RecoXplainer: An Extensible Toolkit for Explainable Recommender Systems

Ludovik **Coba**, Roberto **Confalonieri**, Markus **Zanker**

MQ2 Tutorial at AAAI-21

unibz
Fakultät für Informatik
Facoltà di Scienze e Tecnologie informatiche
Faculty of Computer Science

Resources

- Web page and slides
  http://www.inf.unibz.it/~rconfalonieri/aaai21/
- Repository
  https://github.com/ludovikcoba/recoxplainer

Outline

- Part I: An introduction to Explainable Recommender Systems
- Part II: RecoXplainer

Outline

- An introduction to Explainable Recommender Systems
  - Explainability/Explainable AI
  - Recommender Systems
  - Explanations in Recommender Systems

Outline

- RecoXplainer
  - Overview of the Toolkit
  - Model-based Explanations
  - Post-hoc Explanations
  - Evaluation of Explanations
  - Hands-on Session

Part II
RecoXplainer

RecoXplainer

A library for generating explainable recommendations implemented in Python

- Model-based and Post-hoc explanation algs.
- Standardized evaluation protocol (quality of explanations)

RecoXplainer

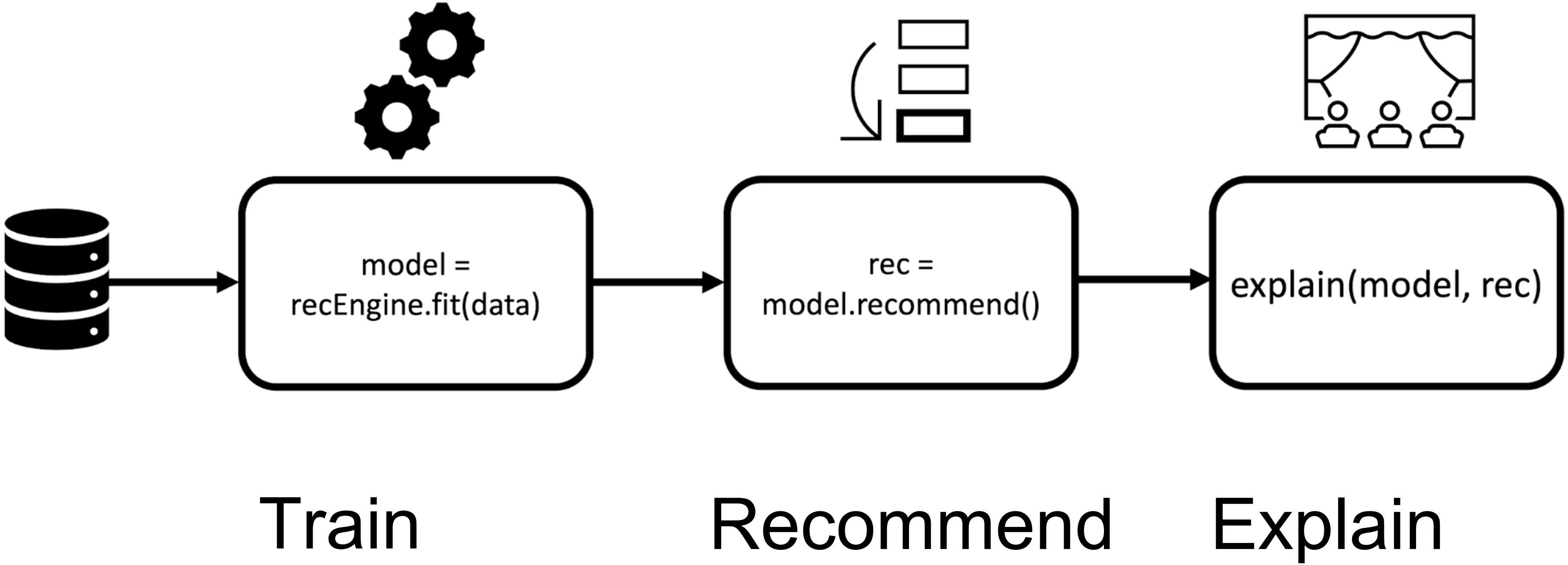
A library for generating explainable recommendations implemented in Python

