Counterfactual Evaluation for Recommender Systems

Nicolò Felicioni
nicolo.felicioni@polimi.it

POLITECNICO MILANO 1863

RecSys Lab.

Counterfactual Evaluation

- Counterfactual reasoning: thinking about alternatives to events that have already occurred

Counterfactual Evaluation

- Counterfactual reasoning: thinking about alternatives to events that have already occurred
- Intersection between Machine Learning and Causal Inference

Counterfactual Evaluation

- Counterfactual reasoning: thinking about alternatives to events that have already occurred
- Intersection between Machine Learning and Causal Inference
- The theory can be applied in a lot of domains:

Counterfactual Evaluation

- Counterfactual reasoning: thinking about alternatives to events that have already occurred
- Intersection between Machine Learning and Causal Inference
- The theory can be applied in a lot of domains:
    - RecSys/Computational advertisement

Counterfactual Evaluation

- Counterfactual reasoning: thinking about alternatives to events that have already occurred
- Intersection between Machine Learning and Causal Inference
- The theory can be applied in a lot of domains:
    - RecSys/Computational advertisement
    - Medicine

Counterfactual Evaluation

- Counterfactual reasoning: thinking about alternatives to events that have already occurred
- Intersection between Machine Learning and Causal Inference
- The theory can be applied in a lot of domains:
  - RecSys/Computational advertisement
  - Medicine
  - Economics (2021 Nobel prize)
  - etc.

Counterfactual Evaluation

- Counterfactual reasoning: thinking about alternatives to events that have already occurred
- Intersection between Machine Learning and Causal Inference
- The theory can be applied in a lot of domains:
    - RecSys/Computational advertisement
    - Medicine
    - Economics (2021 Nobel prize)
    - etc.
- Today we focus on RecSys Evaluation

Evaluation in RecSys

Evaluation in RecSys

- Suppose we are running an online platform with a recommender

Evaluation in RecSys

- Suppose we are running an online platform with a recommender
- A recommender is already up and running (**baseline**)

Evaluation in RecSys

- Suppose we are running an online platform with a recommender
- A recommender is already up and running (**baseline**)
- The R&D team develops a new recommender

Evaluation in RecSys

- Suppose we are running an online platform with a recommender
- A recommender is already up and running (**baseline**)
- The R&D team develops a new recommender
- We would like to assess its quality (in terms of some metric)

Online Evaluation

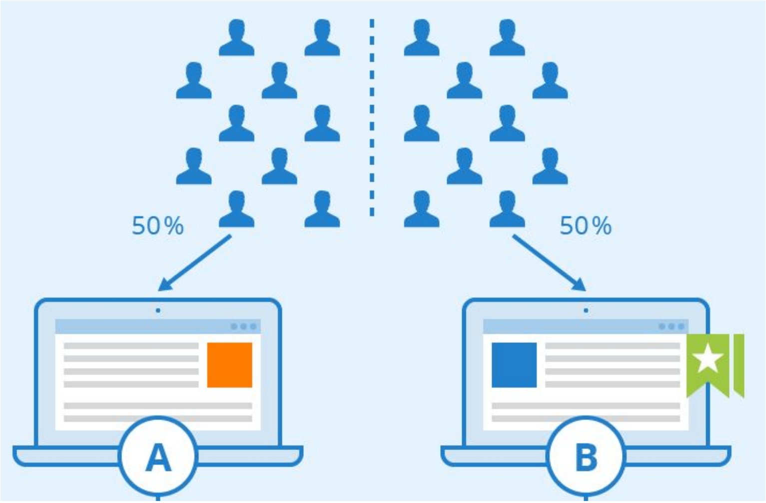
- First idea: use the new recommender with real users

Online Evaluation

- First idea: use the new recommender with real users
- This is called **Online** Evaluation

Online Evaluation

- First idea: use the new recommender with real users
- This is called **Online** Evaluation
- Main way to do it: A/B testing



