CS657A: Information Retrieval
Probabilistic Retrieval Model

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs657/

2nd semester, 2021-22

Tue 1030-1145, Thu 1200-1315

Probabilistic Retrieval Task

- Every document has a *probability* of being relevant for a query
- $P(R = 1|d, q)$ or $P(R|d, q)$
- Probability ranking principle (PRP) ranks documents according to their probability of being relevant to the query

Binary Loss

- Simplest model is binary loss or $1/0$ loss
- Penalty is 1 for not retrieving a relevant document
- Penalty is 1 for retrieving an irrelevant document
- Bayes decision rule is to return documents that are more likely to be relevant than irrelevant
- Document d is relevant if and only if $P(R|d,q) > P(\overline{R}|d,q)$
- If k documents are returned, it is best to return according to $P(R|d,q)$
- This minimizes the *expected loss* or Bayes risk
- This assumes that all probabilities $P(R|d,q)$ can be estimated correctly

Retrieval Model

- Retrieval errors have costs
- Cost of not retrieving a relevant document is C_1
- Cost of retrieving an irrelevant document is C_0
- For a document d, loss is

$$L(d) = C_0.P(\overline{R}|d,q) - C_1.P(R|d,q)$$

- To minimize loss, the next document retrieved should be d such that for any other document e, $L(d) \leq L(e)$

Binary Independence Model

- Binary Independence Model (BIM) to estimate the probabilities
- Boolean or binary model
- Document d and query q are binary vectors of terms that appear in them
- d is represented as $\vec{d} = (d_1, \ldots, d_m)$
  - $d_i = 1$ if term i is present in d
  - $d_i = 0$ otherwise
- q is similarly represented as $\vec{q}$
- Terms are independent of each other
- Relevance of document d is independent of relevances of other documents

Bayes Assumption

- Using Bayes' rule

$$P(R=1|\vec{d},\vec{q}) = P(R|\vec{d},\vec{q}) = \frac{P(\vec{d}|R,\vec{q}).P(R|\vec{q})}{P(\vec{d}|\vec{q})}$$

$$P(R=0|\vec{d},\vec{q}) = P(\overline{R}|\vec{d},\vec{q}) = \frac{P(\vec{d}|\overline{R},\vec{q}).P(\overline{R}|\vec{q})}{P(\vec{d}|\vec{q})}$$

- $P(R|\vec{q})$ and $P(\overline{R}|\vec{q})$ are the prior probabilities of a document being relevant or irrelevant to the query
  - $P(R|\vec{q}) + P(\overline{R}|\vec{q}) = 1$
- Odds of a document being relevant is the ratio of its probability to its complement

$$O(R|\vec{q}) = \frac{P(R|\vec{q})}{P(\overline{R}|\vec{q})}$$

Ranking

- Ranking by probability remains the same as ranking by *odds*

$$O(R|\vec{d},\vec{q}) = \frac{P(R|\vec{d},\vec{q})}{P(\overline{R}|\vec{d},\vec{q})} = \frac{P(R|\vec{q})}{P(\overline{R}|\vec{q})}.\frac{P(\vec{d}|R,\vec{q})}{P(\vec{d}|\overline{R},\vec{q})} = O(R|\vec{q}).\frac{P(\vec{d}|R,\vec{q})}{P(\vec{d}|\overline{R},\vec{q})}$$

- $O(R|\vec{q})$ is the same for all documents and, hence, does not affect ranking
- $\vec{d}$ consists of terms $d_1, \dots, d_m$
- Naïve Bayes assumption: Terms are *conditionally independent* of each other
- Thus,

$$O(R|\vec{d},\vec{q}) \propto \frac{P(\vec{d}|R,\vec{q})}{P(\vec{d}|\overline{R},\vec{q})} = \prod_{t=1}^{m} \frac{P(d_t|R,\vec{q})}{P(d_t|\overline{R},\vec{q})}$$

- $P(d_t|R,\vec{q})$ is the probability of the term d_t appearing in a document that is relevant to q

Manipulating Terms

- Each term d_t is either 0 or 1 for a document d
- Separating the terms

$$O(R|\vec{d},\vec{q}) = \prod_{\forall d_t=1} \frac{P(d_t=1|R,\vec{q})}{P(d_t=1|\overline{R},\vec{q})} . \prod_{\forall d_t=0} \frac{P(d_t=0|R,\vec{q})}{P(d_t=0|\overline{R},\vec{q})}$$

