Задача ранжирования
(learning to rank)

Раньше: $\mathbb{Y} = \mathbb{R}$

$\mathbb{Y} \in \{1, \dots K\}$

$X = \{x_1, \dots x_\ell\}$

$R \subset \{1, \dots \ell\} \times \{1, \dots \ell\}$

$(i, j) \in R \Rightarrow a(x_i) > a(x_j)$

например: $x = (q, d)$

$((q, d_1), (q, d_2)) \in R \Rightarrow a(q, d_1) > a(q, d_2)$

$q: \; d_1 > d_2 > \dots > d_n$

---

① Метрики качества ранжирования

Будем считать метрики для каждого запроса q, а затем усреднять по всем запросам

q:	d_1	1		0	
	d_2	0		0	
	d_3	0	$\frac{2}{5}$	0	$\frac{2}{5}$
	d_4	1		1	
	d_5	0		1	

1) $\text{precision@K}(q) = \frac{1}{K} \sum_{i=1}^{K} [y(i) = 1]$

$\frac{1}{K} \sum_{i=1}^{K} \frac{1}{i} [y(i) = 1]$

2) $AP@k(q) = \sum_{i=1}^{k} \left(\frac{(y_{(i)})}{\sum_{j=1}^{k}(y_{(j)})}\right) \underline{precision@i(q)}$

average precision

1	$1 \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{2}\left(\frac{5}{3}\right) = \frac{5}{6}$
0	
1	
0	
0	

0	$\frac{1}{4} \cdot \frac{1}{2} + \frac{2}{5} \cdot \frac{1}{2} < \frac{5}{6}$
0	
0	
1	
1	

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

---

3) $DCG@k(q) = \sum_{i=1}^{k} \underline{g(y_{(i)})}\ \underline{d(i)} \to \max$

discounted cumulative gain

$g(y) = 2^{y} - 1$ $\quad y$

$d(i) = \boxed{\frac{1}{\log(i+1)}}$

$$nDCG@k(q) = \frac{DCG@k(q)}{\max DCG@k(q)}$$

1	2
0	1
2	0
0	0
0	0

DCG, если мы идеально отранжируем заданные к документов

---

4) pFound

$y_{(i)} \in [0, 1]$ - вероятность того, что польз. найдет ответ в документе

p_i - вер-ть, что польз. дойдет до i-й позиции

$p_1 = 1$

$$p_{i+1} = p_i\,(1 - y(i))\,(1 - p_{out})$$

$$pFound@k(q) = \sum_{i=1}^{k} p_i\, y(i)$$

② Подходы к ранжированию

1) pointwise

$$x_i \to y_i$$

оценки релевантности

$$a(x_i) \approx y_i$$

	m_1	m_2	m_3
y_i	5	4	1
$a(x)$	1	0.5	0

2) pairwise

$$R: \quad (i,j) \in R \Rightarrow a(x_i) < a(x_j)$$

$$\sum_{(i,j)\in R} [a(x_j) - a(x_i) < 0] \to \min_a$$

$$\leq \sum_{(i,j)\in R} L(a(x_j) - a(x_i)) \to \min$$

$$L(z) = \log(1 + e^{-z})$$

RankNet

$$a(x) = \langle w, x \rangle, \quad L(z) = \log(1 + e^{-\sigma z})$$

град. шаг:

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

Теперь хотим оптимизировать более сложную метрику (напр., nDCG) F

$$w := w + \lambda \frac{\sigma}{1+\exp(\sigma\langle x_j - x_i, w\rangle)} \underbrace{|\Delta F_{ij}|}_{} (x_j - x_i)$$

как изменится F, если поменять x_i и x_j местами

LambdaRank

3) listwise

ListNet

$$DCG@k \to \max_a$$

$$\| \quad \frac{1}{|Q|}\sum_{q\in Q}\sum_{i=1}^{k} 2^{y(i)} \cdot \frac{1}{\log(i+1)} \to \max_a$$

вместо этого:

$$\sum_{\pi} P_a(\pi) \cdot \underbrace{DCG@k(\pi)}_{} \to \max_a$$

все перестановки на документах

1 0 1 0 0	0.95
1 1 0 0 0	0.01
1 0 0 1 0	⋮
. . .	
0 0 0 1 1	0.01

$\sum = 1$

q: $(d_1, \dots d_{n_q})$ – надо отранжировать

$y_1, \dots y_{n_q}$ – истинные оценки

$\boxed{z_1, \dots z_{n_q}}$ – оценки из модели $a(x)$

$$P_z(\pi) = \prod_{j=1}^{n_q} \frac{\varphi(z_{\pi(j)})}{\sum_{k=j}^{n_q} \varphi(z_{\pi(k)})}$$

: макс. вероятность имеет перест., сортирующая документы по z_i

$$\varphi(z) = e^{z}$$

$$\boxed{\sum_{\pi} \boxed{P_z(\pi)} \cdot DCG(\pi) \to \max}$$

x_i выше x_j

$z_i > z_j$

если $x_i \leftrightarrow x_j$, то

$P \downarrow$

			$P_z(\pi)$
1	0	0	0.1
0	1	0	0.8
0	0	1	0.1

1	0	0	0.8
0	1	0	0.1
0	0	1	0.1