Information Retrieval (5LN712)

Probabilistic Information Retrieval

Ali Basirat

Department of Linguistics and Philology
Uppsala University

May 13, 2020

Table of contents

1 Introduction

2 Basic Probability Theory

3 The Probability Ranking Principle

4 The Binary Independence Model (BIM)

5 Summary

Introduction

- To use probabilistic classifiers to distinguish between relevant and non-relevant documents
- What is the probability of the occurrence of a term in a relevant or non-relevant document
- The Boolean and vector space models do not deal with the uncertainty involved in a query
- The uncertainty about the relevance of a query and a document can be measured by the probability tools
- The probability theory provide a principled foundation for reasoning under uncertainty

Table of contents

Introduction

Basic Probability Theory

The Probability Ranking Principle

The Binary Independence Model (BIM)

Summary

1 Introduction

2 Basic Probability Theory

3 The Probability Ranking Principle

4 The Binary Independence Model (BIM)

5 Summary

Random Variable

- A *sample space* is a set of all possible outcomes of an experiment
- An *event* is the set of outcomes of an experiment (a subset of a sample space)
- A *variable* represents an event
- The complement of an event A, denoted by $\bar{A}$, includes all elements of the sample space that are not in A
- A *random variable* maps an event to a real number

Random Variable

Example

Example

- Experiment: we roll a dice
- Sample space: the six possible states
- Event A: only one dot is seen
- $\bar{A}$: two, or, three, ..., or six dots is seen
- The random variable **A** is the number of dots seen $(A : \mathbf{A} = 1)$

Probability

- The probability of a random variable tells us about the degree of certainty that the corresponding event happen in the real world.
- How probable an event is?
- A probability is a real value between zero and one
- The probability of zero means the event does not happen
- The probability of one means the event happens surly

Example

- We roll a dice. Let **A** be a random variable that represents the number of dots seen.
- What is the probability of the event A to see only one dot?

Example

- We roll a dice. Let **A** be a random variable that represents the number of dots seen.
- What is the probability of the event A to see only one dot?
- $P(A) = P(\mathbf{A} = 1) = \frac{1}{6}$

Example

- We roll a dice. Let **A** be a random variable that represents the number of dots seen.
- What is the probability of the event A to see only one dot?
- $P(A) = P(\mathbf{A} = 1) = \frac{1}{6}$
- What is the probability of $\bar{A}$?

Example

- We roll a dice. Let **A** be a random variable that represents the number of dots seen.
- What is the probability of the event A to see only one dot?
- $P(A) = P(\mathbf{A} = 1) = \frac{1}{6}$
- What is the probability of $\bar{A}$?
- $1 - P(A)$

Conditional Probability

- We may want to estimate the probabilities based on a subset B of the sample space
- What is the probability of an event A if we know that another event B occurred ($P(A|B)$).

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example

- We roll two dices. If the first one shows an odd number. What is the probability that the sum of the two numbers is 6?
- $S = \{(x, y)|x = 1, \ldots, 6, y = 1, \ldots, 6\}$
- **B**: the first dice shows an odd number
- $B = \{(1, x), (3, x), (5, x)|x = 1, \ldots, 6\}$, $P(B) = \frac{18}{36}$
- **A**: the sum of two dices is 6
- $A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$, $P(A) = \frac{5}{36}$
- $A \cap B = \{(1, 5), (3, 3), (5, 1)\}$, $P(A \cap B) = \frac{3}{36}$
- $P(A|B) = \frac{3}{18}$
- $P(B|A) = \frac{3}{5}$

Joint Events

- Two (or more) events occur together
- The joint event of the two events A and B is the intersection of the two events $A \cap B$
- The probability of the joint event $A \cap B$ is represented by $P(A \cap B)$ or $P(A, B)$
- The joint probability $P(A, B)$ can be calculated by the chain rule

