

Метрики в задачах ранжирования и матчинга.

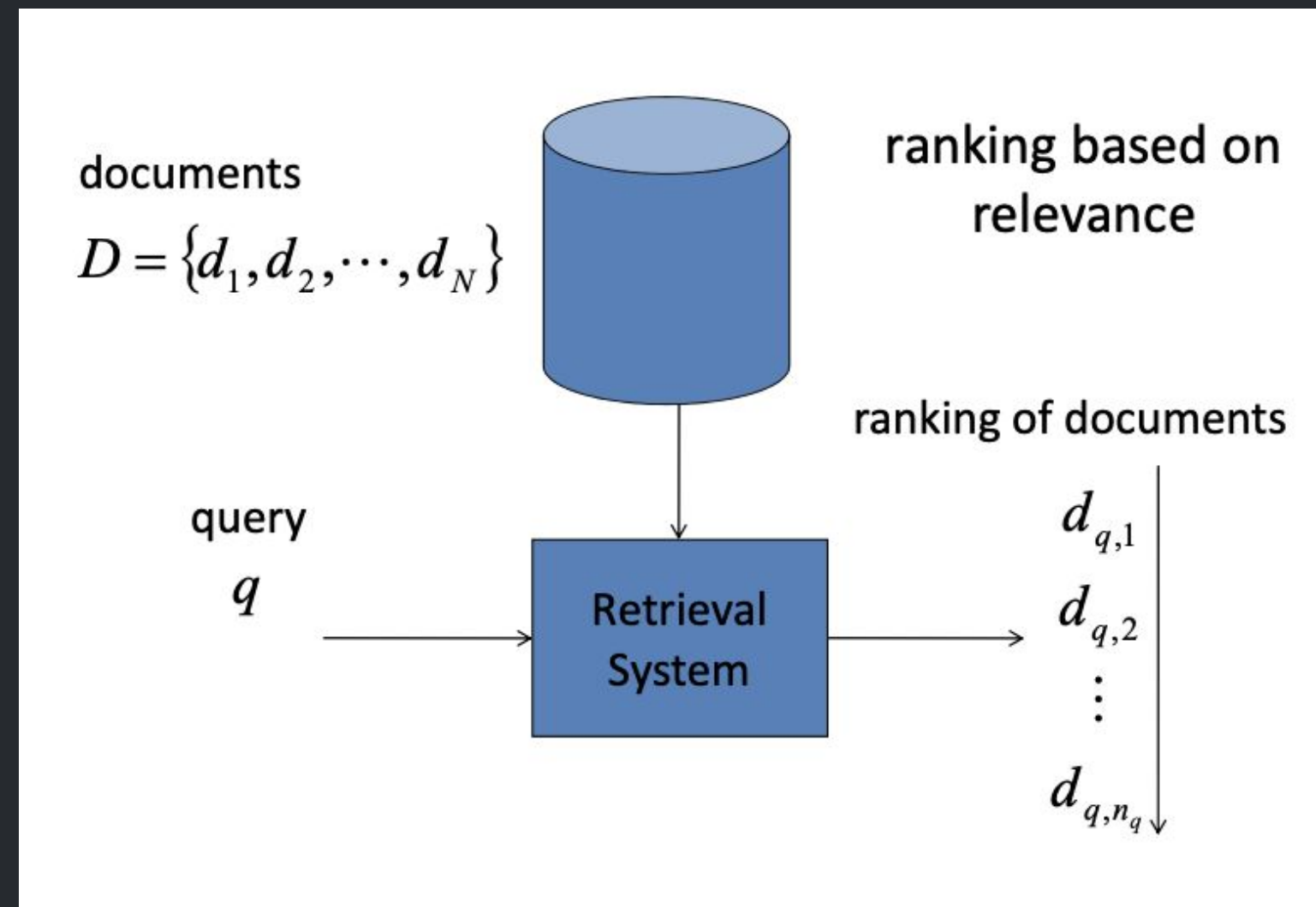
KARPOV.COURSES

План лекции

- метрики качества с точки зрения классификации
- особенности метрик качества для матчинга
- метрики качества с точки зрения ранжирования

Матчинг с точки зрения ML

Learning to **rank** (LTR) - ранжирование



Что измерять в ранжировании?

