Evaluation

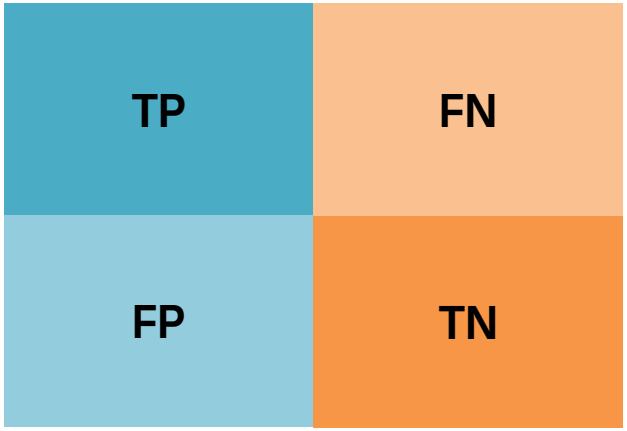
Recommender Systems 2023

Dr.-Ing. Benedikt Loepp

Standard metrics

Precision & recall

- Metrics originally known from **information retrieval**
- Judging quality of a recommendation set
  - *Considered as **positive** by the system (= recommended):*
    - *True positives (TP)*: correctly recommended
    - *False positives (FP)*: incorrectly recommended
  - *Considered as **negative** by the system (= not recommended):*
    - *True negatives (TN)*: correctly not recommended
    - *False negatives (FN)*: incorrectly not recommended



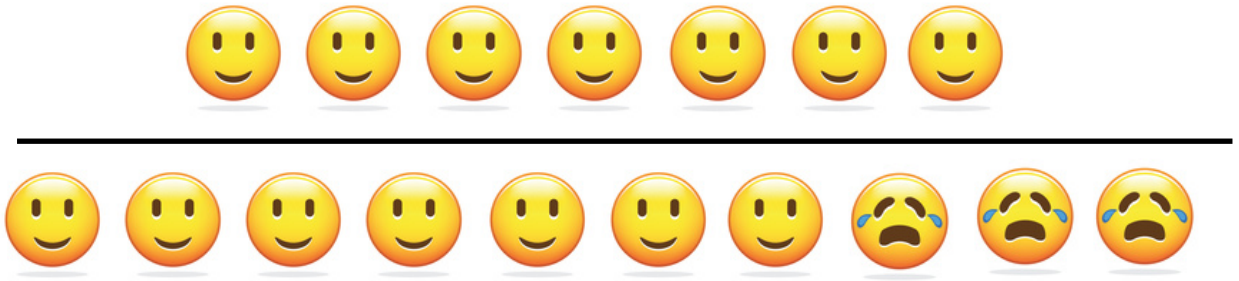
Precision – Example

- **10 recommendations** are presented
  - **7 products** fit the user's taste
  - But **3 products don't**
- Proportion of **correctly recommended** items of **all presented** recommendations

$$precision = \frac{TP}{TP + FP}$$

Precision – Example

- **10 recommendations** are presented
  - **7 products** fit the user's taste
  - But **3 products don't**



$$precision = \frac{TP}{TP + FP}$$

Precision – Example

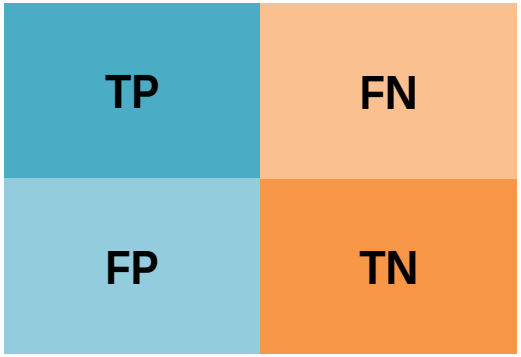
- **10 recommendations** are presented
  - 7 products fit the user's taste
  - But 3 products don't