- Denote $p_t = P(d_t = 1|R,\vec{q})$ of probability of a term appearing in a relevant document
- Then, $P(d_t = 0|R,\vec{q}) = 1 - p_t$
- Denote $u_t = P(d_t = 1|\overline{R},\vec{q})$ of probability of a term appearing in an irrelevant document
- Then, $P(d_t = 0|\overline{R},\vec{q}) = 1 - u_t$

	Relevant (R)	Irrelevant ($\overline{R}$)
Term present $d_t = 1$	p_t	u_t
Term absent $d_t = 0$	$1 - p_t$	$1 - u_t$

Query Terms

- Consider terms that does not appear in query, i.e., $q_t = 0$
- Assume that they are equally likely to occur in relevant versus irrelevant documents
  - When $q_t = 0$, $p_t = u_t$
- Thus, they need not be considered for ranking

$$O(R|\vec{d}, \vec{q}) = \prod_{\forall d_t=1} \frac{p_t}{u_t} . \prod_{\forall d_t=0} \frac{1-p_t}{1-u_t} \propto \prod_{\forall d_t=1, q_t=1} \frac{p_t}{u_t} . \prod_{\forall d_t=0, q_t=1} \frac{1-p_t}{1-u_t}$$

- Further manipulating,

$$\begin{aligned} O(R|\vec{d}, \vec{q}) &= \prod_{\forall d_t=1, q_t=1} \frac{p_t}{u_t} . \prod_{\forall d_t=0, q_t=1} \frac{1-p_t}{1-u_t} \\ &= \prod_{\forall d_t=1, q_t=1} \left(\frac{p_t}{u_t} \Big/ \frac{1-p_t}{1-u_t} \right) . \prod_{\forall d_t=0, q_t=1} \frac{1-p_t}{1-u_t} . \prod_{\forall d_t=1, q_t=1} \frac{1-p_t}{1-u_t} \\ &= \prod_{\forall d_t=1, q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} . \prod_{\forall q_t=1} \frac{1-p_t}{1-u_t} \end{aligned}$$

Odds Ratio

- $\prod_{\forall q_t=1} \frac{1-p_t}{1-u_t}$ is constant for all documents
- Therefore, finally,

$$O(R|\vec{d},\vec{q}) \propto \prod_{\forall d_t=1, q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

- Ranking is done by retrieval status value (RSV) which is logarithm of the odds

$$RSV_d = \log O(R|\vec{d},\vec{q}) = \sum_{\forall d_t=1, q_t=1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{\forall d_t=1, q_t=1} c_t$$

- For each term, c_t is the *logarithm* of odds ratio

$$c_t = \log \frac{p_t/(1-p_t)}{u_t/(1-u_t)} = \log \frac{p_t}{1-p_t} - \log \frac{u_t}{1-u_t}$$

Estimates

- Contingency table of counts of documents for a term

	Relevant (R)	Irrelevant ($\overline{R}$)	Total
Term present $d_t = 1$	r	$df_t - r$	df_t
Term absent $d_t = 0$	$\lvert R\rvert - r$	$(n - df_t) - (\lvert R\rvert - r)$	$n - df_t$
Total	$\lvert R\rvert$	$n - \lvert R\rvert$	n

- Using p_t, u_t, etc.,

$$c_t = \log \frac{\frac{r}{|R|-r}}{\frac{df_t-r}{n-df_t-|R|+r}}$$

- To counter zero counts, a pseudo-count of 0.5 is added

$$c_t = \log \frac{\frac{r+0.5}{|R|-r+0.5}}{\frac{df_t-r+0.5}{n-df_t-|R|+r+0.5}}$$

Example

- d_1: Indians had an advanced system of surgery including rhinoplasty
- d_2: Surprisingly, many ancient civilizations had well designed systems in place for games and sports
- d_3: Ancient Indians invented the Pythagoras' theorem, binary system and many other mathematical systems
- d_4: Indian civilization is one of the most ancient ones
- d_5: In India, when systems become old, they tend to disintegrate
- Query q: ancient Indian system
- Relevant: d_1, d_2, d_3; Irrelevant: d_4, d_5

Term	r	$\|R\| - r$	df_t	$df_t - r$	$n - df_t - \|R\| + r$	c_t
ancient	2	1	3	1	1	$\log \frac{2.5/1.5}{1.5/1.5} = +0.74$
Indian	2	1	4	2	0	$\log \frac{2.5/1.5}{2.5/0.5} = -1.59$
system	3	0	4	1	1	$\log \frac{3.5/0.5}{1.5/1.5} = +2.81$

Example (contd.)

