相关性的评价指标

王树森

http://wangshusen.github.io/

相关性的评价指标

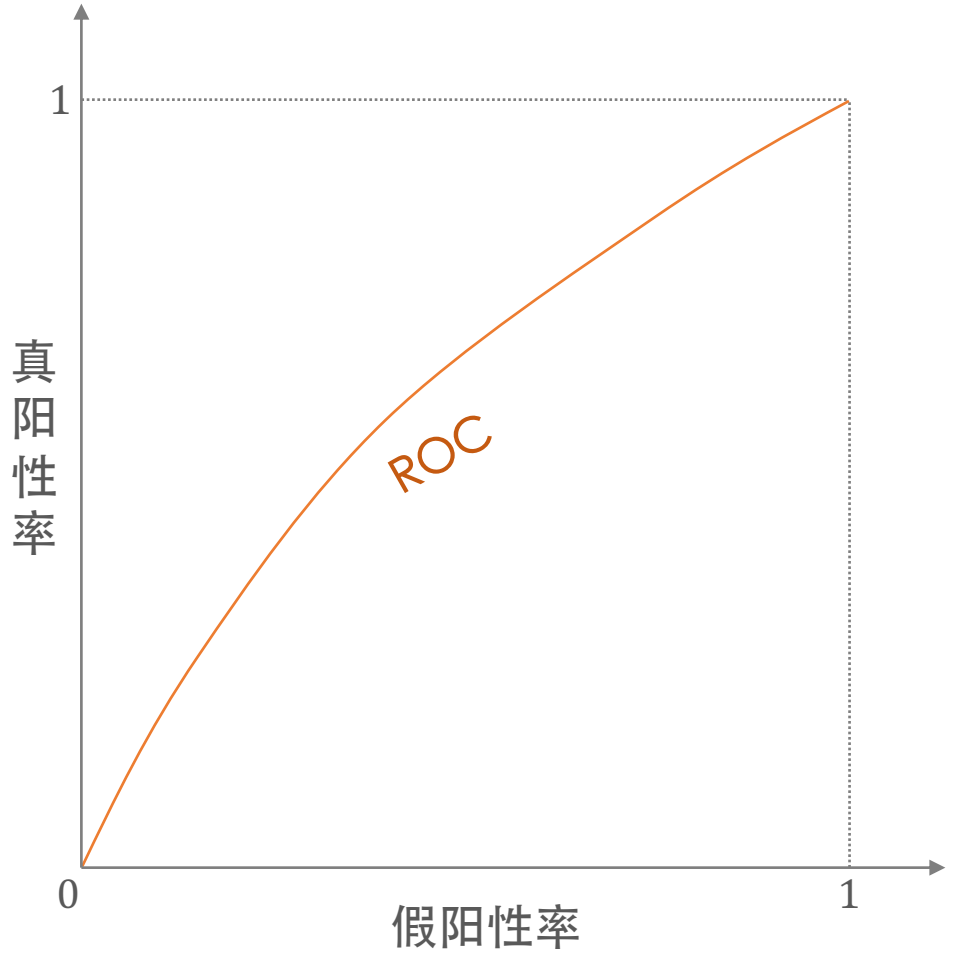
- Pointwise 评价指标：Area Under the Curve（AUC)。
- Pairwise 评价指标：正逆序比 (Positive to Negative Ratio, PNR)。
- Listwise 评价指标： Discounted Cumulative Gain (DCG)。
- 用 AUC 和 PNR 作为离线评价指标，用 DCG 评价模型在线上排序的效果。