Online Evaluation

Online Evaluation

✅ Statistically sound procedure

Online Evaluation

✅ Statistically sound procedure

❌ High risk

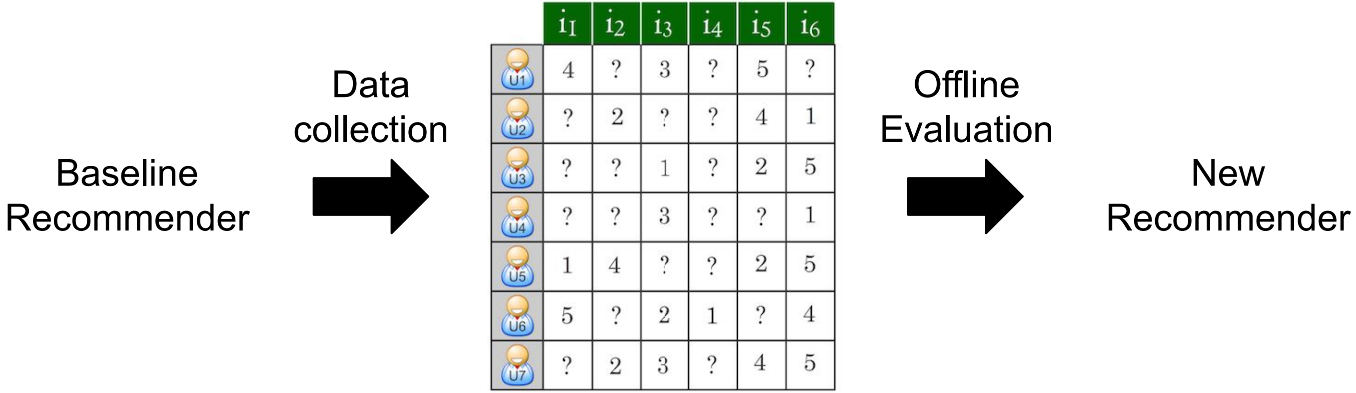
- Providing bad recommendations to real user can harm the platform!
- We would like to evaluate the recommender <u>*offline*</u> first

Offline Evaluation

- First, let the baseline recommender collect data
- Then, evaluate the new recommender on the offline logged dataset

Offline Evaluation

- First, let the baseline recommender collect data
- Then, evaluate the new recommender on the offline logged dataset



	i_1	i_2	i_3	i_4	i_5	i_6
U1	4	?	3	?	5	?
U2	?	2	?	?	4	1
U3	?	?	1	?	2	5
U4	?	?	3	?	?	1
U5	1	4	?	?	2	5
U6	5	?	2	1	?	4
U7	?	2	3	?	4	5

Offline Evaluation

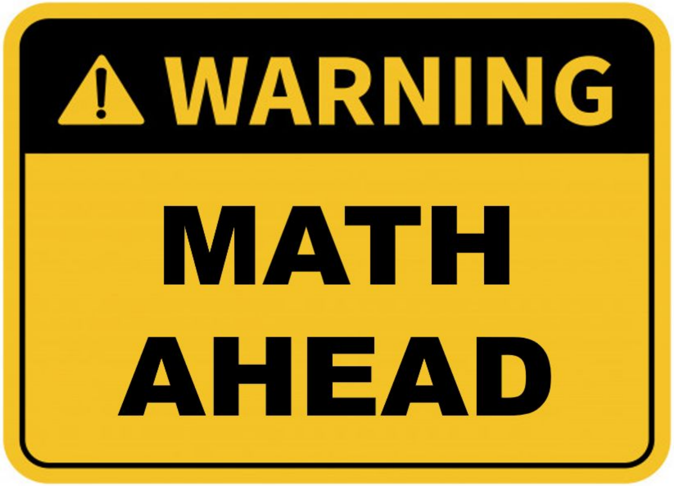
Offline Evaluation

✅ A lot cheaper

Offline Evaluation

- ✅ A lot cheaper
- ❌ Potentially high bias
    - The matrix may be heavily influenced by the baseline recommender!
    - Furthermore, ratings are not Missing-at-Random