- Model-based and Post-hoc explanation algs.
- Standardised evaluation protocol (quality of explanations)

(Main) Characteristics:

- Unified, extendable, easy-to-use
- Replication and reproduction of best practices

RecoXplainer – Overview



RecoXplainer - Recommenders

Collaborative recommender algorithms:

- Alternative Least Square (ALS)
- Bayesian Personalised Ranking (BPR-MF)
- Generalised Matrix Factorisation (GMF)
- Multi-Layer Perceptron (MLP)

Alternative Least Square

Alternative Least Square was introduced to solve the implicit feedback prediction problem

$$\mathcal{L}(\hat{R}) = \mathcal{L}(P,Q) = \sum_{ui}(r_{ui} - p_u \cdot q_i^T)^2 + \lambda(\|p_u\|_F^2 + \|q_i\|_F^2)$$

Optimisation loop:

1. Initialising users and items latent representations
2. Solving least-squares for P and then for Q (alternated)
3. Repeat 2. until convergence

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In 2008 Eighth IEEE International Conference on Data Mining. IEEE, 263–272. https://doi.org/10.1109/ICDM.2008.22

Bayesian Personalised Ranking

BPR-MF is an optimisation criterion that aims to find a personalised total order $>_u \subset I^2$ for any user $u \in U$ and pairs of item $(i, j) \in I^2$

$$\mathcal{L}(\hat{R}) = \mathcal{L}(P, Q) = \sum_{u \in U, i \in I_u^+, j \in I/I_u^+} ln(\sigma(\hat{r}_{ui} - \hat{r}_{uj})) - \lambda(\|p_u\|_F^2 + \|q_i\|_F^2)$$

$p(i >_u j) = \sigma(r'_{ui} - r'_{uj})$, where r'_{ui} is a predicted user interaction defined as the product $p_u \cdot {q_i}^T$

S. Rendle, C. Freudenthaler, Z. Gantner, S-T. Lars. 2009. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of UAI '09). 452– 461. http://portal.acm.org/citation.cfm?id=1795114.1795167

Generalised Matrix Factorisation

GMF adapts MF to a neural network

Given latent feature vectors p_u and q_i

$$\phi(p_u, q_i) = p_u \odot q_i$$

A prediction is calculated as

$$\hat{r}_{ui} = a_{out}(\mathbf{h}^T(p_u \odot q_i))$$

X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T-S. Chua. 2017. Neural Collaborative Filtering. In Proc. of WWW ’17. 173–182. https://doi.org/10.1145/3038912.3052569

Multi-Layer Perceptron

MLP learns to predict new items by first concatenating q_i and p_u

$$z_1 = \phi_1(p_u, q_i) = \begin{bmatrix} p_u \\ q_i \end{bmatrix}$$

Followed by several hidden layers

$$z_2 = \phi_2(z_1) = f_2(W_2^T z_1 + b_2),$$
$$\ldots\ldots$$
$$z_N = \phi_N(z_{N-1}) = f_N(W_N^T z_{N-1} + b_N).$$

A prediction is calculated as

$$\hat{r}_{ui} = \sigma(h^T \odot z_N)$$

X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T-S. Chua. 2017. Neural Collaborative Filtering. In Proc. of WWW ’17. 173–182. https://doi.org/10.1145/3038912.3052569

RecoXplainer - Explainability

Model-based

- ALS Explain
- Explainable Matrix Factorisation

Post-hoc Explanations (via proxy)

- kNN
- Association Rules

Model-based Explainability

Model-based explanations are obtained by constraining the loss function

- ALS Explain, Explainable Matrix Factorization

Model-based Explainability

Model-based explanations are obtained by constraining the loss function

- ALS Explain, Explainable Matrix Factorization

Pros:

- No interpretable proxies needed

Cons:

- Model loses flexibility

ALS Explain

An explanation method that leverages the linearity present in the matrix factorization and the update rules

A prediction is generated as a 'linear combination' of past interactions (*item-style explanation*)

Requested as an additional feature to SPARK

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In 2008 Eighth IEEE International Conference on Data Mining. IEEE, 263–272. https://doi.org/10.1109/ICDM.2008.22