Pointwise 评价指标

二分类评价指标

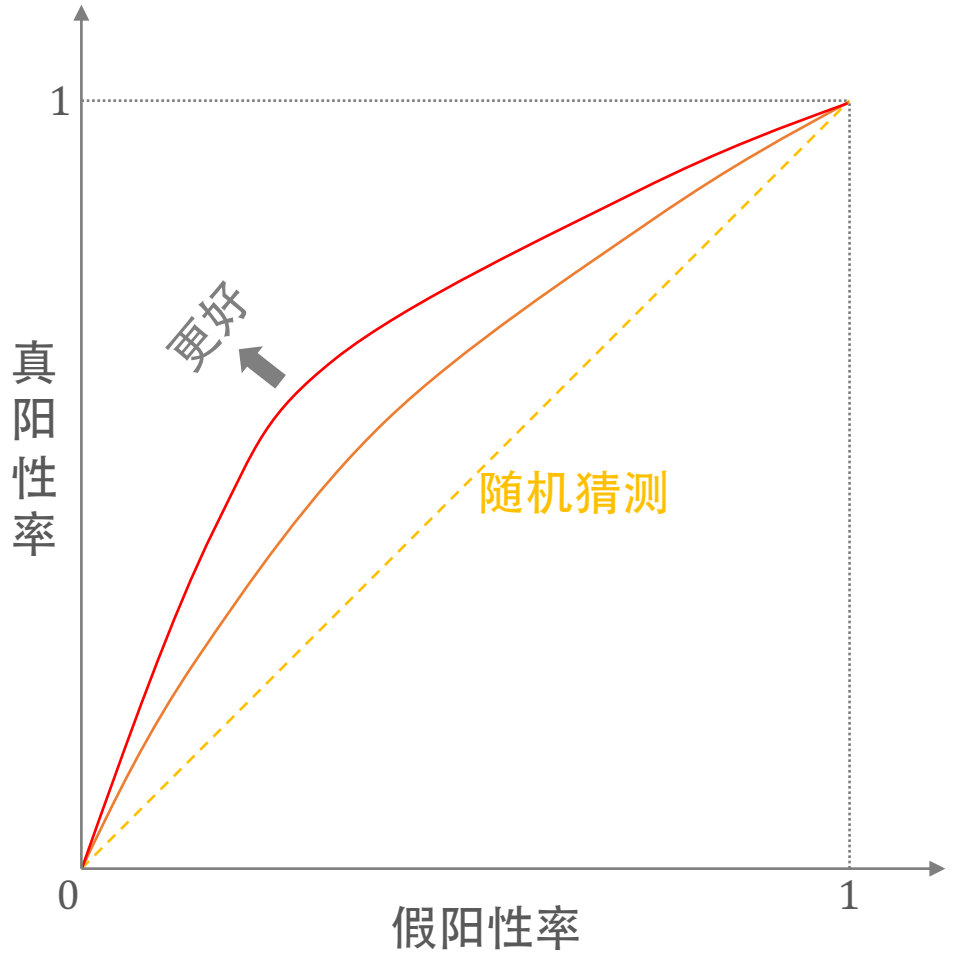
- 把测试集相关性档位转化为 0/1。
  - 高、中两档合并，作为标签 $y = 1$。
  - 低、无两档合并，作为标签 $y = 0$。
- 相关性模型输出预测值 $p \in [0, 1]$。
- 用 AUC 评价模型的预测是否准确。