- Качество / Точность – насколько аккуратна система ранжирования?
 - Измеряем возможности системы ранжировать релевантные документы выше нерелевантных
- Эффективность – насколько быстро выдается ответ? Сколько ресурсов нужно для формирования ответа?
 - Измеряем затраты на память и время формирования ответа
- Удобство использования – насколько полезна система для решения задач?
 - Пользовательские ощущения, UX

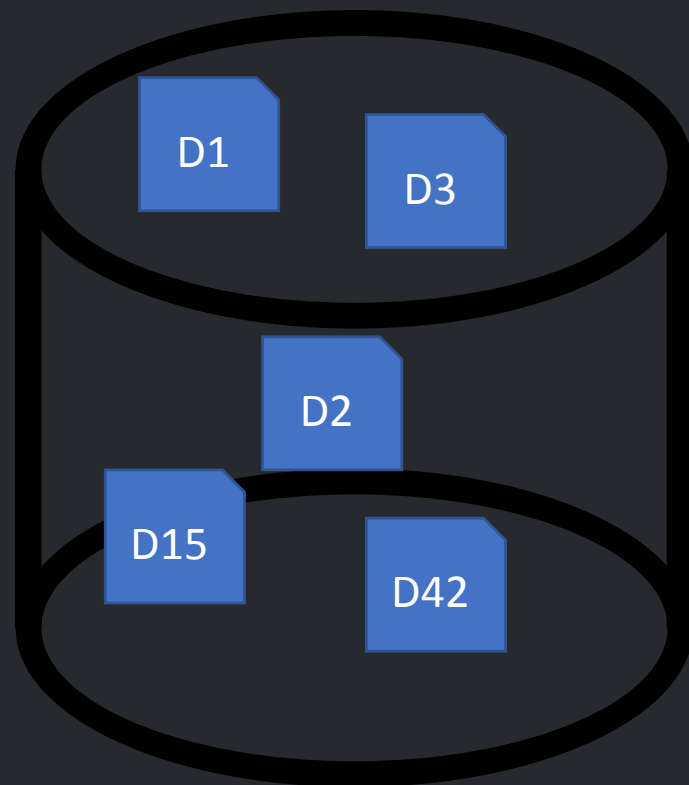
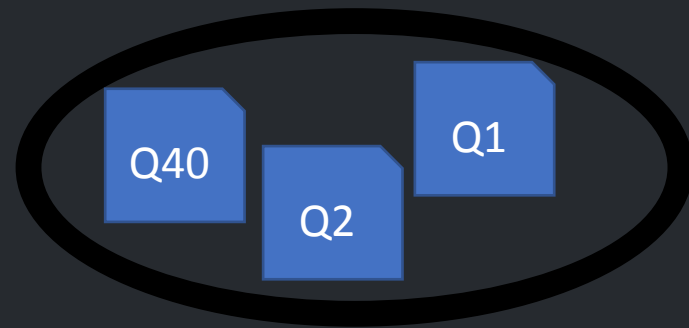
Оценка качества ранжирования

Методология оценки Кранфилда (Cranfield Evaluation Methodology):

- Зафиксированный набор документов
- Зафиксированный набор запросов
- Оценки релевантности пар (в идеале оценки даются пользователями системы)
- Наборы должны быть **репрезентативными**

Оценка качества ранжирования

Запросы



Документы

Оценки

релевантности

Q1 D2 +

Q1 D3 -

Q1 D4 -

...