ALS Explain

Traditional ALS $\min \sum_{i,j \in R} c_{ij}(r_{ij} - u_i v_j^T)^2 + \beta(\|u_i\|^2 + \|v_j\|^2)$

ALS Explain is obtained by:

- Replacing user factors with item factors

$$\hat{r_{ij}} = v_j^T u_i = v_j^T (V^T C^i V + \beta I)^{-1} V^T C^i p(i))$$

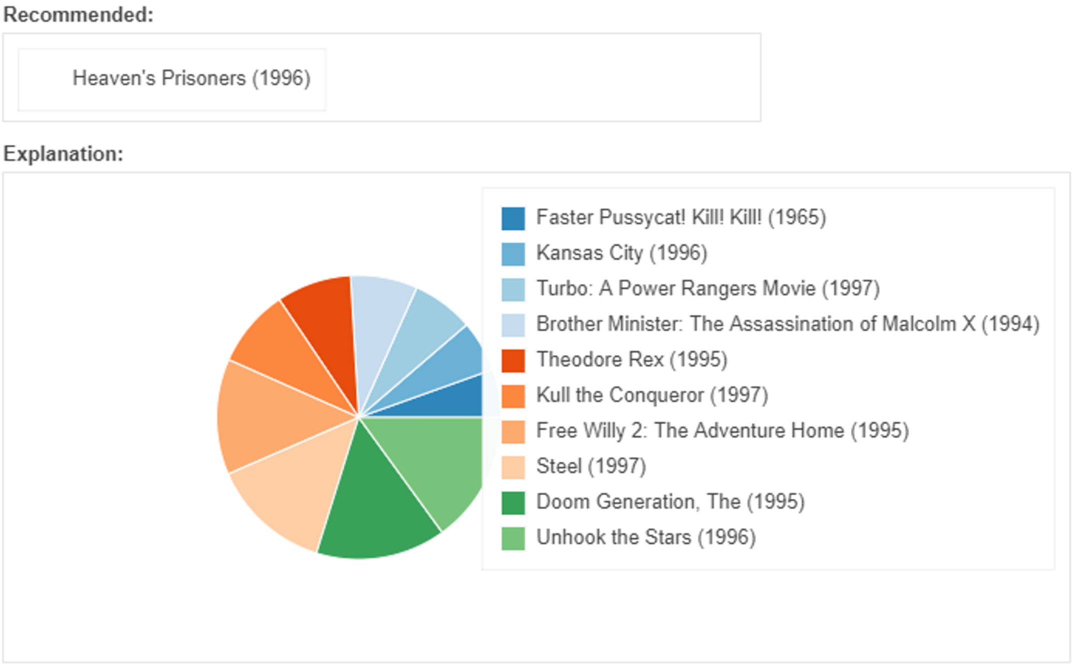
- Defining $W^i = (V^T C^i V + \beta I)^{-1}$ and $s_{jk}^i = y_j^T W^i y_k$

- A prediction is generated as $\hat{r}_{ij} = \sum_{k: r_{ik} > 0} s_{jk}^i c_{ik}$

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In 2008 Eighth IEEE International Conference on Data Mining. IEEE, 263–272. https://doi.org/10.1109/ICDM.2008.22

ALS Explain

"Explains recommended items based on **(similar) previously** interacted items"



Explainable MF

It adds an extra **soft-constraint** to the traditional Matrix Factorization formula

Soft-constraint holds the information of how explainable **item** j is for **user** i (based on how frequently an item j has been highly rated)

It generates *user-style explanation* (item-style also supported)

First implementation in Python (to the best of our knowledge)

B. Abdollahi and O. Nasraoui. 2017. Using Explainability for Constrained Matrix Factorization. In Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys ’17. 79–83. https://doi.org/10.1145/3109859.3109913