A realistic example


WARNING
MATH
AHEAD

Example

- We run a movie streaming platform

Example

- We run a movie streaming platform

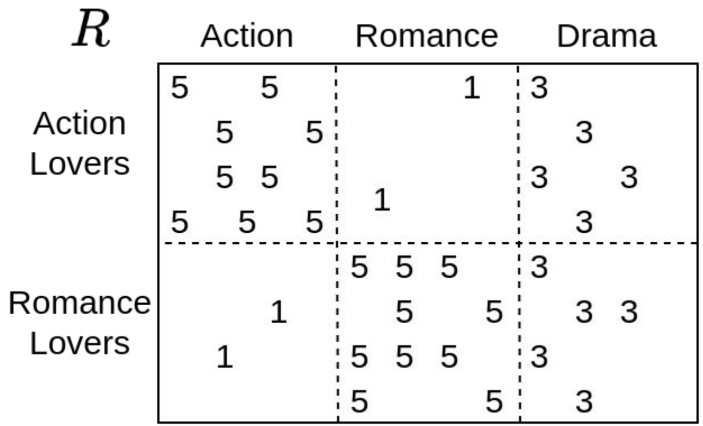
G	Action	Romance	Drama
Action Lovers	5	1	3
Romance Lovers	1	5	3

Example

- We run a movie streaming platform

G	Action	Romance	Drama
Action Lovers	5	1	3
Romance Lovers	1	5	3

- We collect the URM from the baseline recommender:



Example

- We would like to compare two recommenders (Mean Absolute Error)

G	Action	Romance	Drama
Action Lovers	5	1	3
Romance Lovers	1	5	3

Example

- We would like to compare two recommenders (Mean Absolute Error)

G	Action	Romance	Drama
Action Lovers	5	1	3
Romance Lovers	1	5	3

$\hat{R}_1$	Action	Romance	Drama
Action Lovers	5	1	5
Romance Lovers	1	5	5

Example

- We would like to compare two recommenders (Mean Absolute Error)

G	Action	Romance	Drama
Action Lovers	5	1	3
Romance Lovers	1	5	3

$\hat{R}_1$	Action	Romance	Drama
Action Lovers	5	1	5
Romance Lovers	1	5	5

$\hat{R}_2$	Action	Romance	Drama
Action Lovers	5	5	3
Romance Lovers	5	5	3

Example

- We would like to compare two recommenders (Mean Absolute Error)

G	Action	Romance	Drama
Action Lovers	5	1	3
Romance Lovers	1	5	3

$\lvert G - \hat{R}_1 \rvert$	Action	Romance	Drama
Action Lovers	0	0	2
Romance Lovers	0	0	2

$\lvert G - \hat{R}_2 \rvert$	Action	Romance	Drama
Action Lovers	0	4	0
Romance Lovers	4	0	0

Example

- But we don't know G, we can only use the **URM**

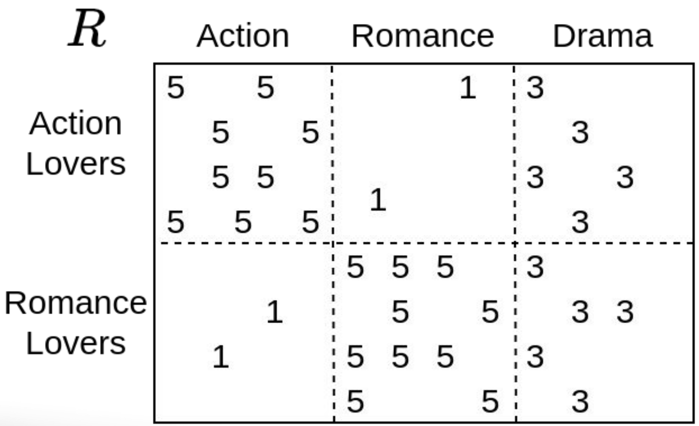
R	Action							Romance							Drama						
Action Lovers	5				5								1		3						
			5				5										3				
			5		5										3				3		
	5			5			5		1								3				
Romance Lovers								5		5		5			3						
					1					5				5			3		3		
			1					5		5		5			3						
								5						5			3				