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

Total Probability Rule

- A set of events $A_1, A_2, \ldots, A_n$ partition a sample space if they are mutually disjoint and their union is the entire sample space
- If $A_1, A_2, \ldots, A_n$ partition a sample space, the probability of an event B in the sample space is:

$$P(B) = P(B, A_1) + \cdots + P(B, A_n) = \sum_{i=1}^{n} P(B, A_i)$$

Total Probability Rule

Example

Example

Any event A in a sample space and its complementary event $\bar{A}$ partition the sample space. Hence, the probability of any event B in the sample space is:

$$P(B) = P(B, A) + P(B, \bar{A}) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

Bayes' Rule

- General case: If $A_1, \ldots, A_n$ partition the sample space

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B, A_i)}$$

- Special case: If $\bar{A}$ is the complement event of A

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

Bayes' Rule

Interpretation

- The probabilities $P(A_i)$ are the prior probabilities.
- The prior probability is an initial estimate of how likely A_i is when we do not have any other information
- The Bayes' rule tells us how the prior probabilities change if another event B has occurred.
- The posterior probability $P(A_i|B)$ measures the probability of A_i after the evidence B is taken into account.

Odds

- The odds of an event A is the ratio of the probability of the event to the probability of its complement.

$$O(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}$$

- The odds of an event tells us how likely the event will take place

Table of contents

1 Introduction

2 Basic Probability Theory

3 The Probability Ranking Principle

4 The Binary Independence Model (BIM)

5 Summary

The Probability Ranking Principle

- Documents are ranked based on their relevance probability to a query
- For each pair of document d and query q, a Bernoulli indicator (random variable) R is defined that takes a value of 1 if d is a relevant document to q
- The document d_1 is more relevant to the query q than a document d_2 if

$$P(\mathbf{R} = 1|d_1, q) > P(\mathbf{R} = 1|d_2, q)$$

The 0/1 Loss

- When making a decision about the relevance of documents, no extra cost is considered about possible failures
- A document is either relevant or not with not decision making cost

The 0/1 Loss

Ranked Retrieval

- For each query, the documents are sorted in the descending order of $P(\mathbf{R} = 1|d, q)$
- The top k documents with highest relevance probability are shown to the user

The 0/1 Loss

Set Retrieval

- If a set of documents (instead of an ordered list) is going to be returned, then a document d is relevant to a query q if and only if

$$P(\mathbf{R} = 1|d, q) > P(\mathbf{R} = 0|d, q)$$

To Include Retrieval Costs

- Let C_0 be the cost of retrieval of a non-relevant document
- Let C_1 be the cost of not retrieving a relevant document
- The cost of retrieval of a document is:

$$C_0 P(\mathbf{R} = 0|d) + C_1(1 - P(\mathbf{R} = 1|d))$$

- The constant C_1 can be eliminated from the cost:

$$C_0 P(\mathbf{R} = 0|d) - C_1 P(\mathbf{R} = 1|d)$$

- Among a set of documents d', the next document to retrieve is one with the minimum retrieval cost

Table of contents

1 Introduction

2 Basic Probability Theory

3 The Probability Ranking Principle

4 The Binary Independence Model (BIM)

5 Summary

The Binary Independence Model

- Making some assumptions to estimate the probability function $P(\mathbf{R}|d,q)$
- We need to know how terms contribute to the relevancy state of a document
- Documents and queries are represented as binary term incidence vectors
- It is assumed that terms occur independently in documents
- Another assumption: the relevance of documents are independent of each other

The Binary Independence Model

- We use Bayes rule to estimate the document relevance probabilities

$$P(\mathbf{R}=1|\vec{x},\vec{q}) = \frac{P(\vec{x}|\mathbf{R}=1,\vec{q})P(\mathbf{R}=1|\vec{q})}{P(\vec{x}|\vec{q})}$$

$$P(\mathbf{R}=0|\vec{x},\vec{q}) = \frac{P(\vec{x}|\mathbf{R}=0,\vec{q})P(\mathbf{R}=0|\vec{q})}{P(\vec{x}|\vec{q})}$$

The Binary Independence Model

Interpretations

- $P(\vec{x}|\mathbf{R}=1,\vec{q})$: the probability that the vector representation of a document relevant to q is $\vec{x}$
- $P(\vec{x}|\mathbf{R}=0,\vec{q})$: the probability that the vector representation of a document not relevant to q is $\vec{x}$
- $P(\mathbf{R}=1|\vec{q})$: the prior probability of retrieving a relevant document
- $P(\mathbf{R}=0|\vec{q})$: the prior probability of retrieving a non-relevant document

