# Оценка качества рекомендательных систем

Июнь 2017

# Задачи рекомендательной системы

- предсказать рейтинг конкретного товара
- ранжировать товары по привлекательности для пользователя

# Как измерять?

- опросы пользователей
- offline ( на отложенных данных)
- online (A/B-testing)

# Как проводить эксперименты?

- кросс-валидация:

  данные разбиваются случайным образом на K частей. Обучение проходит на K-1 части и валидируется на оставшейся. И так K раз.

- Отсечка по времени.

# Оценка точности предсказания

# Оценка точности предсказания

$$MAE: \quad \frac{1}{|K|} \sum_{\{i,j\} \in K} |\hat{r}_{ij} - r_{ij}|,$$

$$MSE: \quad \frac{1}{|K|} \sum_{\{i,j\} \in K} (\hat{r}_{ij} - r_{ij})^2,$$

$$RMSE: \quad \sqrt{\frac{1}{|K|} \sum_{\{i,j\} \in K} (\hat{r}_{ij} - r_{ij})^2}$$

# Специфика **MAE** и **RMSE**

**CASE 1: Evenly distributed errors**

| ID | Error | \|Error\| | Error^2 |
|---|---|---|---|
| 1 | 2 | 2 | 4 |
| 2 | 2 | 2 | 4 |
| 3 | 2 | 2 | 4 |
| 4 | 2 | 2 | 4 |
| 5 | 2 | 2 | 4 |
| 6 | 2 | 2 | 4 |
| 7 | 2 | 2 | 4 |
| 8 | 2 | 2 | 4 |
| 9 | 2 | 2 | 4 |
| 10 | 2 | 2 | 4 |

| MAE | RMSE |
|---|---|
| 2.000 | 2.000 |

**CASE 2: Small variance in errors**

| ID | Error | \|Error\| | Error^2 |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 |
| 6 | 3 | 3 | 9 |
| 7 | 3 | 3 | 9 |
| 8 | 3 | 3 | 9 |
| 9 | 3 | 3 | 9 |
| 10 | 3 | 3 | 9 |

| MAE | RMSE |
|---|---|
| 2.000 | 2.236 |

**CASE 3: Large error outlier**

| ID | Error | \|Error\| | Error^2 |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 |
| 10 | 20 | 20 | 400 |

| MAE | RMSE |
|---|---|
| 2.000 | 6.325 |

https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

# Специфика MAE и RMSE

**CASE 4: Errors = 0 or 5**

| ID | Error | \|Error\| | Error^2 |
|---|---|---|---|
| 1 | 5 | 5 | 25 |
| 2 | 0 | 0 | 0 |
| 3 | 5 | 5 | 25 |
| 4 | 0 | 0 | 0 |
| 5 | 5 | 5 | 25 |
| 6 | 0 | 0 | 0 |
| 7 | 5 | 5 | 25 |
| 8 | 0 | 0 | 0 |
| 9 | 5 | 5 | 25 |
| 10 | 0 | 0 | 0 |

| var | MAE | RMSE |
|---|---|---|
| 6.944 | 2.500 | 3.536 |

**CASE 5: Errors = 3 or 4**

| ID | Error | \|Error\| | Error^2 |
|---|---|---|---|
| 1 | 3 | 3 | 9 |
| 2 | 4 | 4 | 16 |
| 3 | 3 | 3 | 9 |
| 4 | 4 | 4 | 16 |
| 5 | 3 | 3 | 9 |
| 6 | 4 | 4 | 16 |
| 7 | 3 | 3 | 9 |
| 8 | 4 | 4 | 16 |
| 9 | 3 | 3 | 9 |
| 10 | 4 | 4 | 16 |

| var | MAE | RMSE |
|---|---|---|
| 0.278 | 3.500 | 3.536 |

https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

# Связь **MAE** и **RMSE**

$$MAE \leq RMSE \leq MAE \cdot \sqrt{|K|}$$

https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

# decision-support

метрики

# Precision and Recall

| | **Recommended** | **Not Recommended** | |
|---|---|---|---|
| **Preferred** | True-Positive (tp) | False-Negative (fn) | #tp+#fn |
| **Not Preferred** | False-Positive (fp) | True-Negative (tn) | #fp+#tn |
| | #tp+#fp | | |

https://en.wikipedia.org/wiki/Precision_and_recall

# Precision

$$Precision: P = \frac{\#tp}{\#tp + \#fp}$$

Количество  угаданных товаров из всех рекомендованных.

# RECALL (True Positive Rate)

$$Recall: R = \frac{\#tp}{\#tp + \#fn}$$

Количество угаданных товаров из всех “хороших” для пользователя

(из всех, что мы должны были угадать)

# FPR (False Positive Rate)

$$FalsePositiveRate = \frac{\#fp}{\#fp + \#tn}$$

Ошибка первого рода - вероятность отвергнуть правильную гипотезу.