二分类评价指标



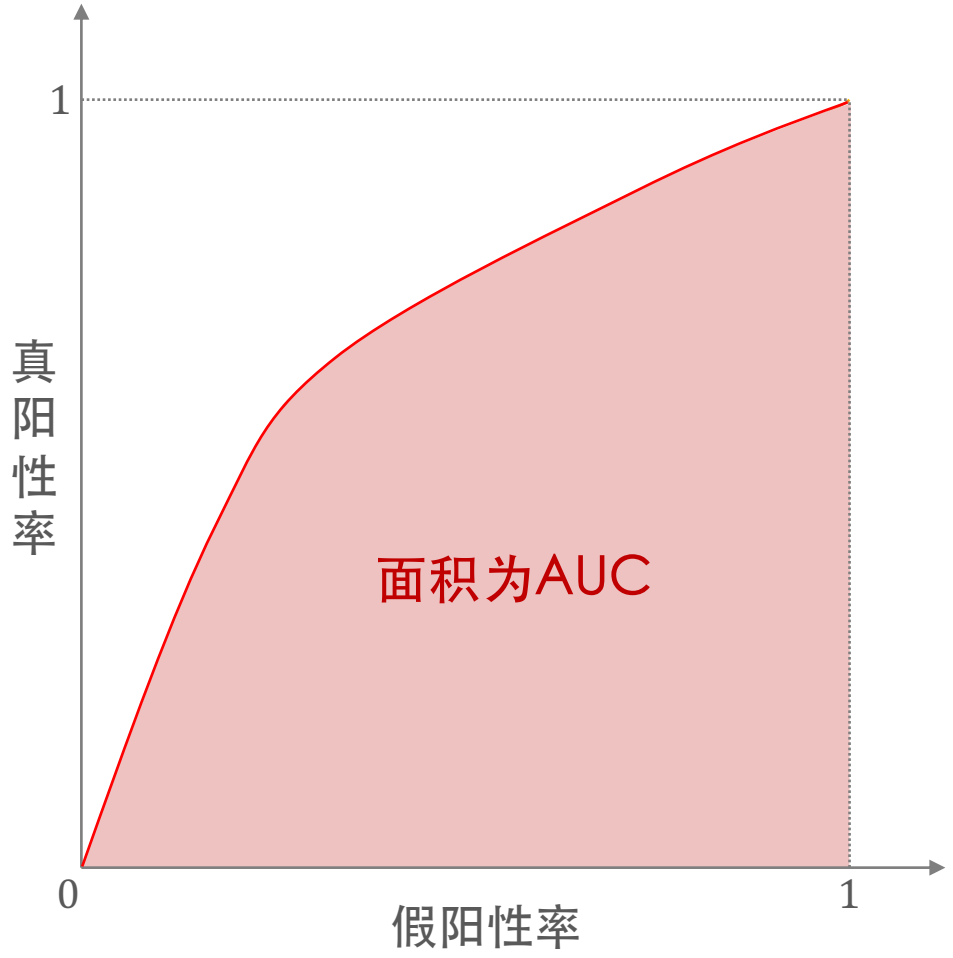
- 把测试集相关性档位转化为 0/1。
  - 高、中两档合并，作为标签 $y = 1$。
  - 低、无两档合并，作为标签 $y = 0$。
- 相关性模型输出预测值 $p \in [0, 1]$。
- 用 AUC 评价模型的预测是否准确。

