ЗАДАЧА РАНЖИРОВАНИЯ

РАНЖИРОВАНИЕ

МФТИ

- Дано:
  - Объекты $x_1, \ldots, x_\ell$
  - Порядок на некоторых парах: $\{(i, j) : x_i < x_j\}$
- Найти:
  - Ранжирующую модель $a(x)$, такую что

$$x_i < x_j \Rightarrow a(x_i) < a(x_j)$$

РАНЖИРОВАНИЕ ПОИСКОВОЙ ВЫДАЧИ

МФТИ

- Объекты — пары (запрос, документ)
- Порядок задан только внутри одного запроса:
  $(\text{запрос}, \text{документ}_1) < (\text{запрос}, \text{документ}_2)$
- Как правило, порядок задают асессоры

РЕКОМЕНДАЦИИ

МФТИ

- Объекты — пары (пользователь, товар)
- Порядок задан только для пар одного и того же пользователя
- Порядок определяется оценками или покупками пользователя

ОСОБЕННОСТИ ЗАДАЧИ

МФТИ

- Объекты не являются независимыми — целевая переменная зависит от пар объектов
- Сложные метрики качества
- Много способов сформировать выборку

МЕТРИКИ КАЧЕСТВА РАНЖИРОВАНИЯ

РАНЖИРОВАНИЕ

МФТИ

- Дано:
  - Объекты $x_1, \ldots, x_\ell$
  - Порядок на некоторых парах: $\{(i, j) : x_i < x_j\}$
- Найти:
  - Ранжирующую модель $a(x)$, такую что

$$x_i < x_j \Rightarrow a(x_i) < a(x_j)$$

ТОЧНОСТЬ РАНЖИРОВАНИЯ

МФТИ

- $y(q, d) \in \{0, 1\}$ — релевантность документа d запросу q
- $a(q, d)$ — оценка релевантности
- $d_q^{(i)}$ — i-й документ в порядке оценки релевантности
- Точность среди первых k документов:

$$\text{Precision@}k(q) = \frac{1}{k} \sum_{i=1}^{k} y\left(q, d_q^{(i)}\right)$$

ТОЧНОСТЬ РАНЖИРОВАНИЯ

МФТИ

- Точность среди первых k документов:

$$\text{Precision@}k(q) = \frac{1}{k}\sum_{i=1}^{k} y\left(q, d_q^{(i)}\right)$$

- Не учитывает позиции релевантных документов

ТОЧНОСТЬ РАНЖИРОВАНИЯ

МФТИ

- Средняя точность (average precision, AP):

$$\text{AP@}k(q) = \frac{\sum_{i=1}^{k} y\left(q, d_q^{(i)}\right) \text{Precision@}i(q)}{\sum_{i=1}^{k} y\left(q, d_q^{(i)}\right)}$$

- Достигает максимума, если все релевантные документы находятся выше всех нерелевантных

ТОЧНОСТЬ РАНЖИРОВАНИЯ

МФТИ

- Средняя AP по запросам:

$$\text{MAP@}k = \frac{1}{|Q|} \sum_{q \in Q} \text{AP@}k(q)$$

DCG

МФТИ

- $y(q, d) \in \mathbb{R}$ — релевантность документа d запросу q
- $a(q, d)$ — оценка релевантности
- $d_q^{(i)}$ — i-й документ в порядке оценки релевантности
- Discounted cumulative gain (DCG):

$$\text{DCG@}k(q) = \sum_{i=1}^{k} \frac{2^{y\left(q, d_q^{(i)}\right)} - 1}{\log(i+1)}$$

DCG

МФТИ

- Метрику DCG принято нормировать:

$$n\text{DCG@}k(q) = \frac{\text{DCG@}k(q)}{\textbf{max}\,\text{DCG@}k(q)}$$

- **max** DCG@$k(q)$ — DCG при идеальном ранжировании

РЕЗЮМЕ

МФТИ

- Два подхода к оцениванию качества — точность и DCG
- Точность учитывает долю релевантных документов в топе
- Модификация: средняя точность
- DCG учитывает релевантность документа и его позицию

МЕТОДЫ РАНЖИРОВАНИЯ

ПОДХОДЫ К РАНЖИРОВАНИЮ

МФТИ

- Pointwise
- Pairwise
- Listwise

ПОТОЧЕЧЕНЫЙ ПОДХОД

МФТИ

- $y(q, d) \in \mathbb{R}$ — релевантность документа d запросу q
- $a(q, d)$ — оценка релевантности
- Будем предсказывать $y(q, d)$ методами регрессии
- Например:

$$\sum_{i=1}^{\ell} (a(q, d) - y(q, d))^2 \to \textbf{min}$$

ПОПАРНЫЙ ПОДХОД

МФТИ

- Минимизируем количество дефектных пар:

$$\sum_{x_i < x_j} [a(x_j) - a(x_i) < 0] \to \textbf{min}$$

ПОПАРНЫЙ ПОДХОД

МФТИ

- Минимизируем количество дефектных пар:

$$\sum_{x_i < x_j} L(a(x_j) - a(x_i)) \to \textbf{min}$$

- $L(M)$ — гладкая функция
- $L(M) = \log(1 + e^{-M})$ — метод RankNet

LISTWISE-ПОДХОД

МФТИ

- RankNet, шаг стохастического градиентного спуска для линейной модели:

$$w := w + \eta \frac{1}{1 + \exp(\langle w, x_j - x_i \rangle)} (x_j - x_i)$$

- Как оптимизировать NDCG вместо $L(M)$?

LISTWISE-ПОДХОД

МФТИ

- Домножим стохастический градиент по паре (x_i, x_j) на изменение NDCG при перестановке x_i и x_j местами:

$$w := w + \eta \frac{1}{1 + \exp(\langle w, x_j - x_i \rangle)} (x_j - x_i) \cdot$$
$$\cdot |\Delta \text{NDCG}_{ij}| (x_j - x_i)$$

- Эмпирическое наблюдение: такая модификация действительно приводит к оптимизации NDCG

РЕЗЮМЕ

МФТИ

- Три подхода: pointwise, pairwise, listwise
- Наиболее часто используется попарный подход