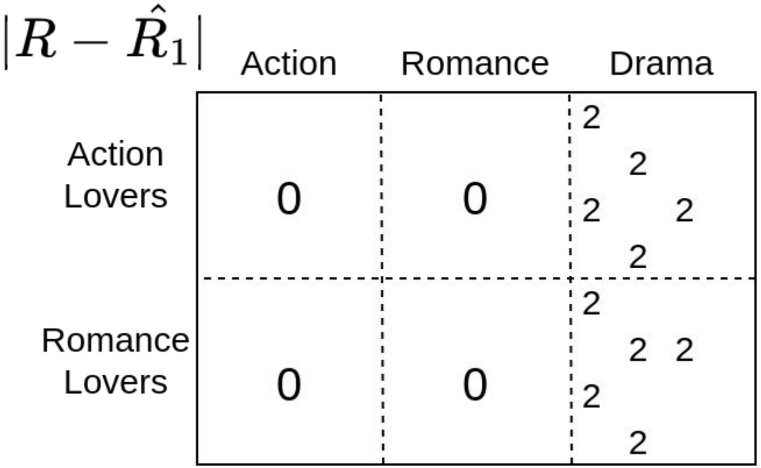
$\hat{R}_1$	Action	Romance	Drama
Action Lovers	5	1	5
Romance Lovers	1	5	5

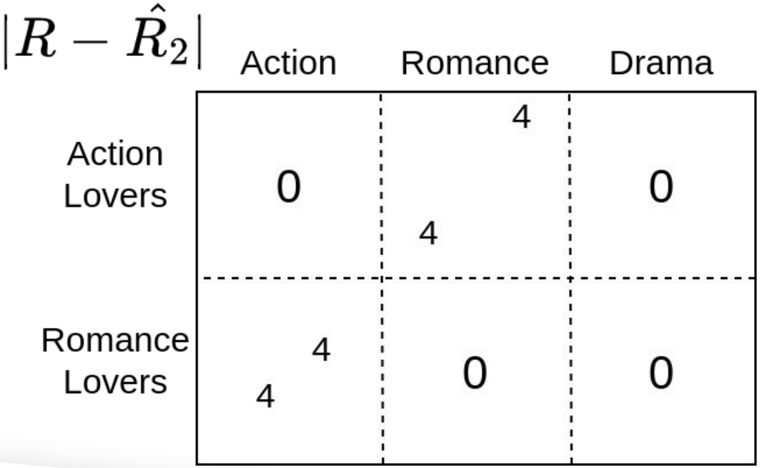
$\hat{R}_2$	Action	Romance	Drama
Action Lovers	5	5	3
Romance Lovers	5	5	3

Example

- But we don't know G, we can only use the **URM**







Example

- But we don't know G, we can only use the **URM**

R

	Action	Romance	Drama
Action Lovers	5 5 5 5 5 5 5 5 5	1 1	3 3 3 3 3
Romance Lovers	1 1	5 5 5 5 5 5 5 5 5 5	3 3 3 3 3

$|R - \hat{R}_1|$

	Action	Romance	Drama
Action Lovers	0	0	2 2 2 2 2
Romance Lovers	0	0	2 2 2 2 2

20

$|R - \hat{R}_2|$

	Action	Romance	Drama
Action Lovers	0	4 4	0
Romance Lovers	4 4	0	0

16

What is the problem?

Why do we obtain wrong results with offline evaluation with the observed URM?

What is the problem?

Why do we obtain wrong results with offline evaluation with the observed URM?

- Missing values are not Missing-at-Random!

What is the problem?

Why do we obtain wrong results with offline evaluation with the observed URM?

- Missing values are not Missing-at-Random!
- The baseline recommender may have biased the data collection

What is the problem?

Why do we obtain wrong results with offline evaluation with the observed URM?

- Missing values are not Missing-at-Random!
- The baseline recommender may have biased the data collection
- Users tend to rate items they like

How to debias the results?

If we knew the probability of observing a rating:

P	Action	Romance	Drama
Action Lovers	0.8	0.1	0.5
Romance Lovers	0.1	0.8	0.5

We could de-bias the results

Counterfactual Evaluation

- What would have happened if the users had observed all the items?

Counterfactual Evaluation

- What would have happened if the users had observed all the items?

- We answer this counterfactual question with the **Inverse Propensity Scoring** (IPS) technique

Counterfactual Evaluation

- What would have happened if the users had observed all the items?
- We answer this counterfactual question with the **Inverse Propensity Scoring** (IPS) technique
- If a rating has a probability p of being observed, we balance this by re-weighting the corresponding error by 1/p

Counterfactual Evaluation

- What would have happened if the users had observed all the items?