二分类评价指标



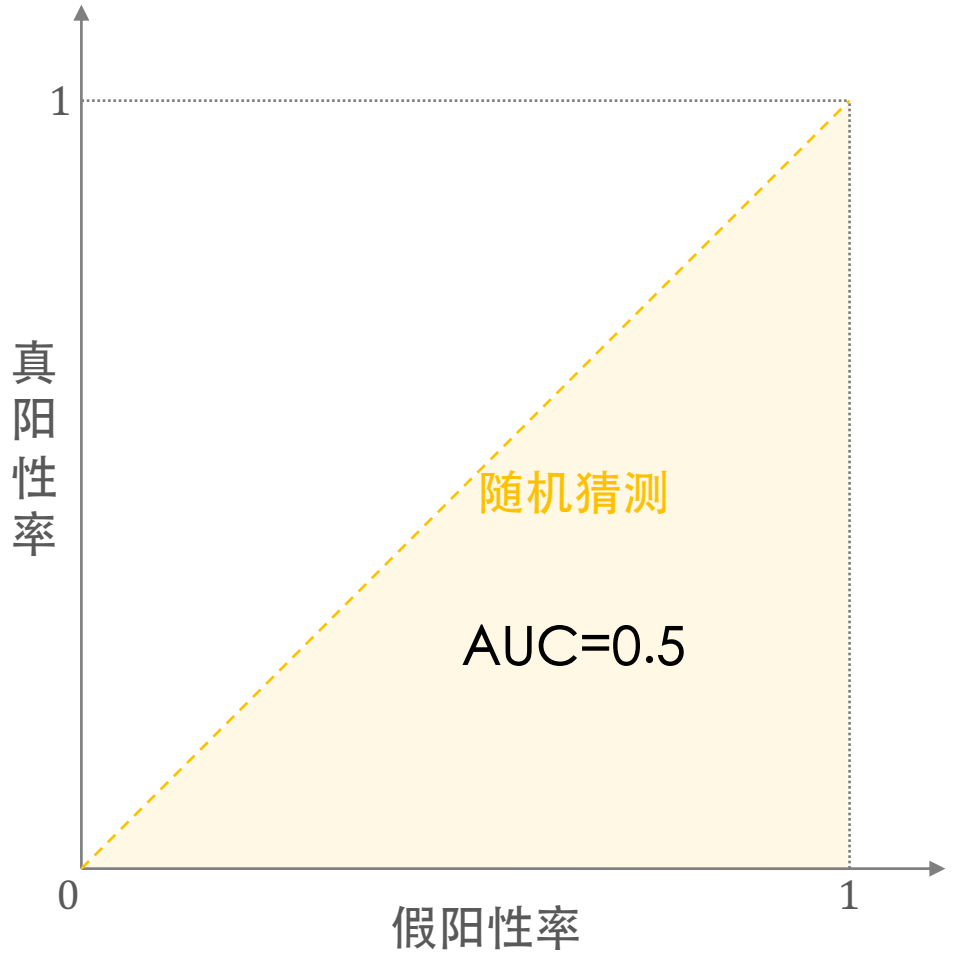
- 把测试集相关性档位转化为 0/1。
  - 高、中两档合并，作为标签 $y = 1$。
  - 低、无两档合并，作为标签 $y = 0$。
- 相关性模型输出预测值 $p \in [0, 1]$。
- 用 AUC 评价模型的预测是否准确。

二分类评价指标



- 把测试集相关性档位转化为 0/1。
  - 高、中两档合并，作为标签 $y = 1$。
  - 低、无两档合并，作为标签 $y = 0$。
- 相关性模型输出预测值 $p \in [0, 1]$。
- 用 AUC 评价模型的预测是否准确。

二分类评价指标



- 把测试集相关性档位转化为 0/1。
  - 高、中两档合并，作为标签 $y = 1$。
  - 低、无两档合并，作为标签 $y = 0$。
- 相关性模型输出预测值 $p \in [0, 1]$。
- 用 AUC 评价模型的预测是否准确。

Pairwise 评价指标

正逆序比 (PNR)

序号
1
2
3
4
5
6

- 根据模型估计的相关性分数 p 对文档做排序。（不知道真实相关性分数。）
- 左边例子中有 6 篇文档，它们的分数满足 $p_1 \geq p_2 \geq \cdots \geq p_6$。

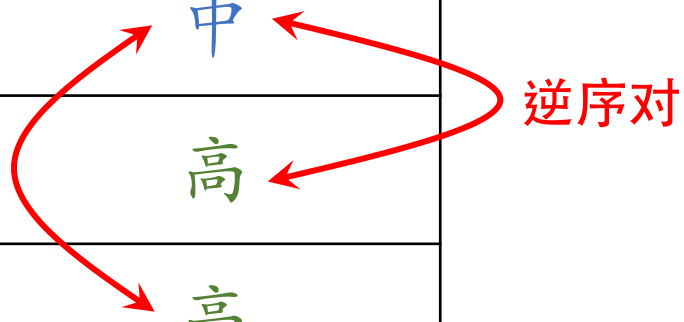
正逆序比 (PNR)

序号	真实相关性
1	高
2	中
3	高
4	高
5	中
6	低

- 根据模型估计的相关性分数 p 对文档做排序。（不知道真实相关性分数。）
- 左边例子中有 6 篇文档，它们的分数满足 $p_1 \geq p_2 \geq \cdots \geq p_6$。

正逆序比 (PNR)