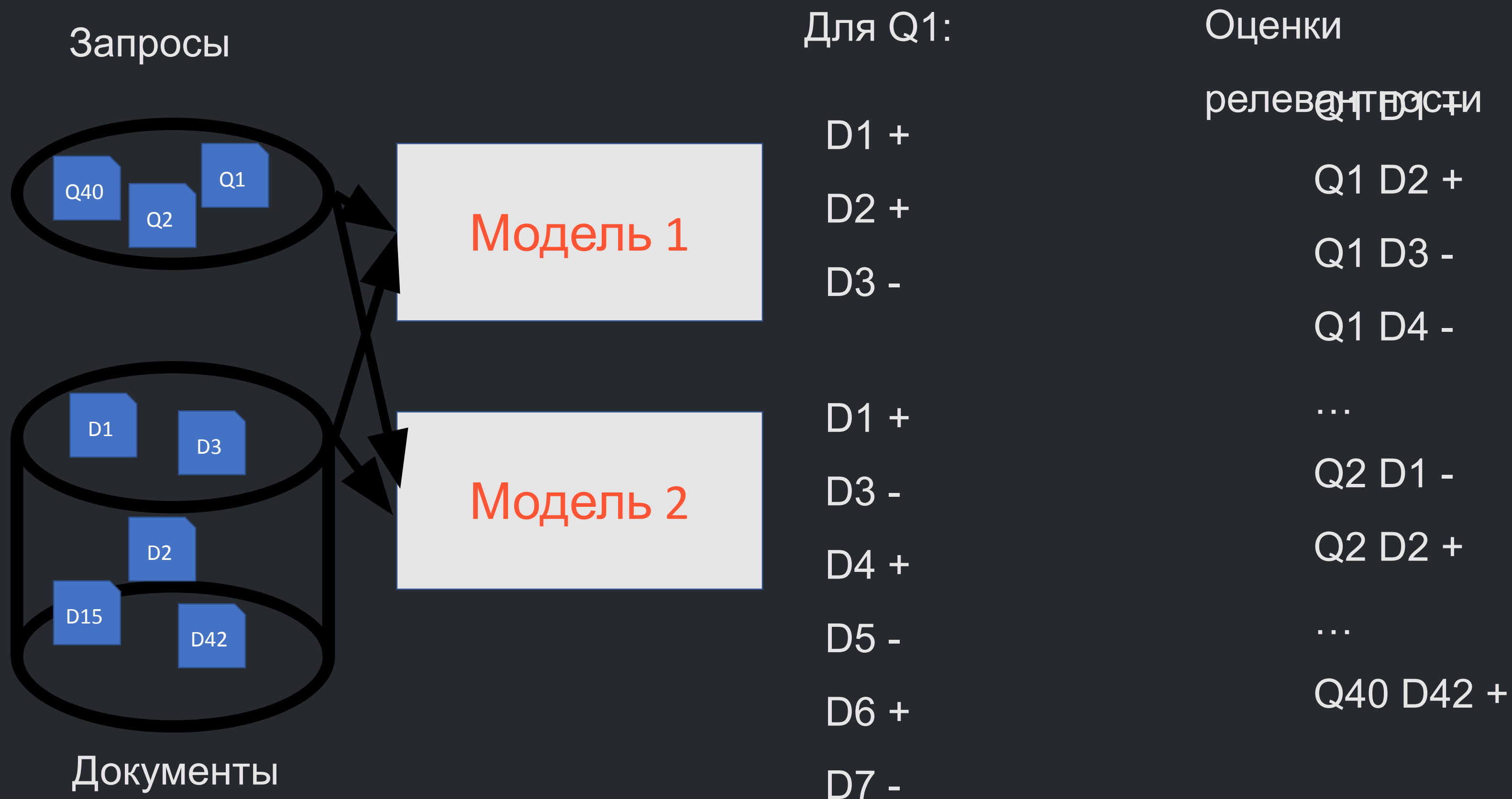
Q2 D1 -

Q2 D2 +

...