The Binary Independence Model

Ranking

- Instead of ranking documents based on the relevance probability, we rank them based on their odds of relevance
- We use odds because it helps eliminating the denominator $P(\vec{x}|\vec{q})$ from the calculations

$$
\begin{aligned}
O(\mathbf{R}|\vec{x},\vec{q}) &= \frac{P(\mathbf{R}=1|\vec{x},\vec{q})}{P(\mathbf{R}=0|\vec{x},\vec{q})} \\
&= \frac{P(\vec{x}|\mathbf{R}=1,\vec{q})P(\mathbf{R}=1|\vec{q})}{P(\vec{x}|\mathbf{R}=0,\vec{q})P(\mathbf{R}=0|\vec{q})} \\
&= O(\mathbf{R}|\vec{q})\frac{P(\vec{x}|\mathbf{R}=1,\vec{q})}{P(\vec{x}|\mathbf{R}=0,\vec{q})}
\end{aligned}
$$

The Binary Independence Model

Ranking

- Assuming that the occurrences of words in a document are independent of each other:

$$P(\vec{x}|\mathbf{R}=1,\vec{q}) = \prod_{t=1}^{M} P(x_t|\mathbf{R}=1,\vec{q})$$

- So, the odd of relevance is:

$$O(\mathbf{R}|\vec{x},\vec{q}) = O(\mathbf{R}|\vec{q}) \prod_{t=1}^{M} \frac{P(x_t|\mathbf{R}=1,\vec{q})}{P(x_t|\mathbf{R}=0,\vec{q})}$$

The Binary Independence Model

Ranking

- x_t is a Boolean variable that can be either 0 or 1
- Hence, the odd values can be decomposed into:

$$O(\mathbf{R}|\vec{x},\vec{q}) = O(\mathbf{R}|\vec{q}) \prod_{t:x_t=1} \frac{P(x_t=1|\mathbf{R}=1,\vec{q})}{P(x_t=1|\mathbf{R}=0,\vec{q})} \prod_{t:x_t=0} \frac{P(x_t=0|\mathbf{R}=1,\vec{q})}{P(x_t=0|\mathbf{R}=0,\vec{q})}$$

The Binary Independence Model

Ranking

- Let $p_t = P(x_t = 1|\mathbf{R} = 1, \vec{q})$ be the probability of the occurrence of the term x_t in a document relevant to q
- Let $u_t = P(x_t = 1|\mathbf{R} = 0, \vec{q})$ be the probability of the occurrence of the term x_t in a document non-relevant to q

	$\mathbf{R} = 1$	$\mathbf{R} = 0$
$x_t = 1$	p_t	u_t
$x_t = 0$	$1 - p_t$	$1 - u_t$

The Binary Independence Model

Ranking

- We assume that the terms that do not occur in the query are equally likely to be seen in both relevant and non-relevant document (if $q_t = 0$ then $u_t = p_t$)
- We only consider terms that appear in the query

$$O(\mathbf{R}|\vec{x},\vec{q}) = O(\mathbf{R}|\vec{q}) \prod_{t:x_t=q_t=1} \frac{p_t}{u_t} \prod_{t:x_t=0,q_t=1} \frac{1-p_t}{1-u_t}$$

- The left product is over query terms found in the document
- The right product is over query terms not found in the document

The Binary Independence Model

Ranking

- If we include the query terms that are found in the document to the right product

$$\prod_{t:x_t=0,q_t=1} \frac{1-p_t}{1-u_t} \rightarrow \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

- We should simultaneously divide the left product by $\frac{1-p_t}{1-u_t}$ to cancel out the effect of the above modification

$$\prod_{t:x_t=q_t=1} \frac{p_t}{u_t} \rightarrow \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

The Binary Independence Model

Ranking

- The odd of the relevant then will be

$$O(\mathbf{R}|\vec{x},\vec{q}) = O(\mathbf{R}|\vec{q}) \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