Explainable MF

- Determining the **similarity** between two users

$$\mathrm{sim}(u,v) = \frac{\sum_{\forall i \in I_u \cap I_v} (r_{u,i} - \overline{r}_u)(r_{v,i} - \overline{r}_v)}{\sqrt{\sum_{\forall i \in I_u} (r_{u,i} - \overline{r}_u)^2} \sqrt{\sum_{\forall i \in I_v} (r_{v,i} - \overline{r}_v)^2}}$$

users

item

Catalogue of u

rating of user v on i

mean rating of user v

- NN(u) is the set of N users with highest similarity

B. Abdollahi and O. Nasraoui. 2017. Using Explainability for Constrained Matrix Factorization. In Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17. 79–83. https://doi.org/10.1145/3109859.3109913

Explainable MF

Extra soft-constraint to the traditional matrix factorisation formula

$$\min \sum_{i,j \in R} (r_{ij} - u_i v_j^T)^2 + \frac{\beta}{2}(\|u_i\|^2 + \|v_j\|^2) + \lambda \|u_i - v_j\|^2 E_{ij},$$

E_{ij} tells how explainable item j is for user i measuring how frequently an item j has been highly rated

$$E_{ij} = \sum_{\substack{\forall r \in R \\ r \geq P_\tau}} r * |NN^k(i)_{jr}|,$$

$NN^k(i)_{jr}$ corresponds to the set of nearest neighbours of target user i who 'positively' rated j (r above P_τ threshold)

B. Abdollahi and O. Nasraoui. 2017. Using Explainability for Constrained Matrix Factorization. In Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17. 79–83. https://doi.org/10.1145/3109859.3109913

Explainable MF – User-style

“Explains recommended items based on **similar users**”

You were recommenderd ItemID-985 because similar users to you rated this item as follows:

Rating	Similar users’ ratings
★	0
★★	0
★★★	0
★★★★	11
★★★★★	22
Average Rating:	4.5 ★★★★½

Explainable MF

$$min \sum_{i,j \in R} (r_{ij} - u_i v_j^T)^2 + \frac{\beta}{2}(\|u_i\|^2 + \|v_j\|^2) + \underbrace{\lambda\|u_i - v_j\|^2 E_{ij},}_{\substack{\text{Soft constraint} \\ \text{(explainability)}}}$$

Popularity bias

Explainable MF

$$min \sum_{i,j \in R} (r_{ij} - u_i v_j^T)^2 + \frac{\beta}{2}(\|u_i\|^2 + \|v_j\|^2) + \lambda \|u_i - v_j\|^2 E_{ij},$$

$$\underbrace{+\delta \|u_i - v_j\| N_{ij}}$$

Soft constraint
(novelty)

dissimilarity

$$\frac{\sum_{\forall k \in I_i} d(i,k)}{|I_i|}$$

Post-hoc Explanations

Post-hoc explanations of a **black-box model** are obtained by means of an **interpretable proxy**

Black-box algorithms: ALS, BPR, GMF, MLP

Interpretable proxies: Association Rules, kNN

Post-hoc Explanations

Post-hoc explanations of a **black-box model** are obtained by means of an **interpretable proxy**

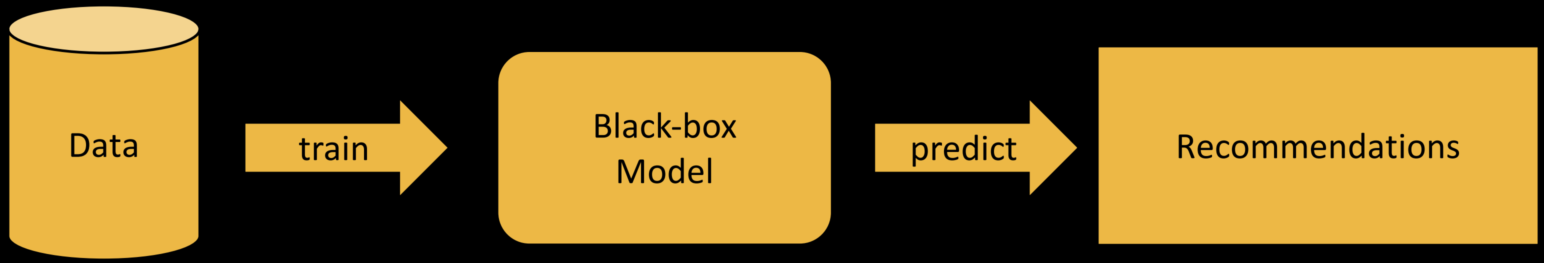
Pros:

- No under-the-hood reworking of the black-box

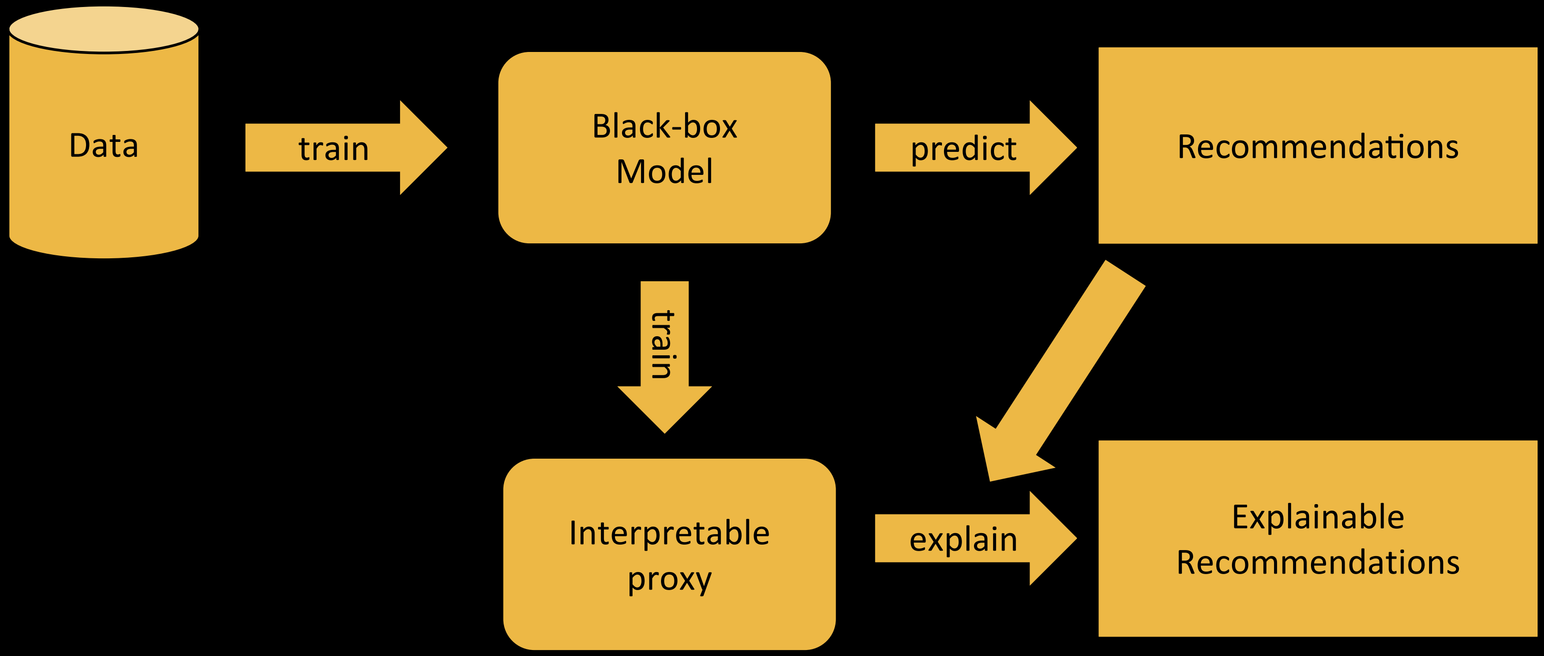
Cons:

- Additional training step, not complete
- Accuracy-interpretability trade-off

Post-hoc Explanations



Post-hoc Explanations



Proxies – Association Rules

Association rule mining algorithms

Detect rules of the form $X \rightarrow Y$ (e.g., beer → diapers) from a set of transactions $T = \{t_1, t_2, \ldots t_n\}$ over a catalogue I

Measure quality by means of **support**, **confidence** used as a threshold to cut off unimportant rules

G. K. Sarwar, J. Konstan, J. Riedl. 2000. Analysis of recommendation algorithms for e-commerce. 158–167. https://doi.org/10.1145/352871.352887