Q40 D42 +

Оценка качества ранжирования



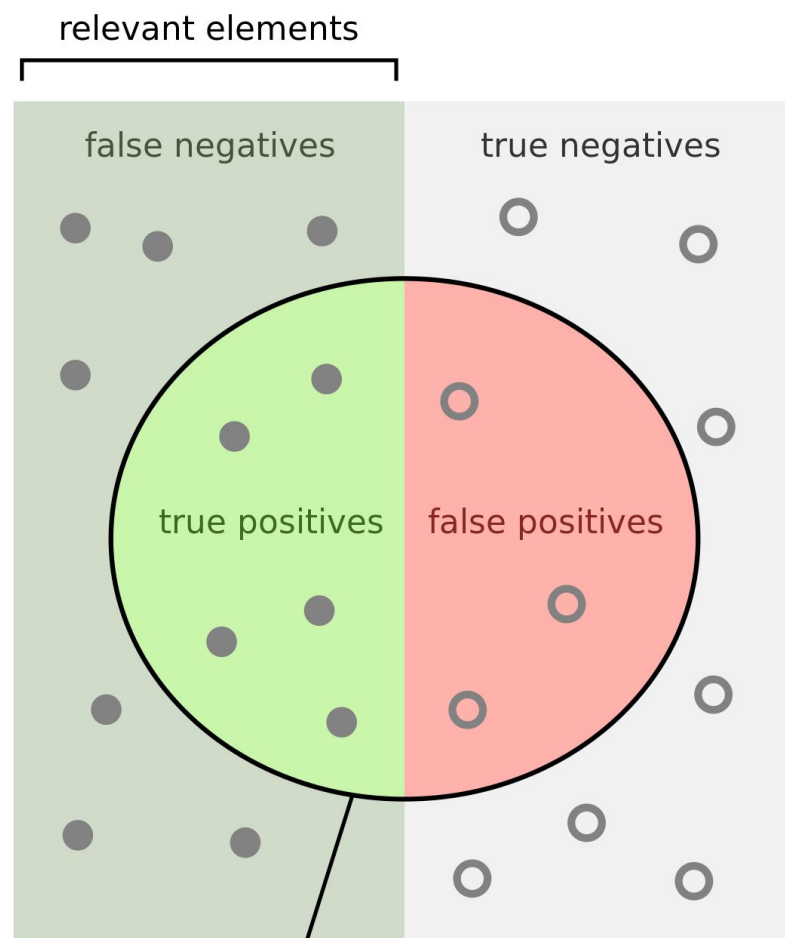
Оценка качества ранжирования



Оценка качества ранжирования

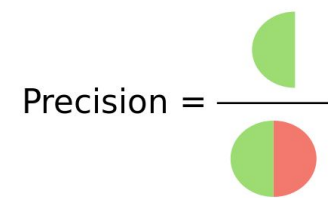


Оценка качества ранжирования



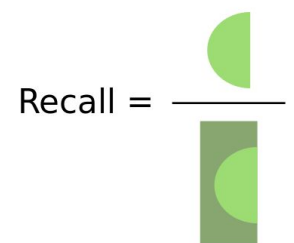
selected elements

How many selected items are relevant?



Precision =

How many relevant items are selected?



Recall =

- Доля правильных ответов (accuracy)
- Точность, полнота (Precision, Recall)

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Ограничение на расчет в Top-K (@K),
Precision@5

