Метрики оценки качества ранжирования

Для начала будем рассматривать бинарные оценки релевантности:

$$y_j^{(i)} \in \{0, 1\} \quad \forall i \in \{1, \dots, m\}, j \in 1, \dots, n^{(i)}$$

Зачастую нас интересует качество не на всех документах, а только на top k документов, которые выдаёт система. Одной из самых простых оценок качества является $\text{Precision@}k$:

$$\text{Precision}^{(\text{i})}@k = \frac{1}{k} \sum_{j=1}^{n^{(i)}} y^{(i)}(r_j^{(i)})$$

$$\text{Precision@}k = \frac{1}{m} \sum_{i=1}^{m} \text{Precision}^{(i)}@k$$

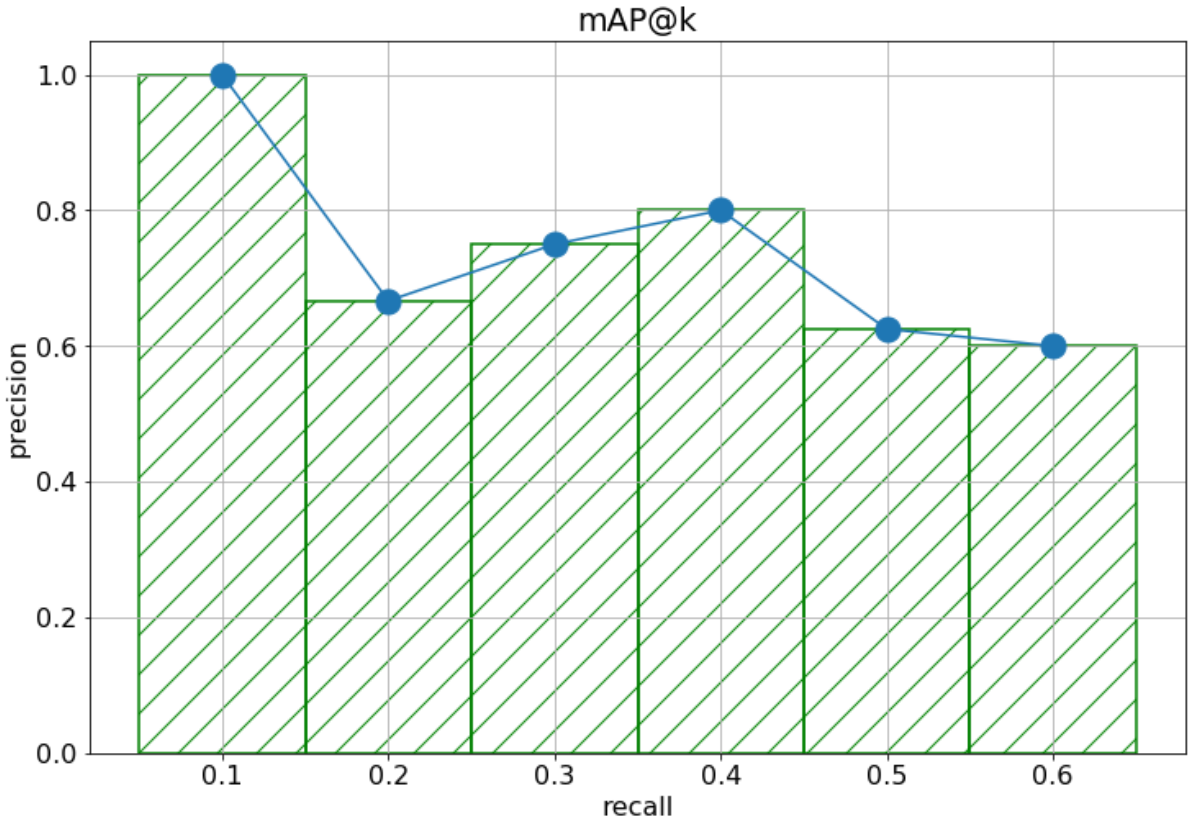
У этой метрики есть существенный недостаток, т.к. она не обращает никакого внимания на документы, так у двух списков с релевантностями $[1, 0, 1, 0, 0]$ и $[0, 1, 0, 0, 1]$ будет одинаковое значение $\text{Precision@}k = 0.4$.