序号	真实相关性
1	高
2	中
3	高
4	高
5	中
6	低



- 有 k 篇文档，则有 $\binom{k}{2} = \frac{k!}{2!\times(k-2)!}$ 种方式将文档两两组合。
- 例中 $k = 6$，有 $\binom{6}{2} = 15$ 种组合。有 2 个逆序对，13 个正序对。
- 正逆序比为：

$$\text{PNR} = \frac{13}{2}.$$

正逆序比 (PNR)

序号	真实相关性
1	高
2	中
3	高
4	高
5	中
6	低

逆序对

序号	真实相关性
1	高
2	高
3	高
4	低
5	中
6	中

Listwise 评价指标

Pairwise指标 vs Listwise指标

- 有 n 篇候选文档，根据模型打分做降序排列，把文档记作 $d_1, \cdots, d_n$。（此时不知道真实相关性分数。）
- $d_1, \cdots, d_n$ 的真实相关性分数为 $y_1, \cdots, y_n$。（人工标注相关性档位，档位映射到 $[0, 1]$ 区间上的实数。）
- 理想的排序：$y_1 \geq y_2 \geq \cdots \geq y_n$，即模型打分的序与真实相关性分数的序一致。（此时 pairwise 和 listwise 指标都最大化。）
- 逆序对会导致 pairwise 和 listwise 指标减小。
  - 逆序对出现的位置不影响 pairwise 指标。
  - 逆序对越靠前，对 listwise 指标造成的损失越大。

Cumulative Gain (CG)

- 有 n 篇候选文档，根据模型打分做降序排列；它们的真实相关性分数为 $y_1, \cdots, y_n$。
- 只关注排在前 k $(k \ll n)$ 的文档，它们最可能获得曝光，对用户体验的影响最大。
- Cumulative Gain: $\mathrm{CG@}k = \sum_{i=1}^{k} y_i$.
- $\mathrm{CG@}k$ 何时最大化？
  - 真实相关性分数 y 最高的 k 篇文档被模型排在前 k。
  - 前 k 篇文档的序不重要，它们之间可以存在逆序对。

Cumulative Gain (CG)

序号	真实相关性
1	1
2	0.7
3	0.3
4	1
5	0.7
6	0.7
7	0.3
8	0
9	0.7
10	0

问题：求 CG@4

- 定义： $\mathrm{CG}@k = \sum_{i=1}^{k} y_i$.
- $\mathrm{CG}@4 = 1 + 0.7 + 0.3 + 1 = 3$

Cumulative Gain (CG)

序号	真实相关性
1	1
2	0.7
3	0.3
4	1
5	0.7
6	0.7
7	0.3
8	0
9	0.7
10	0

问题：求 CG@4

- 定义： $\mathrm{CG@}k = \sum_{i=1}^{k} y_i$.
- $\mathrm{CG@4} = 1 + 0.7 + 0.3 + 1 = 3$

问题：什么样的排序最大化 CG@4？

Cumulative Gain (CG)

序号	真实相关性
1	1
2	1
3	0.7
4	0.7
5	0
6	0.3
7	0.7
8	0.3
9	0
10	0.7

问题：求 CG@4

- 定义： $\mathrm{CG@}k = \sum_{i=1}^{k} y_i$.
- $\mathrm{CG@4} = 1 + 0.7 + 0.3 + 1 = 3$

问题：什么样的排序最大化 CG@4？

- $\mathrm{CG@4} = 1 + 1 + 0.7 + 0.7 = 3.4$

Cumulative Gain (CG)

序号	真实相关性
1	1
2	1
3	0.7
4	0.7
5	0
6	0.3
7	0.7
8	0.3
9	0
10	0.7

问题：求 CG@4

- 定义： $\mathrm{CG@}k = \sum_{i=1}^{k} y_i$.
- $\mathrm{CG@4} = 1 + 0.7 + 0.3 + 1 = 3$

问题：什么样的排序最大化 CG@4？

- $\mathrm{CG@4} = 1 + 1 + 0.7 + 0.7 = 3.4$

Cumulative Gain (CG)