# 4 алгоритма

| A | Rec | NotRec | |
|---|---|---|---|
| **good** | 63 | 37 | TPR = 0.63 |
| **bad** | 28 | 72 | FPR = 0.28 |

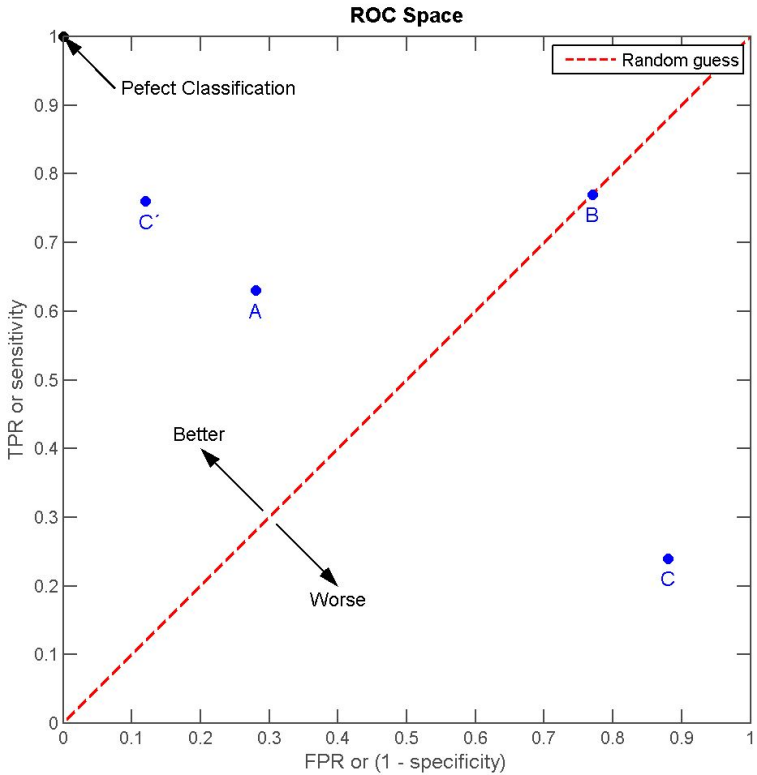
| B | Rec | NotRec | |
|---|---|---|---|
| **good** | 77 | 23 | TPR = 0.77 |
| **bad** | 77 | 23 | FPR = 0.77 |

| C | Rec | NotRec | |
|---|---|---|---|
| **good** | 24 | 76 | TPR = 0.24 |
| **bad** | 88 | 12 | FPR = 0.88 |

| C’ | Rec | NotRec | |
|---|---|---|---|
| **good** | 76 | 24 | TPR = 0.76 |
| **bad** | 12 | 88 | FPR = 0.12 |

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

# 4 алгоритма




https://en.wikipedia.org/wiki/Receiver_operating_characteristic

# В реальном мире

**Precision -> Precision@K**

**Recall -> Recall@K**

**+ усреднение по всем пользователям**

# Ранговые метрики

# Ранговые метрики

- MRR (Mean Reciprocal Rank)
- Доля правильно уопрядоченных пар
- SRC (Spearman’s Rank Correlation)
- DCG (Discount Cumulative Gain)

# MEAN RECIPROCAL RANK

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Q - набор запросов

rank_i - ранк первого релевантного объекта в i-ом запросе.

плохо - не берет в учет остальные объекты

# Доля правильно упорядоченных пар

$$\frac{TrueOrder}{\frac{n(n-1)}{2}}$$

# Spearman's Rank Correlation

$$SPR = \frac{\sum_i (r_1(i) - \overline{r})(r_2(i) - \overline{r})}{\sqrt{\sum_i (r_1(i) - \overline{r})^2} \sqrt{\sum_i (r_2(i) - \overline{r})^2}}$$

плохо - не учитывают позицию элемента в списке

# Discounted Cumulative Rate

$$DCG(R) = \sum u(i)d(i)$$

$$nDCG(R) = \frac{DCG(R)}{DCG(trueR)}$$

# Discounted Cumulative Rate

$$DCG(R) = \sum u(i)d(i)$$

$$d(i) = \frac{1}{\max(1, \log_2 r(i))}$$

$$d(i) = \frac{1}{2^{\frac{r(i)-1}{\alpha-1}}}$$

# “Бизнес-метрики”

# Бизнес - метрики

- coverage
- diversity
- serendipity

# Coverage

- % товаров, которые рекомендуются пользователям
- % товаров, для которых RecSys умеет делать предсказания

# Diversity

Разнообразие рекомендаций:

- удалять из топа слишком похожие
- всегда отсавлять первую рекомендацию
- заменить похожие на (n+1) рекомендацию
- кластеризовать выдачу

# Serendipity

- Штрафовать популярные
- “Заставить” потреблять менее популярные

Спасибо за внимание.