Оценка качества ранжирования

- F1, F_β-меры

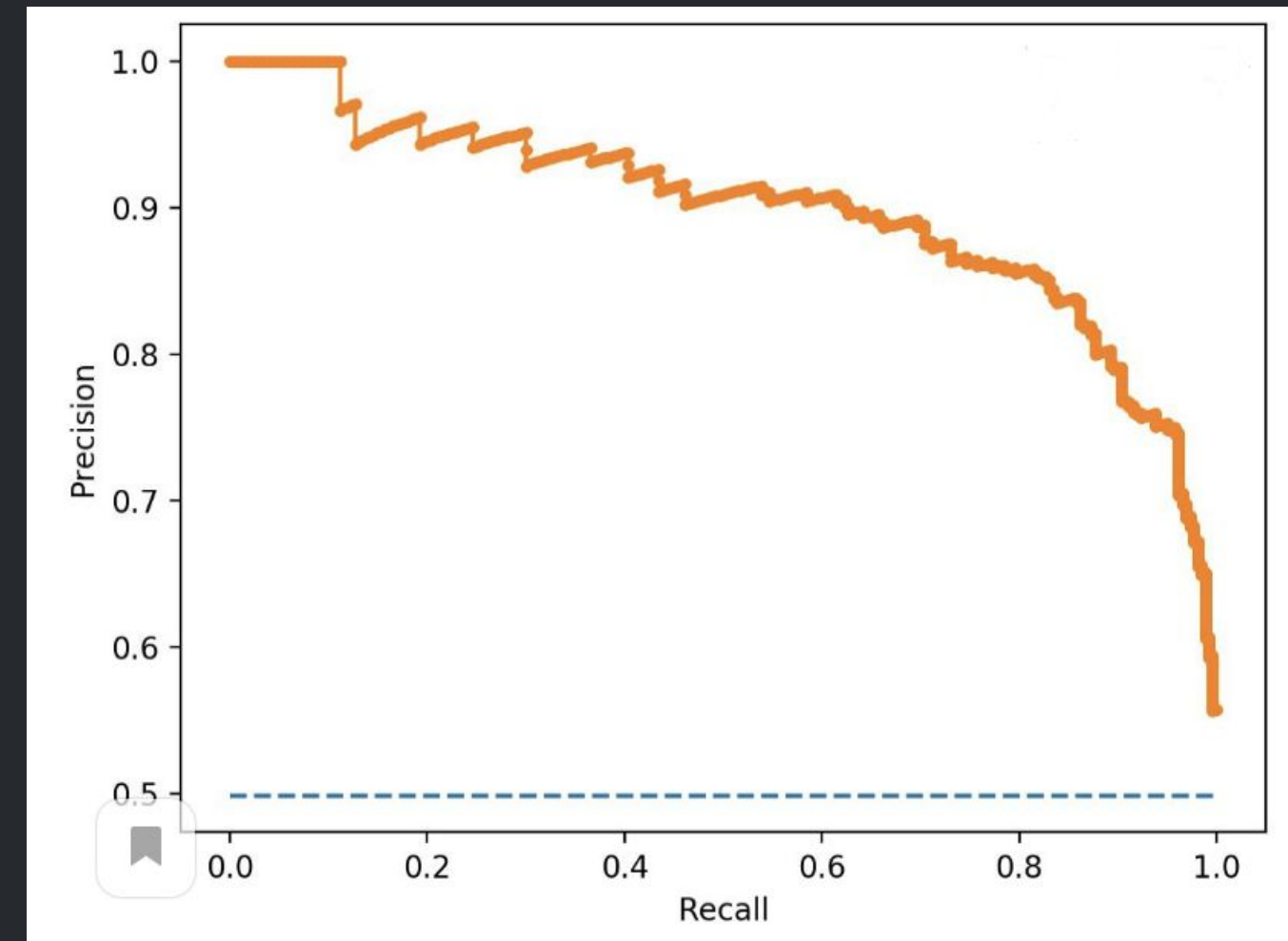
$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Почему просто не брать $0.5 * P + 0.5 * R$?