- Odds of documents
- d_1: $0 - 1.59 + 2.81 = +1.22$
- d_2: $+0.74 + 0 + 2.81 = +3.55$
- d_3: $+0.74 - 1.59 + 2.81 = +1.96$
- d_4: $+0.74 - 1.59 + 0 = -0.85$
- d_5: $0 - 1.59 + 2.81 = +1.22$
- Ranking: d_2, d_3, (d_1, d_5), d_4
- Negative influences can be bounded as 0
- d_1: $0 + 0 + 2.81 = +2.81$
- d_2: $+0.74 + 0 + 2.81 = +3.55$
- d_3: $+0.74 + 0 + 2.81 = +3.55$
- d_4: $+0.74 + 0 + 0 = +0.74$
- d_5: $0 + 0 + 2.81 = +2.81$
- Ranking: (d_2, d_3), (d_1, d_5), d_4

Estimates of Probabilities

- Training data may be absent for all the terms
- For irrelevant documents,

$$\log \frac{1-u_t}{u_t} \approx \log \frac{n-df_t}{df_t} \approx \log \frac{n}{df_t}$$

- This is the standard *idf*
- For relevant documents,

$$p_t = 0.5$$

- $\log \frac{p_t}{1-p_t} = 0$ and has no effect
- p_t generally increases with df_t
- One model is $p_t = \frac{1}{3} + \frac{2}{3}.\frac{df_t}{n}$
- Better is to use relevance feedback

Relevance Feedback

- Asking users repetitively to improve the estimates
- Start with $p_t^{(0)} = 0.5$
- $u_t/(1 - u_t)$ remains same as n/df_t
- Retrieve initial set of documents $V^{(0)}$ using these parameters
- Ask user for feedback on this
- Mark relevant ones as $VR^{(0)}$
- Find $VR_t^{(0)}$ among them that contains term t
- Then, next guess of p_t is

$$p_t = \frac{|VR_t|}{|VR|}$$

- Pseudo-counts are generally added

$$p_t = \frac{|VR_t| + 0.5}{|VR| + 1}$$

- Iterations stop when estimates do not improve much

Pseudo-Count Prior

- Whether a document is relevant or not is a Bernoulli trial
- Robustly estimating a distribution parameter requires conjugate priors
- Priors model the prior belief about the pseudo-counts
- A batsman scores 100 in his first innings
- Is he the world's best scoring batsman?
- Suppose, without any *evidence*, a batsman is supposed to score at an average of 35 runs per innings
- Also suppose, at least 10 innings are required before averages stabilize
- Using prior, average is *not* $100/1$ but $(100 + 35 \times 10)/(1 + 10) = 41$
- Conjugate prior for Bernoulli distribution is beta distribution
- Pseudo-count of τ is added to estimate the next p_t:

$$p_t^{(k+1)} = \frac{|VR_t| + \tau . p_t^{(k)}}{|VR| + \tau}$$

- $\tau = 5$ is a good pseudo-count

Non-Binary Model

- Binary models ignore term weights and document lengths
- One of the most important and successful *non-binary* models is Okapi BM25 or simply BM25
- Still assumes independence of terms, though

Global and Local Weights

- Binary version only considers the idf, i.e., the global weight of a term
- Local weights are captured by tf's
- Final weight of a term is

$$w_t = l_t.g_t$$

2-Poisson Eliteness Model

- For a term, certain documents are "elite"
  - Essentially, they are the relevant documents
- There is a probability of a document being elite for a term
- Probability depends on term frequency
- Consider a document of length n to have n slots
- In each slot, independently, a particular term has a probability of appearing – Bernoulli distribution
- Term frequency, thus, follows the binomial distribution
- Approximated by Poisson distribution
- Depending on eliteness, parameters of Poisson vary
- Thus, 2 Poisson models