序号	真实相关性
1	1
2	1
3	0.7
4	0.7
5	0
6	0.3
7	0.7
8	0.3
9	0
10	0.7

问题：求 CG@4

- 定义：$\mathrm{CG@}k = \sum_{i=1}^{k} y_i$.
- $\mathrm{CG@4} = 1 + 0.7 + 0.3 + 1 = 3$

问题：什么样的排序最大化 CG@4？

- $\mathrm{CG@4} = 1 + 1 + 0.7 + 0.7 = 3.4$
- 交换前 4 文档顺序不影响 CG@4。

Discounted Cumulative Gain (DCG)

- 有 n 篇候选文档，根据模型打分做降序排列。它们的真实相关性分数为 $y_1, \cdots, y_n$。
- Discounted Cumulative Gain:

$$\mathrm{DCG}@k = \sum_{i=1}^{k} \frac{y_i}{\log_2(i+1)}.$$

Discounted Cumulative Gain (DCG)

- 有 n 篇候选文档，根据模型打分做降序排列，它们的真实相关性分数为 $y_1, \cdots, y_n$。
- Discounted Cumulative Gain:

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{y_i}{\log_2(i+1)}.$$

- $\mathrm{DCG@}k$ 何时最大化？
  - 真实相关性分数 y 最高的 k 篇文档被模型排在前 k。
  - 前 k 篇文档不存在逆序对。

Discounted Cumulative Gain (DCG)

序号	真实相关性
1	1
2	0.7
3	0.3
4	1
5	0.7
6	0.7
7	0.3
8	0
9	0.7
10	0

问题：求 DCG@4

- $\mathrm{DCG@4} = \sum_{i=1}^{4} \frac{y_i}{\log_2(i+1)}$

$$= \frac{1}{\log_2 2} + \frac{0.7}{\log_2 3} + \frac{0.3}{\log_2 4} + \frac{1}{\log_2 5}$$

Discounted Cumulative Gain (DCG)

序号	真实相关性
1	1
2	0.7
3	0.3
4	1
5	0.7
6	0.7
7	0.3
8	0
9	0.7
10	0

问题：求 DCG@4

- $\mathrm{DCG@4} = \sum_{i=1}^{4} \frac{y_i}{\log_2(i+1)}$

$$= \frac{1}{\log_2 2} + \frac{0.7}{\log_2 3} + \frac{0.3}{\log_2 4} + \frac{1}{\log_2 5} = 2.02$$

Discounted Cumulative Gain (DCG)

序号	真实相关性
1	1
2	0.7
3	0.3
4	1
5	0.7
6	0.7
7	0.3
8	0
9	0.7
10	0

问题：求 DCG@4

- $\mathrm{DCG@4} = \sum_{i=1}^{4} \frac{y_i}{\log_2(i+1)}$

$$= \frac{1}{\log_2 2} + \frac{0.7}{\log_2 3} + \frac{0.3}{\log_2 4} + \frac{1}{\log_2 5} = 2.02$$

问题：什么样的排序最大化 DCG@4？

Discounted Cumulative Gain (DCG)

序号	真实相关性
1	1
2	1
3	0.7
4	0.7
5	0
6	0.7
7	0.3
8	0
9	0.3
10	0.7

问题：求 DCG@4

- $\mathrm{DCG@4} = \sum_{i=1}^{4} \frac{y_i}{\log_2(i+1)}$

$$= \frac{1}{\log_2 2} + \frac{0.7}{\log_2 3} + \frac{0.3}{\log_2 4} + \frac{1}{\log_2 5} = 2.02$$

问题：什么样的排序最大化 DCG@4？

- $\mathrm{DCG@4} = \frac{1}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0.7}{\log_2 4} + \frac{0.7}{\log_2 5}$

Discounted Cumulative Gain (DCG)

