	**0.501**	0.068	0.230	0.047	0.397	0.062	0.175
GRU4REC	0.085	0.470	0.068	0.233	0.046	0.400	0.065	0.181
SR	0.089	0.488	0.052	0.193	0.038	0.342	0.060	0.174

Main outcomes (II)

- “Proving” progress is simple and difficult at the same time
  - There is no consistent ranking of the algorithms across the datasets
  - The ranking furthermore depends on the particular measurement method (consider only next or all) and the evaluation metric (hit rate or Mean Reciprocal Rank)

Main outcomes (3)

- Domain-specifics can play a role
- For the e-commerce datasets, it is for example sufficient to retain only the last few days for training



More on domain specifics

- Findings for the e-commerce domain:
  - Short-term intents are much more important than long-term preference models
  - Reminding users can be beneficial both in terms of business value as well for offline accuracy
  - Short-term trends in the consumer community can be leveraged for improved recommendations
  - Recommending items that are currently on sale (discounted) can be effective

Jannach, D., Ludewig, M. and Lerche, L.: "**Session-based Item Recommendation in E-Commerce: On Short-Term Intents, Reminders, Trends, and Discounts**". User-Modeling and User-Adapted Interaction, Vol. 27(3-5). Springer, 2017, pp. 351-392

Improvements that don’t add up

- Follow-up study with newer neural approaches for session-based recommendation
  - Publications from CIKM ‘17, KDD ‘18, CIKM ‘19, WSDM ’19
  - Most of them claim to outperform GRU4REC
- Integrated into common evaluation framework

Improvements that don't add up

Table 5: Results for E-commerce Datasets

Metrics	MAP@20	P@20	R@20	HR@20	MRR@20
			RETAIL		
S-KNN	**0.0283**	**0.0532**	**0.4707**	**0.5788**	0.3370
VS-KNN	0.0278	0.0531	0.4632	0.5745	**0.3395**
GRU4REC	0.0272	0.0502	0.4559	0.5669	0.3237
NARM	0.0239	0.0440	0.4072	0.5549	0.3196
STAMP	0.0229	0.0428	0.3922	0.4620	0.2527
AR	0.0205	0.0387	0.3533	0.4367	0.2407
SR	0.0194	0.0362	0.3359	0.4174	0.2453
NEXTITNET	0.0173	0.0320	0.3051	0.3779	0.2038
CT	0.0162	0.0308	0.2902	0.3632	0.2305
			DIGI		
S-KNN	**0.0255**	**0.0596**	**0.3715**	**0.4748**	0.1714
VS-KNN	0.0249	0.0584	0.3668	0.4729	**0.1784**
GRU4REC	0.0247	0.0577	0.3617	0.4639	0.1644
NARM	0.0218	0.0528	0.3254	0.4188	0.1392
STAMP	0.0201	0.0489	0.3040	0.3917	0.1314
AR	0.0189	0.0463	0.2872	0.3720	0.1280
NEXTITNET	0.0149	0.0380	0.2416	0.2922	0.1424
CT	0.0115	0.0294	0.1860	0.2494	0.1075
SR	0.0113	0.0296	0.1856	0.2349	0.1044

All hope is lost?

- Much room for improvement for neural approaches that only use the item IDs
- Hybrid approaches often seem effective
  - Combination of simple and complex techniques
  - Usage of content information about items
  - Leveraging context information

News Session-Based Recommendations using Deep Neural Networks

Gabriel de Souza Pereira Moreira*
CI&T
Campinas, SP, Brazil
gabrielpm@ciandt.com

Felipe Ferreira
Globo.com
Rio de Janeiro, RJ, Brazil
felipe.ferreira@corp.globo.com

Adilson Marques da Cunha
Brazilian Aeronautics Institute of Technology - ITA
São José dos Campos, SP, Brazil
cunha@ita.br

Reflection

- General issues of applied machine learning
  - Blind “obsession” with accuracy measures
    - Use them whether the measurement is meaningful or not
    - No justification for the choice of metric or cut-off thresholds
  - Limited reproducibility
    - Algorithm code increasingly shared, but not the code for data preprocessing and optimization
  - Choice of baselines
    - Recent non-neural approaches are often not considered. We re-start with new neural baselines that are sometimes not strong.
  - Missing tuning of baselines
    - Leading to pseudo-progress

Future Directions

- Reproducibility
  - Should be very easy in our scientific discipline
  - Publish data and code (including pre-processing and evaluation code)
- Focus on problems that matter
  - Improving 1% on an seemingly arbitrarily chosen accuracy measure on an arbitrarily chosen dataset does not help
    - In particular when the baseline is badly chosen and not optimized.
  - Several studies show that higher accuracy does not necessarily translate into better perceived or more effective recommendations

Future directions

- Considering multiple quality factors and domain-specifics
  - Everyone searches the single best model for a given class of problems, but this does not exist
- Grow the methodological repertoire and understand which recommendations create value
  - User studies
  - Simulation studies
  - Field tests

Example

- Recent studies on the quality perception of different next-track music recommendations
  - Nearest-neighbor methods also lead to recommendations that people like and consider as suitable continuations
  - Neural method falls behind
  - Spotify’s recommendations receive fewer likes, but are very helpful for discovery
  - Discovery as a main quality factor in the domain
  - Optimizing for more “likes” is misleading
  - Offline accuracy vs. user perception trade-offs

Hands-On

git.io/fxTtV

Overall

- Discussed the family of sequence-aware recommenders
- Proposed categorization
- Reviewed algorithmic approaches
- Discussed evaluation aspects and open issues

Thank you for your attention!

- Questions?