Assignment-2

Himanshu S. Bhatt and Samarth Bharadwaj

January 30, 2012

Question-2: Prove the Bayes optimality

For a given document d, if $P(R=1|d) > P(R=0|d)$, we decide that the document is relevant. The probability of error whenever we make a incorrect decision is given as:

$$P(error|d) = \begin{cases} P(R=1|d) & if \ \ d\, is\, irrelevant \\ P(R=0|d) & otherwise \end{cases} \tag{1}$$

Therefore, for a given document d, we can minimize the probability of error by deciding the document to be relevant if $P(R=1|d) > P(R=0|d)$ and irrelevant otherwise. However, it requires that all probabilities are known correctly. Therefore, we minimize the average probability of error given as:

$$P(error) = \sum_{d} P(error|d)P(d) \tag{2}$$

If for every document d, we ensure that the $P(errors|d)$ is as small as possible, then the overall summation will be small. Thus, is justifies the Bayes decision rule for minimizing the probability of error.

$$D\ is\ relevant\ iff, P(R=1|d) > P(R=0|d) \tag{3}$$

(a) Zero-one loss function The loss function $\lambda(\alpha_i|R_j)$ is the loss incurred for deciding a document d as α_i when the actual state of the document is R_j. Therefore, we define conditional risk $R(\alpha_1|R_j)$ as:

$$R(\alpha_i|R_j) = \sum_{j=1}^{2} \lambda(\alpha_i|R_j)P(R_j|d) \tag{4}$$

For a given document d, we can minimize the expected loss by selection the action that minimizes the conditional risk. The fundamental rule is to decide d as relevant if $R(\alpha_1|d) > R(\alpha_2|d)$.

The zero-one loss function is given as:

$$\lambda(\alpha_i|R_j) = \begin{cases} 0 & if \ \ i=j \\ 1 & otherwise \end{cases} \tag{5}$$

This loss function assigns no loss to a correct decision and assigns a unit loss to any error. Therefore, the conditional risk can be rewritten as:

$$R(\alpha_i|d) = \sum_{j=1}^{2} \lambda(\alpha_1|R_j)P(R_j|d) \tag{6}$$

$$= \sum_{j \neq i} P(R_j|d) \tag{7}$$

$$= 1 - P(R_i|d) \tag{8}$$

where $P(R_i|d)$ is the conditional probability that R_i is the correct state of the document. Thus, to minimize the average probability of error, we should maximize the posterior probability $P(R_i|d)$. By minimizing the overall Bayes risk best performance is achieved given as:

$$d\, is\, relevant\, iff, P(R = 1|d) > P(R = 0|d) \tag{9}$$

(b) With cost functions The conditional risk in this case can be written as:

$$R(\alpha_i|d) = \sum_{j=1}^{2} \lambda(\alpha_i|R_j)P(R_j|d) \tag{10}$$

$$R(\alpha_1|d) = \lambda(\alpha_1|R_1)P(R_1|d) + \lambda(\alpha_1|R_2)P(R_2|d) \tag{11}$$

$$R(\alpha_1|d) = CN\ P(R_1|d) + CR\ P(R_2|d) \tag{12}$$

CN is the cost associated with retrieving a relevant document and CR is the cost associated with not retrieving a relevant document. R_1 represents that the document is relevant and R_2 represents that the document is irrelevant. Thus, to minimize the average probability of error, the overall Bayes risk is minimized to achieve the best performance:

$$d\, is\, relevant\, iff, CN\ P(R_1|d) + CR\ P(R_2|d) < CN\ P(R_1|d') + CR\ P(R_2|d') \tag{13}$$

where d’ is the set of documents not yet retrieved.

Question3

Quiz-2

Answer-1

- Stemming Errors: 1) University and universal (Stem-Universe), 2) Organism / organization (Stem-Organ)

- Tokenization: 1) Academic year 2011-12 (tokenized string : academic year 2011) , 2) Meeting is scheduled at 8 AM (Tokenized string: meeting scheduled).

Answer-2 The pseudo code for negation is given below. For a negation operation on a keyword, all documents in which the keyword is present are retrieved.

Algorithm 1 Pseudo code

```
Negative(kw) kw is the given keyword.
For $i$= 1 to N (number of documents)
Do if $D_i$==1
Then $D_i$=0
Else if $D_i$==0
Then $D_i$=1
End if
End For
Output: Negation.
```

Answer-3 The IDF of a term that occurs in all documents is zero. Expression

$$IDF_t = log\frac{N}{df_t} \tag{14}$$

where N is the number of documents in a collection and df is the document frequency of the term t (i.e. the number of documents in which the term occurred).

Answer-4 Yes, Tf-IDF weight for a term can exceed 1. Higher weight represents more discriminative terms.

Answer-5 Table 1 gives the inverted index that can support positional restrictions. The final sorted and merged sorted index is presented.

Often, the co-occurrence of two terms of a phrase is determined with a proximity threshold k. In which case the complexity of search becomes $t \times k$. Where t is the number of terms.

Quiz-3

Answer-1 We use harmonic mean in F-measure because it is always less than or equal to the arithmetic mean and the geometric mean. When the values of two numbers differ greatly, the harmonic mean is closer to their minimum than to their arithmetic mean.

Answer-2

$$Precision = \frac{Truepositives}{Trurepositives + Falsepositives} \tag{15}$$

$$Recall = \frac{Truepositives}{Trurepositives + Falsenegatives} \tag{16}$$

Term	**Posting list**
Am	D_2(8,11), D_3(5)
Be	D_1(5,7), D_2(2,6), D_3(7,9), D_4(9,12)
Da	D_4(4,5,6)
Do	D_1(1,10), D_3(6,8,10), D_4(1,2,3)
I	D_2(7,10), D_3(1,4)
Is	D_1(3,8)
It	D_4(8,11)
Let	D_4(7,10)
Not	D_2(4)
Or	D_2(3)
Think	D_3(2)
Therefore	D_3(3)
To	D_1(1,4,6,9), D_2(1,5)
What	D_2(9)

Table 1: Sorted and merged inverted index.

True positives has to bee maximized for both precision and recall. However, precision can be optimized by reducing the number of false positives whereas, Recall is optimized by reducing the number of false negatives. Since, it is very difficult to simultaneously optimize both false positives and false negatives. Therefore, the claim by Mr. Big Brain is not possible. Generally, a system is optimized for reducing the number of false positives (optimizing precision) or false negatives (optimizing recall).

Answer-3 While using a search engine, it is not possible to find if you have found all relevant documents or not. However based on the query, if the user gets the information what he was looking for in top few retrieved documents. User is quite satisfied.

Answer-4 Assumption, we assume that the document is relevant when both the users marked it as relevant. Therefore, total number of relevant documents is four.

Given the retrieval order $S_q = \{4, 5, 6, 7, 8\}$, the mean precision is calculated as: $MAP = (1 + 0.5 + 0.33 + 0.25 + 0.2)/4 = 0.57$

The optimal retrieval order can be $S_q = \{3, 4, 11, 12, 5\}$. The mean average precision for this sequence of retrieved documents is calculated as: $MAP = (1 + 1 + 1 + 1 + 0.8)/4 = 1.2$

Answer-5 The reformulated query is not the same as the original one because with each iterations (relevance feedback), we change the weights for query term to move it closer to the relevant documents and farther from irrelevant documents. However, when the input query is itself the optimal query, in such conditions reformulated query may remain same.