Оценка качества ранжирования

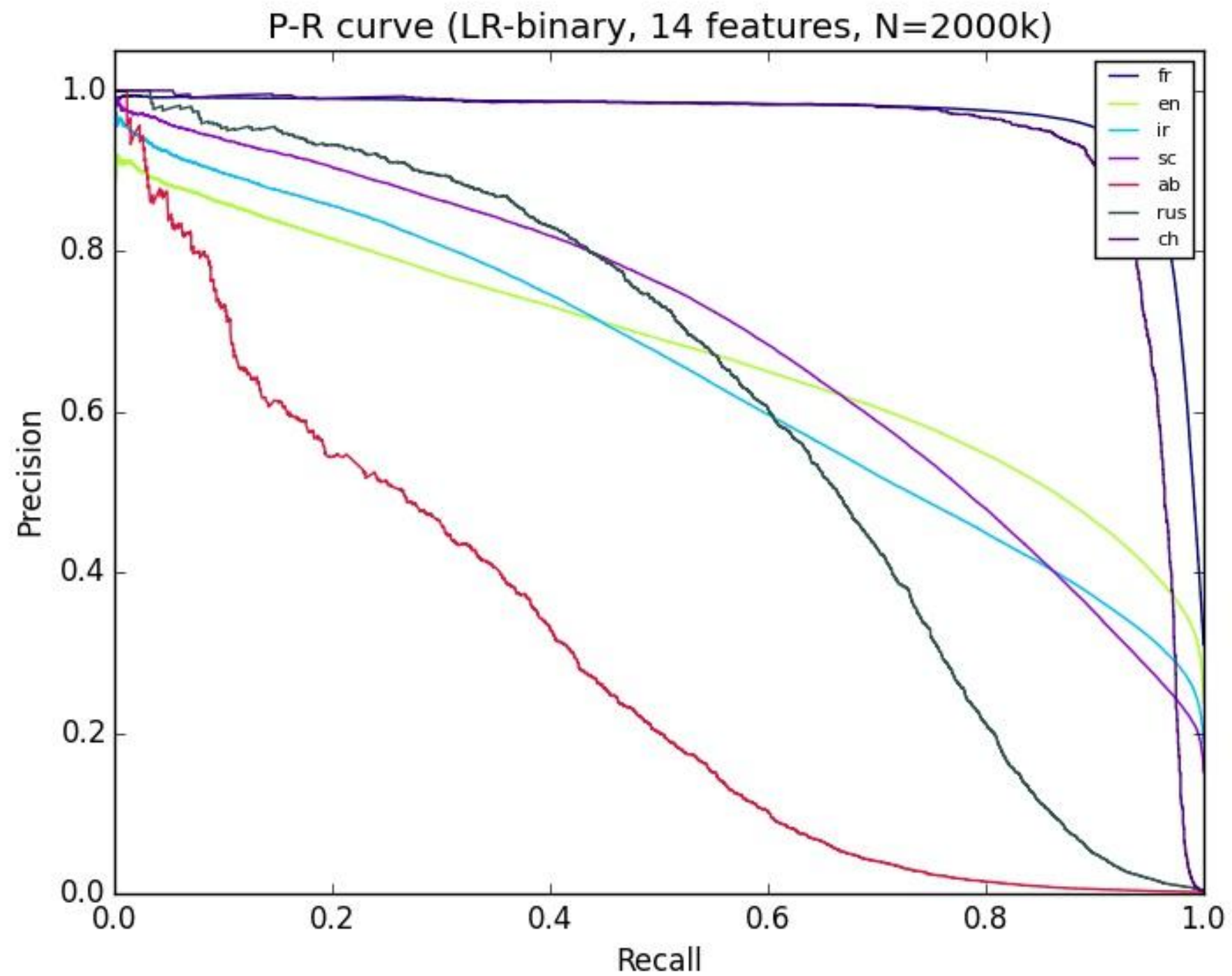
- PR-auc

ID оффера	ID модели	Предсказание формулы	Правильный ответ
a01	1	6.4	1
a01	3	0.7	0
b02	2	0.6	1
c03	2	-0.8	0



- Сортируем предсказания по убыванию релевантности
- Считаем значение точности и полноты по первой паре
- Понижаем значение порога, чтобы выше порога было две пары
- Повторяем до тех пор, пока не добавим все элементы

Оценка качества ранжирования



- PR-auc
- PR-auc @N

Оценка качества ранжирования

Average Precision (AP) – насколько много релевантных объектов сконцентрировано среди самых высокооцененных

$$AP = \sum_K (Recall@k - Recall@k-1) \cdot Precision@k$$

Оценка качества ранжирования

Average Precision (AP) – насколько много релевантных объектов сконцентрировано среди самых высокооцененных

$$AP = \sum_K (Recall@k - Recall@k-1) \cdot Precision@k$$

k	Document ID	Predicted Relevance	Actual Relevance
1	06	0.90	Relevant (1.0)
2	03	0.85	Not Relevant (0.0)
3	05	0.71	Relevant (1.0)
4	00	0.63	Relevant (1.0)
5	04	0.47	Not Relevant (0.0)
6	02	0.36	Relevant (1.0)
7	01	0.24	Not Relevant (0.0)
8	07	0.16	Not Relevant (0.0)