Для учёта порядка как правило исользуются Average Precision и Mean Average Precision соответственно:

$$\text{AP}^{(i)}@k = \frac{\sum_{j=1}^{k} y^{(i)}(r_j) \text{Precision@}j}{\min\left(k, \sum_{j=1}^{n^{(i)}} y_j\right)}$$

$$\text{mAP@}k = \frac{1}{m} \sum_{i=1}^{m} \text{AP}^{(i)}@k$$

$\text{AP}^{(i)}@k$ можно рассматривать как площадь под кривой precision/recall для первых k документов.



Однако эти метрики не подходят для случая грейдированных оценок релевантности, который более распространён в задаче обучения ранжированию:

$$y_j^{(i)} \in \{0, \dots, l\},\ l \in \mathbb{N} \quad \forall i \in \{1, \dots, m\}, j \in 1, \dots, n^{(i)}$$

Здесь нам на помощь приходят метрики DCG и nDCG. DCG определяется как:

$$\mathrm{DCG}^{(i)}@k = \sum_{j=1}^{k} \mathrm{Gain}(y^{(i)}(r_j^{(i)})) \cdot \mathrm{Discount}(j)$$

Как правило:

$$\mathrm{Gain}(y) = 2^y - 1,\ \mathrm{Discount(j)} = \frac{1}{\log_2(j+1)}$$

$$\mathrm{DCG}^{(i)}@k = \sum_{j=1}^{k} \frac{2^{y^{(i)}(r_j^{(i)})} - 1}{\log_2(j+1)}$$

Зачастую используют нормализованную версию DCG:

$$\mathrm{IDCG}^{(i)}@k = \max_{\mathbf{s^{(i)}}} \mathrm{DCG}^{(i)}@k = \sum_{j=1}^{k} \frac{2^{y^{(i)}(r_j^{*(i)})} - 1}{\log_2(j+1)}$$

$$\mathrm{nDCG}@k = \frac{1}{m} \sum_{i=1}^{m} \frac{\mathrm{DCG}^{(i)}@k}{\mathrm{IDCG}^{(i)}@k}$$

Проблема метрики NDCG заключается в том, что j-ый документ вносит одинаковый вклад независимо от того, был ли показан выше более релевантный документ. Из соображений здравого смысла очевидно, что если документ "средней" релевантности находится ниже сильно релевантного документа, то пользователь вряд ли в принципе дойдёт до этого документа, и его конкретное местоположение будет уже гораздо меньше влиять на общее качество выдачи. Решением могут быть метрики, реализующую так называемую каскадную модель, например ERR (http://olivier.chapelle.cc/pub/err.pdf) и pFound (http://romip.ru/romip2009/15_yandex.pdf).

$$R_j^{(i)} \equiv \mathcal{R}(y^{(i)}(r_j^{(i)}))$$

$$\mathcal{R}(y) \equiv \frac{2^y - 1}{2^l}$$

$$R_j^{(i)} = \frac{2^{y^{(i)}(r_j^{(i)})} - 1}{2^l}$$

Для определения ERR нам нужно определение полезности документа на позиции $r_j^{(i)}$, $\phi(r_j^{(i)})$:

$$\phi(1) = 1;\ \phi(r) \to 0, r \to +\infty$$

$$\phi(r) = \frac{1}{r}; \quad \phi(r) = \frac{1}{\log_2(r+1)}$$

ERR определяется как математическое ожидание $\phi(r)$ для позиции r, на которой пользователь закончил просмотр выдачи:

$$\mathrm{ERR}^{(i)} = \int \phi(r^{(i)}) \mathrm{d}P(r^{(i)}) = \sum_{j=1}^{n^{(i)}} \phi(j) \mathbb{P}(j = stop)$$

$$\mathrm{ERR}^{(i)} = \sum_{j=1}^{n^{(i)}} \frac{1}{j} R_j^{(i)} \prod_{l=1}^{j-1} (1 - R_l^{(i)})$$

PFound добавляет в эту модель вероятность того, что пользователю надоест просматривать выдачу ($\mathbb{P}_{\text{break}}$):

$$\text{PFound}^{(i)} = \sum_{j=1}^{n^{(i)}} (1 - \mathbb{P}_{\text{break}})^{j-1} R_j^{(i)} \prod_{l=1}^{j-1} (1 - R_l^{(i)})$$