- We answer this counterfactual question with the **Inverse Propensity Scoring** (IPS) technique

- If a rating has a probability p of being observed, we balance this by re-weighting the corresponding error by 1/p
    - If a rating has a probability 0.1 of being observed, we multiply the corresponding error by 10

R	Action				Romance				Drama			
Action Lovers	5		5					1	3			
		5		5						3		
		5	5						3		3	
	5	5		5	1					3		
Romance Lovers					5	5	5		3			
			1			5		5		3	3	
		1			5	5	5		3			
					5			5		3		

P	Action	Romance	Drama
Action Lovers	0.8	0.1	0.5
Romance Lovers	0.1	0.8	0.5

$\hat{R}_1$	Action	Romance	Drama
Action Lovers	5	1	5
Romance Lovers	1	5	5

$\hat{R}_2$	Action	Romance	Drama
Action Lovers	5	5	3
Romance Lovers	5	5	3

R

	Action	Romance	Drama
Action Lovers	5 5 5 5 5 5 5 5 5	1 1	3 3 3 3 3
Romance Lovers	1 1	5 5 5 5 5 5 5 5 5 5	3 3 3 3 3

P

	Action	Romance	Drama
Action Lovers	0.8	0.1	0.5
Romance Lovers	0.1	0.8	0.5

$|R - \hat{R}_1|$

	Action	Romance	Drama
Action Lovers	0	0	2 2 2 2 2
Romance Lovers	0	0	2 2 2 2 2

$|R - \hat{R}_2|$

	Action	Romance	Drama
Action Lovers	0	4 4	0
Romance Lovers	4 4	0	0

R	Action	Romance	Drama
Action Lovers	5 5 5 5 5 5 5 5 5	1 1	3 3 3 3 3
Romance Lovers	1 1	5 5 5 5 5 5 5 5 5 5	3 3 3 3 3

P	Action	Romance	Drama
Action Lovers	0.8	0.1	0.5
Romance Lovers	0.1	0.8	0.5

$\lvert R - \hat{R}_1 \rvert_{IPS}$	Action	Romance	Drama
Action Lovers	0	0	4 4 4 4 4
Romance Lovers	0	0	4 4 4 4 4

$\lvert R - \hat{R}_2 \rvert_{IPS}$	Action	Romance	Drama
Action Lovers	0	40 40	0
Romance Lovers	40 40	0	0

R

	Action	Romance	Drama
Action Lovers	5 5 5 5 5 5 5 5 5	1 1	3 3 3 3 3
Romance Lovers	1 1	5 5 5 5 5 5 5 5 5 5	3 3 3 3 3

P

	Action	Romance	Drama
Action Lovers	0.8	0.1	0.5
Romance Lovers	0.1	0.8	0.5

$|R - \hat{R}_1|_{IPS}$

	Action	Romance	Drama
Action Lovers	0	0	4 4 4 4 4
Romance Lovers	0	0	4 4 4 4 4

40

$|R - \hat{R}_2|_{IPS}$

	Action	Romance	Drama
Action Lovers	0	40 40	0
Romance Lovers	40 40	0	0

160

Conclusions

Conclusions

- Online Evaluation: reliable, but risky

Conclusions

- Online Evaluation: reliable, but risky
- Offline Evaluation: can be unreliable if used naively

Conclusions

- Online Evaluation: reliable, but risky
- Offline Evaluation: can be unreliable if used naively
- Counterfactual Evaluation: could improve reliability of offline evaluation in RecSys
    - Missing: what to do when P is not known, theoretical guarantees (bias, variance, convergence), other types of estimators, etc.

Conclusions

- Online Evaluation: reliable, but risky
- Offline Evaluation: can be unreliable if used naively
- Counterfactual Evaluation: could improve reliability of offline evaluation in RecSys
  - Missing: what to do when P is not known, theoretical guarantees (bias, variance, convergence), other types of estimators, etc.
- Interesting lecture on the topic: https://www.youtube.com/watch?v=HMo9fQMVB4w

Thanks for the attention

nicolo.felicioni@polimi.it

RecSys Lab.

POLITECNICO
MILANO 1863