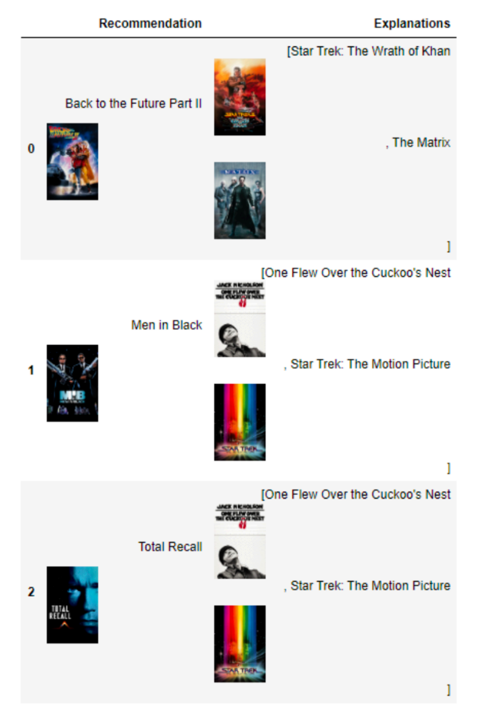
Proxies – Association Rules

Association rules to generate post-hoc explanations

Mine association rules on the generated predictions from a black-box RS

For each user filter the learned transactions such that antecedents are in the training set and consequents are unseen or non-interacted items

The resulting subset is ranked by support/confidence/lift. We keep the top-D consequents



G. Peake, J. Wang. 2018. Explanation mining: Post-hoc interpretability of latent factor models for recommendation systems. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018), 2060–2069. https://doi.org/10.1145/3219819.3220072

Proxies - kNN

kNN identifies the *k*-most similar items for each target *t* and ranks them according to aggregated similarities

Uses cosine similarity $sim(\vec{i}, \vec{j}) = cos(\vec{i}, \vec{j}) = (\vec{i} \cdot \vec{j})/(|\vec{i}| * |\vec{j}|).$

$NN(i)$: neighborhood of an item i choosing the items with the highest similarity value

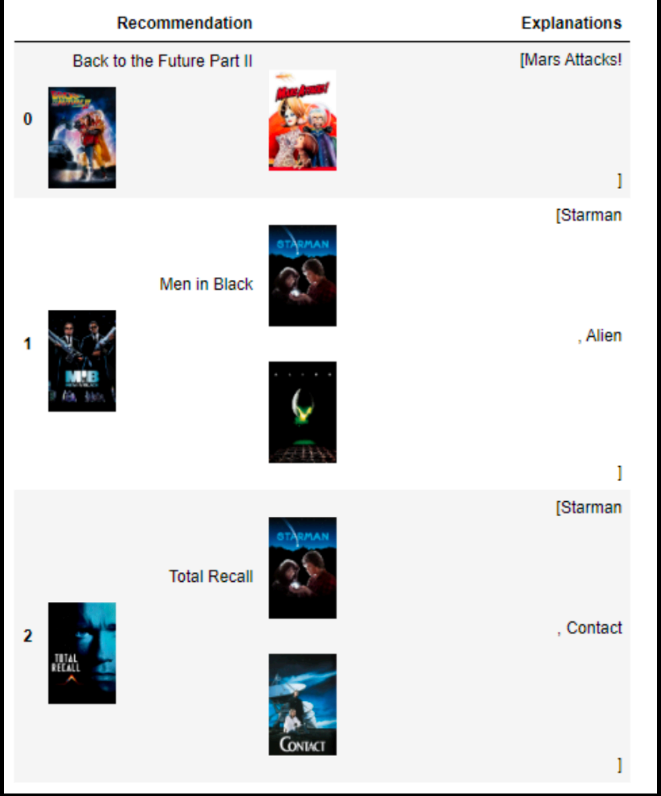
A prediction is generated as

$$p_{u,i} = (\sum_{j \in NN(i)} sim(\vec{i}, \vec{j}) * R_{u,j})/(\sum_{j \in NN(i)} sim(\vec{i}, \vec{j}))$$

