Information Retrieval (5LN712)
Probabilistic Information Retrieval

Ali Basirat

Department of Computer and Information Science (IDA)
Linköping University
ali.basirat@liu.se

Department of Linguistics and Philology
Uppsala University
ali.basirat@lingfil.uu.se

April 20, 2021
## Table of contents

1. **Introduction**

2. **Basic Probability Theory**

3. **The Probability Ranking Principle**

4. **The Binary Independence Model (BIM)**

5. **Summary**
• 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 provides a principled foundation for reasoning under uncertainty
1 Introduction

2 Basic Probability Theory

3 The Probability Ranking Principle

4 The Binary Independence Model (BIM)

5 Summary
• 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
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 ($A : A = 1$)
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?
- $P(A) = P(A = 1) = \frac{1}{6}$
Example

- We roll a dice. Let $\mathbf{A}$ 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?
- $P(A) = P(A = 1) = \frac{1}{6}$
- 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:

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

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

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

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

$$P(R = 1|d_1, q) > P(R = 1|d_2, q)$$
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 no decision making cost.
• For each query, the documents are sorted in the descending order of $P(R = 1|d, q)$

• The top $k$ documents with highest relevance probability are shown to the user
• 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

$$P(R = 1|d, q) > P(R = 0|d, q)$$
To Include Retrieval Costs

- Let $C_0$ be the cost of retrieval of a non-relevant document (false-positive)
- Let $C_1$ be the cost of not retrieving a relevant document (false-negative)
- The cost of retrieval of a document is:

$$
C_0 P(R = 0|d) + C_1 (1 - P(R = 1|d))
$$

- The constant $C_1$ can be eliminated from the cost:

$$
C_0 P(R = 0|d) - C_1 P(R = 1|d)
$$

- 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 $P(R|d, q)$
- 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

\[
P(R = 1 | \tilde{x}, \tilde{q}) = \frac{P(\tilde{x} | R = 1, \tilde{q}) P(R = 1 | \tilde{q})}{P(\tilde{x} | \tilde{q})}
\]

\[
P(R = 0 | \tilde{x}, \tilde{q}) = \frac{P(\tilde{x} | R = 0, \tilde{q}) P(R = 0 | \tilde{q})}{P(\tilde{x} | \tilde{q})}
\]
The Binary Independence Model

Interpretations

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

Introductory

Basic Probability Theory

The Probability Ranking Principle

The Binary Independence Model (BIM)

Summary

• 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 $P(\bar{x}|\bar{q})$ from the calculations.

$$O(R|\bar{x}, \bar{q}) = \frac{P(R = 1|x, q)}{P(R = 0|x, q)} = \frac{P(\bar{x}|R = 1, \bar{q}) P(R = 1|\bar{q})}{P(\bar{x}|R = 0, \bar{q}) P(R = 0|\bar{q})} = O(R|\bar{q}) \frac{P(\bar{x}|R = 1, \bar{q})}{P(\bar{x}|R = 0, \bar{q})}$$
• Assuming that the occurrences of words in a document are independent of each other:

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

• So, the odd of relevance is:

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

\[
O(R|\vec{x}, \vec{q}) = O(R|\vec{q}) \prod_{t:x_t=1} \frac{P(x_t = 1|R = 1, \vec{q})}{P(x_t = 1|R = 0, \vec{q})} \prod_{t:x_t=0} \frac{P(x_t = 0|R = 1, \vec{q})}{P(x_t = 0|R = 0, \vec{q})}
\]
The Binary Independence Model

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

<table>
<thead>
<tr>
<th></th>
<th>$R = 1$</th>
<th>$R = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_t = 1$</td>
<td>$p_t$</td>
<td>$u_t$</td>
</tr>
<tr>
<td>$x_t = 0$</td>
<td>$1 - p_t$</td>
<td>$1 - u_t$</td>
</tr>
</tbody>
</table>
The Binary Independence Model

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

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

The left product is over query terms found in the document.

The right product is over query terms not found in the document.
• 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

The odd of the relevant then will be:

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

- The left product is over the query terms found in the document.
- The right product is over all query terms.
- The odd term \( O(R|\vec{q}) \) 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

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)

$$RSV_d = \log \prod_{t: x_t = q_t = 1} \frac{p_t (1 - u_t)}{u_t (1 - p_t)} = \sum_{t: x_t = q_t = 1} \log \frac{p_t (1 - u_t)}{u_t (1 - p_t)}$$
The 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 odds of the term appearing in a relevant document
- Its second component is the logarithm of the odds of the term appearing in a non-relevant document
The Binary Independence Model

• $c_t$ 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:

\[
\begin{array}{c|cc|c}
& R = 1 & R = 0 & \text{total} \\
\hline
x_t = 1 & s & df_t - s & df_t \\
x_t = 0 & S - s & (N - df_t) - (S - s) & N - df_t \\
\hline
\text{total} & S & N - S & N \\
\end{array}
\]

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

\[ p_t = p(x_t = 1| R = 1, \bar{q}) = \frac{S}{S} \]

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

\[ u_t = p(x_t = 1| R = 0, \bar{q}) = \frac{df_t - s}{N - S} \]
• The term weight $c_t$ is:

\[
    c_t = \log \frac{p_t(1 - u_t)}{u_t(1 - p_t)}
    = \log \frac{s((N - S) - (df_t - s))}{(df_t - s)(S - s)}
    = \log \frac{s/(S - s)}{(df_t - s)/((N - df_t) - (S - s))}
\]
We add a smoothing value to the four terms of $c_t$ to avoid the possibility of zeros

$$
\hat{c}_t = \log \frac{(s + 0.5)/(S - s + 0.5)}{(df_t - s + 0.5)/((N - df_t) - (S - s) + 0.5)}
$$
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
  \[ u_t = \frac{df_t}{N} \]
- The inverse of the odds of $u_t$ can then be approximated by the idf of the term $t$
  \[ \log \frac{1 - u_t}{u_t} = \log \frac{N - df_t}{df_t} \approx \log \frac{N}{df_t} \]
- This cannot be easily extended to relevant document
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
Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich.

*Introduction to Information Retrieval.*