Всего релевантных
нашли

1

1

2

3

3

4

4

4

Оценка качества ранжирования

Average Precision (AP) – насколько много релевантных объектов сконцентрировано среди самых высокооцененных

$$AP = \sum_K (Recall@k - Recall@k-1) \cdot Precision@k$$

(Кол-во корректных предсказаний)

k	Document ID	Predicted Relevance	Actual Relevance
1	06	0.90	Relevant (1.0)
2	03	0.85	Not Relevant (0.0)
3	05	0.71	Relevant (1.0)
4	00	0.63	Relevant (1.0)
5	04	0.47	Not Relevant (0.0)
6	02	0.36	Relevant (1.0)
7	01	0.24	Not Relevant (0.0)
8	07	0.16	Not Relevant (0.0)

Всего релевантных
нашли

1

1

2

3

3

4

4

4

Скользкая
сумма

$$0 + 1/1 = 1$$

1

$$1 + 2/3 = 1.67$$

$$1.67 + 3/4 = 2.42$$

2.42

$$2.42 + 4/6 = 3.08$$

3.08

3.08

$$3.08 / 4 = 0.77$$

Оценка качества ранжирования

Average Precision (AP) – насколько много релевантных объектов сконцентрировано среди самых высокооцененных

$$AP = \sum_K (Recall@k - Recall@k-1) \cdot Precision@k$$

$$MAP = \frac{\sum_{q=1}^Q AveP(q)}{Q}$$

Оценка качества ранжирования

Переход от бинарной задачи релевантно/не релевантно к многоуровневой

Уровень релевантности:

1. Не релевантно
2. В целом релевантно
3. Очень релевантно, точное соответствие

Оценка качества ранжирования

Переход от бинарной задачи релевантно/не релевантно к многоуровневой

Уровень релевантности:

1. Не релевантно
2. В целом релевантно
3. Очень релевантно, точное соответствие

D1 3

D2 2

D3 1

D4 1

D5 3

D6 1

D7 2

Оценка качества ранжирования

Переход от бинарной задачи релевантно/не релевантно к многоуровневой

Уровень релевантности:

1. Не релевантно
2. В целом релевантно
3. Очень релевантно, точное соответствие

	Gain	Cumulative Gain
D1	3	3
D2	2	3+2
D3	1	3+2+1
D4	1	3+2+1+1
D5	3	3+2+1+1+3
D6	1	3+2+1+1+3+1
D7	2	3+2+1+1+3+1+2

Оценка качества ранжирования

Переход от бинарной задачи релевантно/не релевантно к многоуровневой

Уровень релевантности:

1. Не релевантно
2. В целом релевантно
3. Очень релевантно, точное соответствие

$$\log_2(k+1)$$

	Gain	Cumulative Gain	Discounted Cumulative Gain
D1	3	3	3
D2	2	3+2	3 + 2/log(3)
D3	1	3+2+1	3 + 2/log(3) + 1/log(4)
D4	1	3+2+1+1	3 + 2/log(3) + 1/log(4) + 1/log(5)
D5	3	3+2+1+1+3	...
D6	1	3+2+1+1+3+1	
D7	2	3+2+1+1+3+1+2	DCG@7 = 3 + 2/log(3) + ... + 2/log(8)

Оценка качества ранжирования

Переход от бинарной задачи релевантно/не релевантно к многоуровневой

Уровень релевантности:

1. Не релевантно
2. В целом релевантно
3. Очень релевантно, точное соответствие

	Gain	Discounted Cumulative Gain
D1	3	3
D2	2	$3 + 2/\log(3)$
D3	1	$3 + 2/\log(3) + 1/\log(4)$
D4	1	$3 + 2/\log(3) + 1/\log(4) + 1/\log(5)$
D5	3	...
D6	1	
D7	2	$DCG@7 = 3 + 2/\log(3) + \dots + 2/\log(8) \sim 7.38$

Оценка качества ранжирования

Переход от бинарной задачи релевантно/не релевантно к многоуровневой

Уровень релевантности:

1. Не релевантно
2. В целом релевантно
3. Очень релевантно, точное соответствие

	Gain	Discounted Cumulative Gain
D1	3	3
D2	2	$3 + 2/\log(3)$
D3	1	$3 + 2/\log(3) + 1/\log(4)$
D4	1	$3 + 2/\log(3) + 1/\log(4) + 1/\log(5)$
D5	3	...
D6	1	
D7	2	$DCG@7 = 3 + 2/\log(3) + \dots + 2/\log(8) \sim 7.38$
		$IdealDCG@7 = 3 + 3/\log(3) + \dots + 1/\log(8) \sim 7.83$