Proxies - kNN

kNN to generate post-hoc explanations

- For each user and recommendation from the black-box model find the kNN items
- Filter the neighbours to be in the training set of the user
- Filter only unseen interactions, and use the similarity score to rank items
- Draw the top-D predictions and their corresponding explanations



	Recommendation	Explanations
0	Back to the Future Part II	[Mars Attacks!]
1	Men in Black	[Starman , Alien]
2	Total Recall	[Starman , Contact]

Evaluation

Formal definition of interpretability is used as a proxy for quantifying the explanation quality

RecoXplainer (currently) supports two categories of offline Evaluation Metrics

- Mean Explainability Precision (MEP)
- Model Fidelity

F. Doshi-Velez and B. Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (2 2017). http://arxiv.org/abs/1702.08608

Mean Explainability Precision

It evaluates if a model behaves as expected

Given a recommendation list L_u for a given user u:

$$MEP = \frac{1}{|U|} \times \sum_{u \in U} \frac{|\{i : i \in L_u, E_{ui} > 0\}|}{N},$$

Where U is the set of users, and E_{ui} is a formalisation of the definition of interpretability

B. Abdollahi and O. Nasraoui. 2017. Using Explainability for Constrained Matrix Factorization. In Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys ’17. 79–83. https://doi.org/10.1145/3109859.3109913

Model Fidelity

It evaluates explainability via the proxy

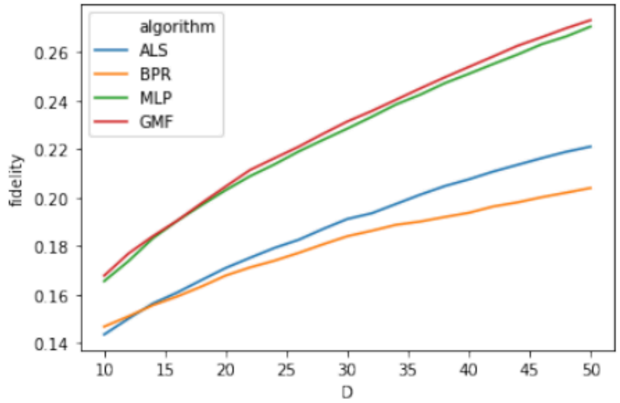
It measures the 'faithfulness' of the proxy to the black-box model:

$$\text{Model Fidelity} = \frac{|L \cap ProxPred|}{|L|},$$

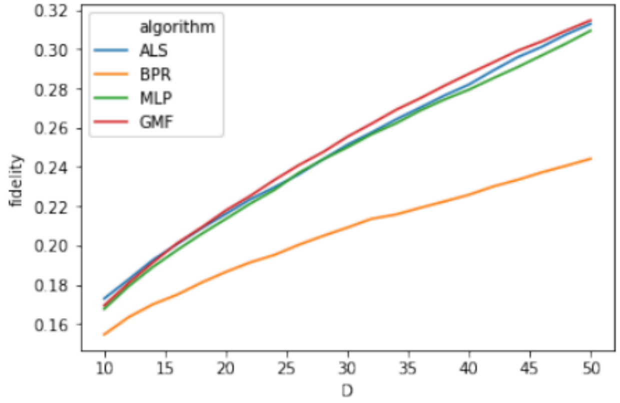
L are the recommendations from the black-box model, and $ProxPred$ are the proxy predictions.

G. Peake and J. Wang. 2018. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018), 2060–2069. https://doi.org/10.1145/3219819.3220072

Model Fidelity



(a) Association Rules



(b) Nearest neighbours (kNN)

Thank you!
Questions?

Part II
Hands-on Session

Hands-on Session

Jupyter notebooks

1. ALS explain
2. Explainable Matrix Factorisation
3. Post-hoc Explanations
4. Extensions

Wrapping up

RecoXplainer: a unified, extendable, easy-to-use Python library to develop explainable RecSys

Code available at: https://github.com/ludovikcoba/recoxplainer

Looking for use-cases

Who we are

Dr. Ludovik **Coba**
ludovik.coba@unibz.it

Dr. Roberto **Confalonieri**
rconfalonieri@unibz.it

Prof. Markus **Zanker**
markus.zanker@unibz.it

Thank you!
Questions?