Saturation with Term Frequency

- Raw term frequency is too drastic to serve as local weight
- It should saturate, i.e., asymptotically approach a maximum
- Desirable properties
  - $f(tf) = 0$ when $tf = 0$
  - $f(tf) \propto tf$
  - $f(tf) \rightarrow max$ as $tf \rightarrow \infty$
- Many possible functions
- Easier to understand

$$f(tf) = \frac{tf}{k_1 + tf}$$

- Final form

$$f(tf) = \frac{(k_1 + 1)tf}{k_1 + tf}$$

- Has the desirable property that $f(tf) = 1$ when $tf = 1$

Document Length

- Documents can be longer due to two issues
- Verbosity
  - Dear Prof., I was wondering if perhaps you might have possibly gotten the chance to potentially find the time to may be look at the draft paper (which I am attaching again just in case).
  - Read.
- Scope
  - Read. Submit write-up. P.S.: With graphs.
- Generally, a mixture of both the issues
- Term frequency should be adjusted with document length

Document Length Normalization

- Document length is sum of all raw term frequencies
- Longer documents should not be allowed more weight
- Thus, for document of length L_d, normalized to

$$\frac{L_d}{L_{avg}}$$

- Another parameter b to control effect of normalization

$$1 - b + b.\frac{L_d}{L_{avg}}$$

  - $b = 1$: full normalization
  - $b = 0$: no normalization
- Normalization added to k_1 of term weight

BM25

- Full model

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{n}{df_t} \right) \cdot \frac{(k_1 + 1) tf_d}{k_1 (1 - b + b . \frac{L_d}{L_{avg}}) + tf_d}$$

when ground truth is not available

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{\frac{r + 0.5}{|R| - r + 0.5}}{\frac{df_t - r + 0.5}{n - df_t - |R| + r + 0.5}} \right) \cdot \frac{(k_1 + 1) tf_d}{k_1 (1 - b + b . \frac{L_d}{L_{avg}}) + tf_d}$$

when ground truth is available

- k_1 between 1.2 and 2.0; $b = 0.75$

Earlier Versions

- BM0

$$RSV_d = \sum_{\forall t \in q} 1$$

- BM1 ($k_1 = 0$)

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{n}{df_t} \right)$$

- BM11 ($b = 1$)

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{n}{df_t} \right) \cdot \frac{(k_1 + 1) tf_d}{k_1 . \frac{L_d}{L_{avg}} + tf_d}$$

- BM15 ($b = 0$)

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{n}{df_t} \right) \cdot \frac{(k_1 + 1) tf_d}{k_1 + tf_d}$$

Fuller Variants of BM25

- Term frequencies in query may also be considered

$$\frac{(k_3 + 1)tf_q}{k_3 + tf_q}$$

- Further global correction with document length

$$k_2.|Q|.\frac{L_{avg} - L_d}{L_{avg} + L_d}$$

- Fullest variant of BM25

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{n}{df_t} \right) . \frac{(k_1 + 1)tf_d}{k_1(1 - b + b.\frac{L_d}{L_{avg}}) + tf_d} . \frac{(k_3 + 1)tf_q}{k_3 + tf_q} + k_2.|Q|.\frac{L_{avg} - L_d}{L_{avg} + L_d}$$

- Experimentally, these additions are not very successful

BM25+

- Consider a very long document
- L_d/L_{avg} is very large
- Consequently, fraction is almost 0
- Thus, scored similarly with a short document with *no* query term
- Not properly lower bounded
- BM25+

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{n}{df_t} \right) . \left(\frac{(k_1 + 1) tf_d}{k_1 (1 - b + b.\frac{L_d}{L_{avg}}) + tf_d} + \delta \right)$$

BM25F

- Documents have *fields* or streams
- They are weighted differently
- s streams having corresponding weights v_s
- Weighted average of values

$$\widetilde{tf}_d = \sum_{\forall s} v_s.tf_s \qquad \widetilde{L}_d = \sum_{\forall s} v_s.|dl_s| \qquad \widetilde{df}_t = \sum_{\forall s} v_s.df_{t,s}$$

- Simple BM25F

$$RSV_d = \sum_{\forall t \in q} \left(\log \frac{n}{\widetilde{df}_t} \right) . \frac{(k_1 + 1)\widetilde{tf}_d}{k_1(1 - b + b.\frac{\widetilde{L}_d}{\widetilde{L}_{avg}}) + \widetilde{tf}_d}$$

- Integer stream weights is same as repeating a stream its weight number of times
- Variant
    - Even the parameters can be stream-specific
    - In particular, b_s is found to be useful