Оценка качества ранжирования

Переход от бинарной задачи релевантно/не релевантно к многоуровневой

Уровень релевантности:

1. Не релевантно
2. В целом релевантно
3. Очень релевантно, точное соответствие

	Gain	Discounted Cumulative Gain
D1	3	3
D2	2	$3 + 2/\log(3)$
D3	1	$3 + 2/\log(3) + 1/\log(4)$
D4	1	$3 + 2/\log(3) + 1/\log(4) + 1/\log(5)$
D5	3	...
D6	1	
D7	2	

$$\text{DCG}@7 = 3 + 2/\log(3) + \dots + 2/\log(8) \sim 7.38$$
$$\text{IdealDCG}@7 = 3 + 3/\log(3) + \dots + 1/\log(8) \sim 7.83$$
$$\frac{7.38}{7.83} = 0.942$$

$$\text{nDCG}@K = \frac{\text{DCG}@K}{\text{IdealDCG}@K}$$

Оценка качества ранжирования

PFound (Yandex):

Значение метрики будет **оценкой вероятности** найти релевантный результат в выдаче модели

$$pfound = \sum_{i=1}^n pLook[i] * pRel[i]$$

$pLook[i]$ – вероятность просмотреть i -й документ из списка

$pRel[i]$ – вероятность того, что i -й документ окажется релевантным (например, 0%, 50%, 100% для шкалы с тремя уровнями)

Оценка качества ранжирования

PFound (Yandex):

Значение метрики будет **оценкой вероятности** найти релевантный результат в выдаче модели

$$pfound = \sum_{i=1}^n pLook[i] * pRel[i]$$

$pLook[i]$ – вероятность просмотреть i -й документ из списка

$pRel[i]$ – вероятность того, что i -й документ окажется релевантным (например, 0%, 50%, 100% для шкалы с тремя уровнями)

Для расчета $pLook[i]$ используется два предположения:

- результаты ранжирования отсматриваются сверху вниз
- процесс прекращается в случае нахождения релевантного результата либо без каких-то определенных причин («надоело»)

Оценка качества ранжирования

$$p_{found} = \sum_{i=1}^n p_{Look}[i] * p_{Rel}[i]$$

$p_{Look}[i]$ – вероятность просмотреть i -й документ из списка

$p_{Rel}[i]$ – вероятность того, что i -й документ окажется релевантным (например, 0%, 50%, 100% для шкалы с тремя уровнями)

Для расчета $p_{Look}[i]$ используется два предположения:

- результаты ранжирования отсматриваются сверху вниз
- процесс прекращается в случае нахождения релевантного результата либо без каких-то определенных причин («надоело»)

$$p_{Look}[i] = p_{Look}[i - 1] * (1 - p_{Rel}[i - 1]) * (1 - p_{Break})$$

Исторические метрики

Среднеобратный ранг (Mean reciprocal rank, MRR)

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

Среднее гармоническое между рангами

Запрос	Ответы	Правильный ответ	Ранг	Обратный ранг
кочерга	кочерг, кочергей, кочерёг	кочерёг	3	1/3
попадья	попадь, попадей , попадьеёв	попадей	2	1/2
турок	турок , турков, турчан	турок	1	1

$$(1/3 + 1/2 + 1) / 3 = 11/18 \sim 0.61$$

Исторические метрики

Kendall rank correlation coefficient (Kendall's τ)

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}$$

$$\binom{n}{2} = \frac{n(n-1)}{2} \quad \text{- Биномиальный коэффициент}$$

Часто используется в статистике для оценки **ранговых корреляций**

Матчинг и ранжирование

- Имеем привилегию отказаться от выдачи
- Важны только самые-самые первые результаты (1-3)
- Огромный дисбаланс (от нуля до тысяч матчей)
- Финальное решение можно предоставить классификатору
- Отдельные метрики для разных этапов пайплайна
- Метрики могут агрегироваться на уровне одного SKU
- Различие прокси-метрик и бизнес-метрик