TP	FN
FP	TN

$$precision = \frac{TP}{TP + FP}$$

Precision – Example

- **10 recommendations** are presented
  - 7 products fit the user's taste
  - But 3 products don't



$$precision = \frac{TP}{TP + FP} = \frac{7}{7 + 3} = 0.7$$

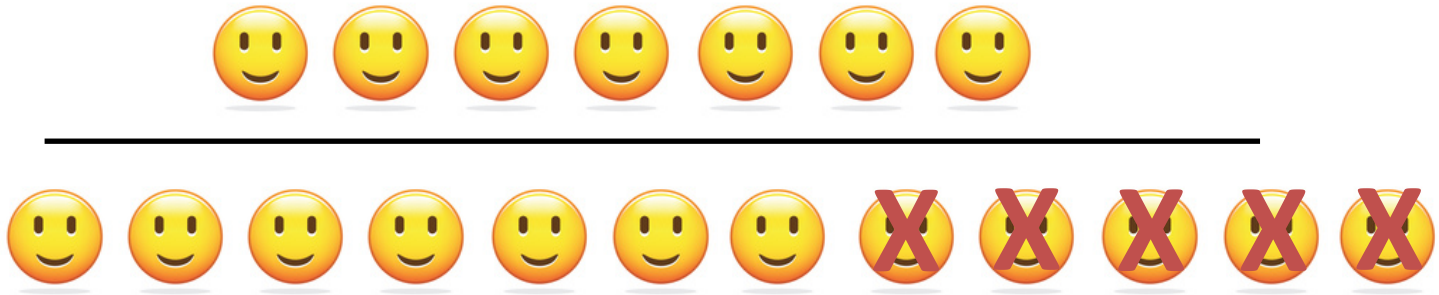
Recall – Example

- System did not recognize **5 other products** that would meet the user's preferences as the **7 correctly presented** products
- Proportion of **correctly recommended** items of **all potentially** recommendable items

$$recall = \frac{TP}{TP + FN}$$

Recall – Example

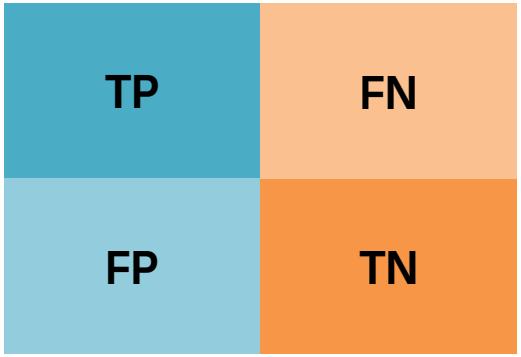
- System did not recognize **5 other products** that would meet the user's preferences as the **7 correctly presented** products



$$recall = \frac{TP}{TP + FN}$$

Recall – Example

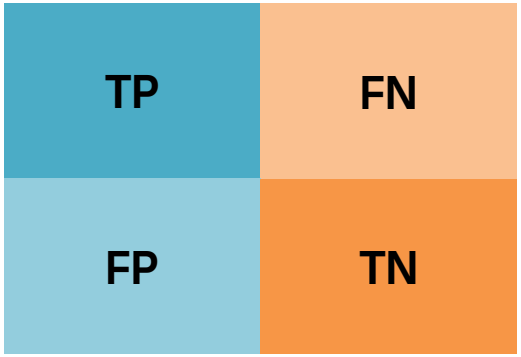
- System did not recognize **5 other products** that would meet the user's preferences as the **7 correctly presented** products



$$recall = \frac{TP}{TP + FN}$$

Recall – Example

- System did not recognize **5 other products** that would meet the user's preferences as the **7 correctly presented** products



$$recall = \frac{TP}{TP + FN} = \frac{7}{7 + 5} = 0.583$$

Correlation between precision & recall

- **Increasing number** of recommended items
  - →Probably **higher recall**
  - →But also **lower precision**

- **Prec@N**
  - Predictions not interesting for all items
  - Users do not view all recommendations
  - →If number of recommendations is fixed: Prec@N (Prec@5, Prec@10, …)

F_1 score

- **F_1 score** determines **trade-off**
  - Harmonic mean of **precision** and **recall**
- Suitable for **comparison**

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

F_1 score – Example

- $precision = 0.7$
- $recall = 0.583$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

F_1 score – Example

- $precision = 0.7$
- $recall = 0.583$
- $F_1 = \frac{2 \cdot 0.7 \cdot 0.583}{0.7+0.583}$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

F_1 score – Example

- $precision = 0.7$
- $recall = 0.583$
- $F_1 = \frac{2 \cdot 0.7 \cdot 0.583}{0.7+0.583} = 0.636$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Precision-recall curves

- Show **split** in rather good and rather bad results by **F_1 score** (for (r, p)-pairs)
- Good for **comparing** lists of different lengths