CS630 Lecture 4: Probabilistic Retrieval

Lecture by Lillian Lee
Scribed by Randy Au, Nick Gerner, Blazej Kot

February 7, 2006

1 Probabilistic Retrieval

The Vector Space Model is essentially an ad hoc approach to information retrieval. If we wanted firmer theoretical ground, we will have to take a different approach using a priori assumptions. One such attempt is derived here.

1.1 Notation

Capital letters represent random variables. Lower case letters represent the values taken by the random variables.

2 The Derivation

Assumption 1: the query, q, is fixed to keep the derivation simpler.

Our ultimate goal is to find a way to rank documents according to the probability of their relevance to the query, as opposed to the VSM model where relevance is simply how "close" q is to d in a vector space.

$$P(relevant|doc)$$

We therefore want to compute:

$$P(R = r|D = d)$$

where: R is a random variable of relevance, and comes from a set such as: {y, n}, {0,0.1,. . . ,1}, and D is a random variable for documents, which might range over the set of documents in the corpus, or all possible documents, real or imagined.

Note: Throughout the derivation $P(D = d) > 0$ holds true. Otherwise we would condition on an event of probability zero, and the universe explodes.

Because the user is most interested in relevant documents, we want to compute:

$$P(R = y|D = d)$$

Issue #1: For a *given* document, and a *given* query, the probability of relevance of that document should be a single value in {0,1}. There's no randomness in the situation. We need a way to justify injecting random variables into the equations.

Possible Solution #1: We could say that, for a given user, on a given day, they might judge the document relevant, while on another day, the same document for the same query is irrelevant, and this process is random. *OR*, over all users, some percentage will find the document relevant while some won't.

The problem with this is that this puts human preferences into the picture.

Possible Solution #2: We can represent uncertainty as a function of document representation. If the user looks at collections of documents, binned in some way, some fraction of the bin would be relevant, and this would be a random probability.

For this derivation, we will proceed using the second solution to avoid the need to model humans preferences.

Let $\mathbf{A}$ be a vector of attribute variables:

$$\mathbf{A} = (\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_m})^{\mathbf{T}}$$

Then, for each document d, there is an associated attribute vector $\mathbf{a}(d)$:

$$\mathbf{a}(d) = (a_1(d), a_2(d), \ldots, a_m(d))^T$$

For example, $a_j(d)$ may be the number of times that the term $v^{(j)}$ appears in d. Attributes can be as complex as desired, *i.e.* $a_{m+1} = length(d)$.

Assumption #2: In this derivation, we assume that $\mathbf{A}_j$ corresponds to the appearance of term $v^{(j)}$ in a document.

So, we want to compute

$$P(R = y | \mathbf{A} = \mathbf{a}(d))$$

We still have no information about relevancy. How can we proceed further, pushing the question of relevancy as late in the process as we can?

Assumption #3: Seeing a given attribute vector, $\mathbf{a}(d)$, has a very low probability. (We can increase the probability somewhat by keeping $\mathbf{a}(d)$ short, but we would lose information.) In any event, it seems reasonable to assume that

$$P(R = y) > P(\mathbf{A} = \mathbf{a}(d))$$

This assumption suggests that we apply a *"Bayes Flip"*, by applying Bayes Rule, to flip the conditional.

$$P(R = y|\mathbf{A} = \mathbf{a}(d)) = \frac{P(\mathbf{A} = \mathbf{a}(d)|R = y)P(R = y)}{P(\mathbf{A} = \mathbf{a}(d))}$$

With respect to computing a score for a document d, $P(R = y)$, being a constant, falls out.

$$\stackrel{rank}{=} \frac{P(\mathbf{A} = \mathbf{a}(d)|R = y)}{P(\mathbf{A} = \mathbf{a}(d))}$$

Remember that $P(\mathbf{A} = \mathbf{a}(d))$ is a tiny value.

NOTE: In the classic presentation of this derivation, an odds ratio is used, and so the denominator term $P(\mathbf{A} = \mathbf{a}(d))$ is eliminated:

$$\frac{P(\mathbf{A} = \mathbf{a}(d)|R = y)}{P(\mathbf{A} = \mathbf{a}(d)|R = n)}$$

In estimation, it is then assumed that for $P(\mathbf{A} = \mathbf{a}(d)|R = n)$, most documents aren't relevant anyway, and so

$$P(\mathbf{A} = \mathbf{a}(d)|R = n) = P(\mathbf{A} = \mathbf{a}(d))$$

which is what we have before.

Assumption #4: The attributes are conditionally independent, **and** terms are independent.

Issue #2: Cooper in "Some inconsistencies and misnomers in probabilistic information retrieval" in SIGIR '91, finds counterexamples that yield logical errors when this assumption is used. The formula we use is a version of a formulation that he was unable to find a counter example for.

$$= \alpha \frac{\prod_j P(\mathbf{A}_j = a_j(d)|R = y)}{\prod_j P(\mathbf{A}_j = a_j(d))}$$

α falls out since it is a constant and we are ranking.

Notice: The product ranging over all term-attributes within a document can allow documents corresponding to unobserved attribute vectors to still have non-zero probabilities.