- The left product is over the query terms found in the document
- The right product is over all query terms
- The odd term $O(\mathbf{R}|\vec{q})$ and the right product are constant for a query
- They have the same value for all documents when processing a particular query

The Binary Independence Model

Ranking

- Documents can be ranked for their relevance to a query based on the product:

$$\prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

- Equivalently, we can rank the documents by their retrieval status value (RSV)

$$\text{RSV}_d = \log \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=q_t=1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

The Binary Independence Model

Ranking

- The logarithm can be decomposed into

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

- Its first component is the logarithm of the odd of the term appearing in a relevant document
- Its second component is the logarithm of the odd of the term appearing in a non-relevant document

The Binary Independence Model

Ranking

- c_t is a weight for the term t
- c_t is zero if the term t is equally likely to appear in both relevant and non-relevant documents
- A positive value of c_t indicates that the term is more likely to appear in relevant documents
- A negative value of c_t indicates that the term is more likely to appear in non-relevant documents
- The document score is then the sum of c_t for all document terms matching the query terms

The Binary Independence Model

Probability Estimation

- For a query q, if we have S relevant documents out of N documents with the following distribution on the terms

	$\mathbf{R}=1$	$\mathbf{R}=0$	total
$x_t=1$	s	$\mathrm{df}_t - s$	df_t
$x_t=0$	$S-s$	$(N-\mathrm{df}_t)-(S-s)$	$N-\mathrm{df}_t$
total	S	$N-S$	N

- The probability of seeing the term x_t in a relevant document is

$$p_t = p(x_t = 1|\mathbf{R}=1,\vec{q}) = \frac{s}{S}$$

- The probability of seeing the term x_t in a non-relevant document is

$$u_t = p(x_t = 1|\mathbf{R}=0,\vec{q}) = \frac{\mathrm{df}_t - s}{N-S}$$

The Binary Independence Model

Probability Estimation

- The term weight c_t is:

$$
\begin{aligned}
c_t &= \log \frac{p_t(1-u_t)}{u_t(1-p_t)} \\
&= \log \frac{s((N-S)-(\mathrm{df}_t-s))}{(\mathrm{df}_t-s)(S-s)} \\
&= \log \frac{s/(S-s)}{(\mathrm{df}_t-s)/((N-\mathrm{df}_t)-(S-s))}
\end{aligned}
$$

The Binary Independence Model

Probability Estimation

- We add a smoothing value to the four terms of c_t to avoid the possibility of zeros

$$\hat{c}_t = \log \frac{(s + 0.5)/(S - s + 0.5)}{(\mathrm{df}_t - s + 0.5)/((N - \mathrm{df}_t) - (S - s) + 0.5)}$$

The Binary Independence Model

Probability Estimation

- The number of relevant documents (S) are often very smaller than the number of non-relevant documents ($N - S$)
- We can estimate the probability u_t from the statistics of the entire collection

$$u_t = \frac{\mathrm{df}_t}{N}$$

- The inverse of the odd of u_t can then be approximated by the idf of the term t

$$\log \frac{1 - u_t}{u_t} = \log \frac{N - \mathrm{df}_t}{\mathrm{df}_t} \approx \log \frac{N}{\mathbf{df}_t}$$

- This cannot be easily extended to relevant document

The Binary Independence Model

Probabilistic Estimates in Relevance Feedback

The probabilities p_t and u_t can be estimated in an iterative process of pseudo relevance feedback

1. Guess initial estimates of p_t and u_t (e.g., $p_t = 0.5$).
2. Retrieve a set of candidate documents based on the current estimates of p_t and u_t.
3. Ask the user to judge the retrieved documents
4. Re-estimate p_t and u_t based on the user judgements.
5. Repeat the process from Step 2 until the user is satisfied.

Table of contents

1 Introduction

2 Basic Probability Theory

3 The Probability Ranking Principle

4 The Binary Independence Model (BIM)

5 Summary

Summary

- How to rank documents based on their probability of relevance
- The binary Independence model for relevance probability
- How to estimate the probabilities
- The probability estimation in an iterative relevance feedback procedure

References I

Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich.
Introduction to Information Retrieval.
Cambridge University Press, 2008.