序号	真实相关性
1	1
2	1
3	0.7
4	0.7
5	0
6	0.7
7	0.3
8	0
9	0.3
10	0.7

问题：求 DCG@4

- $\mathrm{DCG@4} = \sum_{i=1}^{4} \frac{y_i}{\log_2(i+1)}$

$$= \frac{1}{\log_2 2} + \frac{0.7}{\log_2 3} + \frac{0.3}{\log_2 4} + \frac{1}{\log_2 5} = 2.02$$

问题：什么样的排序最大化 DCG@4？

- $\mathrm{DCG@4} = 2.28$

Discounted Cumulative Gain (DCG)

序号	真实相关性
1	1
2	1
3	0.7
4	0.7
5	0
6	0.7
7	0.3
8	0
9	0.3
10	0.7

问题：求 DCG@4

- $\mathrm{DCG@4} = \sum_{i=1}^{4} \frac{y_i}{\log_2(i+1)}$

$$= \frac{1}{\log_2 2} + \frac{0.7}{\log_2 3} + \frac{0.3}{\log_2 4} + \frac{1}{\log_2 5} = 2.02$$

问题：什么样的排序最大化 DCG@4？

- $\mathrm{DCG@4} = 2.28$
- 前 4 出现逆序对会让 DCG@4 减小。

总结

相关性的评价指标

- 相关性有 pointwise、 pairwise、 listwise 评价指标。
- Pointwise： 单独评价每一个 (q, d) 二元组，判断预测的相关性分数与真实标签的相似度。
- Pairwise： 对比 (q, d_1) 和 (q, d_2)，判断两者的序是否正确（正序对或逆序对）。
- Listwise： 对比 $(q, d_1), (q, d_2), \cdots, (q, d_n)$，判断整体的序关系的正确程度。

离线评价指标

- 事先准备人工标注的数据，划分为训练集和测试集。
- 完成训练之后，计算测试集上的 AUC 和 PNR。
- 相关性有 4 个档位，为什么不用多分类的评价指标（Macro F1 和 Micro F1）？
  - 相关性的标签存在序关系：高 > 中 > 低 > 无。
  - 多分类把 4 种标签看做 4 个类别，忽略其中的序关系。
  - 把“高”错判为“中”、或错判为“无”，错误严重程度不同。但被多分类视为同等的分类错误。

线上评价指标

- 一个搜索 session：用户搜索 q，搜索结果页上按顺序展示文档 $d_1, \cdots, d_n$。
- 从搜索日志中抽取一批 session，覆盖高、中、低频查询词。
- 对于每个 session，取排序最高的 k 篇文档 $d_1, \cdots, d_k$。
  - k 的设定取决于用户浏览深度，比如 $k = 20$。
  - 高频查询词前 20 篇文档几乎都是高相关，指标过高。
  - 高频查询词的 k 设置得较大（比如 $k = 40$），低频查询词的 k 设置得较小（比如 $k = 20$）。

线上评价指标

- 一个搜索 session：用户搜索 q，搜索结果页上按顺序展示文档 $d_1, \cdots, d_n$。
- 从搜索日志中抽取一批 session，覆盖高、中、低频查询词。
- 对于每个 session，取排序最高的 k 篇文档 $d_1, \cdots, d_k$。
- 人工标注相关性分数，记作 $y_1, \cdots, y_k$ 。
- 计算 $\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{y_i}{\log_2(i+1)}$，作为该 session 的评价指标。
- 对 $\mathrm{DCG@}k$ 关于所有 session 取平均，评价线上相关性模型。

思考题

- $\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}$ 是教科书中经典的评价指标。
  - 其中 $\text{IDCG@}k$ 是 $\text{DCG@}k$ 的最优值，对应最优的排序。
  - 因此 $\text{NDCG@}k$ 的值介于 0 和 1之间。
- 问题：NDCG 可否代替 DCG 用作线上评价指标？NDCG 有什么缺陷？
- 提示：
  - 先做召回，再做排序。假设召回的结果全是低相关文档。
  - DCG 是高是低？NDCG 是高是低？DCG 与 NDCG 谁更合理？

Thank You!

http://wangshusen.github.io/