We have just about done all that we can to modify the equation while putting the relevancy question off. We are now out of tricks and so must deal with it. The simplest step would be to factor whether terms in the query do or do not appear in a document, given that the query is one clue regarding relevance.

$$= \overbrace{\prod_{j:q_j=1} \frac{P(\mathbf{A}_j = a_j(d)|R = y)}{P(\mathbf{A}_j = a_j(d))}}^{\text{when term is in query}} \cdot \underbrace{\prod_{j:q_j=0} \frac{P(\mathbf{A}_j = a_j(d)|R = y)}{P(\mathbf{A}_j = a_j(d))}}_{\text{when term is not on query}}$$

Notation:

$$q_j = \begin{cases} 1 & \text{if } v^{(j)} \text{ is in q,} \\ 0 & \text{otherwise} \end{cases}$$

Assumption #5: Non-query term-attributes have the same distribution over relevant documents as over all documents.

Issue #3: This would probably be true for non-query terms that aren't related to the query terms, but is not quite so true for non-query terms that are related to the query terms.

This assumption allows us to state for terms not in the query:

$$P(\mathbf{A}_j = a_j(d)|R = y) = P(\mathbf{A}_j = a_j(d))$$

Which allows us to cancel out the non-query terms in our equation, leaving us with:

$$\overset{assumption}{=} \prod_{j:q_j=1} \frac{P(\mathbf{A}_j = a_j(d)|R = y)}{P(\mathbf{A}_j = a_j(d))}$$

And that's as far as we got in this lecture!

(To be continued. . .)

3 Questions

One assumption we made during the first part of our derivation of the probability of relevance was that the distribution of non-query attributes have the same distribution over relevant documents as over non-relevant documents. Let's examine this assumption.

In the following, for simplicity assume that $\mathbf{A}_j$ corresponds to term presence/absence, rather than term counts.

3.1 Question 1

What does this assumption mean, especially with respect to our derivation? I.e., why would we make this assumption and how does it simplify our derivation? In addition to any mathematical formulae, give a short text answer explaining each probability term and how they interact.

3.2 Question 2

Given the following vocabulary of nouns and verbs (and their derivatives, e.g. since dog is in the language, then so would dogs, or chase gives us chases, chased, etc.):

$V = (cat, dog, mouse, bone, seeds, food, toy, chase, hunt, play, forage, eat, sleep, lay, run, bury)$

Suppose our attributes or terms are also given by this vocabulary (derivatives of the terms will be considered as being named by the same attribute)

Consider the following query, corpus and relevance tags. We have included some common terms. But these will not be included in our analysis (even though these do affect semantics)

corpus =

d(1) = "dogs chase cats"
d(2) = "mice forage for seeds"
d(3) = "cats eat mice"
d(4) = "dogs play with bones"
d(5) = "cats play with toys"
d(6) = "cats play with mice"
d(7) = "cats sleep after eating"
d(8) = "dogs run with bones"
d(9) = "dogs bury bones"
d(10) = "cats play with cats"

query = "what do cats play with?"

Then, documents 5,6 and 10 are relevant. Does this corpus and query meet our assumption? Show that

$$P(\mathbf{A}_j = a_j(d)|R = y) = P(\mathbf{A}_j = a_j(d))$$

for the necessary $\mathbf{A}_j$'s, or show that

$$P(\mathbf{A}_j = a_j(d)|R = y) \neq P(\mathbf{A}_j = a_j(d))$$

for at least one necessary $\mathbf{A}_j$ (but several would be more convincing).

3.3 Question 3

Now consider the stability of a system in which the assumption is met. Suppose you were presented with a corpus, a query, and relevance tags such that our assumption is met. (Assume that some of the relevant documents contain some non-query terms.)

3.3.1 Question 3 Part 1

Can you change the query (and corresponding relevance tags) to make the assumption invalid? How or why not?

3.3.2 Question 3 Part 2

Can you add one or more documents (keeping the language, query and relevant documents fixed) to make the assumption invalid? How or why not?

3.3.3 Question 3 Part 3

Can you add documents (keeping the language, query and ratio of relevant documents to all documents fixed) to make the assumption invalid? How or why not?

4 Answers

4.1 Answer 1

The assumption says:

$$\frac{P(\mathbf{A}_j = a | R = y)}{P(\mathbf{A}_j = a)} = 1 \; if \; q_j = 0$$

$P(\mathbf{A}_j = a | R = y) \; = \; probability \; of \; seeing \; term \; v^{(j)} \; in \; a \; document \; given \; that \; it \; is \; relevant$

$$P(\mathbf{A}_j = a) = \; probability \; of \; seeing \; term \; v^{(j)} \; in \; any \; document$$

If the distribution of $v^{(j)}$ over relevant documents equals the distribution of $v^{(j)}$ over all documents, then these probabilities will be equal.

4.2 Answer 2

The non-query terms are dog, mouse, bone, seeds, food, toy, chase, hunt, play, forage, eat, sleep, lay, run, bury

dog isn't in any of the relevant documents, but occurs in many documents

mice occurs in 1/3 of the relevant documents but only 3/10 of all documents (even though this is close, the assumption is that they're perfectly equal)

4.3 Answer 3

4.3.1 Answer 3 Part 1

Yes, the distribution of terms over relevant/all docs could be completely different since the condition is only required to hold on non-query terms, and presumably does not hold for

query terms. Thus, removing a (redundant) term from the query would certainly lead to a violation of the assumption.

4.3.2 Answer 3 part 2

Yes, suppose we add a bunch of new documents each consisting of just one non-query term which occurs in a relevant document, assuming that the non-query term does not appear in all the relevant documents.

4.3.3 Answer 3 part 3

Yes. Given an original relevance ratio of 1/10, add ten new documents, one relevant, nine not, where the relevant document contains a non-query term that didn't occur in any relevant documents previously. Also, some new non-relevant document must not contain the term.