


CS 598cxz (/course/598f16) Fall 2016

Assignment #1: Background, Probability and Statistics

 **Notice:** This assignment is due **Saturday, September 3rd at 11:59pm**.

Please submit your solutions via Compass (<https://compass2g.illinois.edu>). You should submit your assignment as a **PDF**.

1. Information Retrieval Background [25 pts]

 **Info:** The following questions should be review from material covered in the CS 410 prerequisite (or equivalent). If you find yourself having difficulty with these concepts (or have never seen them before) this is OK, but you will have to teach yourself these concepts relatively quickly. This part of the assignment is designed to help you with that.

We encourage you to use the following textbook for reference: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining (<http://dl.acm.org.proxy2.library.illinois.edu/citation.cfm?id=2915031>). The provided link should give you access to the full text after logging in with your NetID and AD password.

(In fact, most of these questions are lifted from the exercises presented in that book!)

- a. [5 pts] Why might using raw term frequency counts with a dot product similarity not produce the best ranking results?
- b. [5 pts] Let d be a document in a corpus. Suppose we add another copy of d to the collection. How does this affect the IDF values for all words in the corpus?
- c. [10 pts] Suppose there are 16 total relevant documents in a collection. Consider the following result, where $+$ indicates a relevant document and $-$ indicates an non-relevant document.

$$\{+, +, -, +, +, -, -, +, -, -\}$$

Calculate the following evaluation measures for this ranked list.

- i. Precision
- ii. Recall
- iii. F_1 score
- iv. Average precision

d. [5 pts] Using the same setup as above, assume that the “gain” of a relevant document is 1 and the “gain” of a non-relevant document is 0. Calculate the following:

- i. Cumulative Gain at 7 documents
- ii. Normalized Discounted Cumulative Gain at 7 documents (use \log_2 for the discounting function)

2. Probabilistic Reasoning and Bayes Rule [25 pts]

Consider the problem of detecting email messages that may carry a virus. This problem can be modeled probabilistically by treating each email message as representing an observation of values of the following 4 random variables:

1. A : whether the message has an attachment (1 for yes);
2. K : whether the sender is unknown to the receiver (1 for yes);
3. L : whether the message is not longer than 10 words (1 for yes); and
4. V : whether the message carries a virus (1 for yes).

Given a message, we can observe the values of A , L , and K , and we want to infer its value of V . In terms of probabilistic reasoning, we are interested in evaluating the conditional probability $p(V \mid A, L, K)$, and we would say that the message carries a virus if $p(V = 1 \mid A, L, K) > p(V = 0 \mid A, L, K)$.

We make a further assumption that $p(A, L, K \mid V) = p(A \mid V)p(L \mid V)p(K \mid V)$ for $V = 0$ and $V = 1$, i.e., given the status whether a message carries a virus, the values of A , K , and L are independent.

a. [3 pts] Suppose we observe 12 samples (Table 1):

sample #	A	K	L	V
0	1	1	1	1
1	1	1	0	1
2	1	1	0	1
3	1	0	0	0
4	1	1	1	0
5	0	1	0	1
6	0	1	1	1
7	0	0	0	0
8	0	1	0	0

9	0	1	1	0
10	0	0	0	0
11	1	0	0	1

Fill in the following table (Table 2) with conditional probabilities using *only* the information present in the above 12 samples.

V	$p(A = 1 \mid V)$	$p(K = 1 \mid V)$	$P(L = 1 \mid V)$	prior $p(V)$
0				1/2
1	4/6			

- b. **[5 pts]** With the independence assumption, use Bayes' rule and probabilities you just computed in part A to compute the probability that a message M with $A = 0$, $K = 1$, and $L = 0$ carries a virus. i.e., compute $p(V = 1 \mid A = 0, K = 1, L = 0)$ and $p(V = 0 \mid A = 0, K = 1, L = 0)$. Would we conclude that message M carries a virus?
- c. **[3 pts]** Now, compute $p(V = 1 \mid A = 0, K = 1, L = 0)$ and $p(V = 0 \mid A = 0, K = 1, L = 0)$ directly from the 12 examples in Table 1, just like what you did in problem A. Do you get the same value as in problem B? Why?
- d. **[2 pts]** Now, ignore Table 1, and consider any possibilities you can fill in Table 2. Are there any constraints on these values that we must respect when assigning these values? In other words, can we fill in Table 2 with 8 arbitrary values between 0 and 1?
- e. **[2 pts]** Can you change your conclusion of problem B (i.e., whether message M carries a virus) by only changing the value A (i.e., if the message has an attachment) in 1 example of Table 1?
- f. **[5 pts]** Note that the conditional independence assumption $p(A, L, K \mid V) = p(A \mid V)p(L \mid V)p(K \mid V)$ helps simplify the computation of $p(A, L, K \mid V)$. In particular, with this assumption, we can compute $p(A, L, K \mid V)$ based on $p(A \mid V)$, $p(L \mid V)$, and $p(K \mid V)$. If we were to specify the values for $p(A, L, K \mid V)$ directly, what is the minimum number of probability values that we would have to specify in order to fully characterize the conditional probability distribution $p(A, L, K \mid V)$? Why? Note that all the probability values of a distribution must sum to 1.
- g. **[5 pts]** Explain why the independence assumption $p(A, L, K \mid V) = p(A \mid V)p(L \mid V)p(K \mid V)$ does not necessarily hold in reality.

3. Maximum Likelihood Estimation [50 pts]

A Poisson distribution is often used to model the word frequency. Specifically, the number of occurrences of a word in a document with fixed length can be assumed to follow a Poisson distribution given by

$$p(X = x) = \frac{u^x e^{-u}}{x!}, u > 0$$

where X is the random variable representing the number of times we have seen a specific word w in a document, and u is the parameter of the Poisson distribution (which happens to be its mean). Now, suppose we observe a sample of counts of a word w , $\{x_1, \dots, x_N\}$, from N documents with the same length (x_i is the counts of w in one document). We want to estimate the parameter u of the Poisson distribution for word w . One commonly used method is the maximum likelihood method, in which we choose a value for u that maximizes the likelihood of our data $\{x_1, \dots, x_N\}$, i.e.,

$$\hat{u} = \arg \max_u p(x_1, \dots, x_N | u), u > 0$$

a. **[35 pts]** Derive a closed form formula for this estimate.

(Hint: Write down the log likelihood of $\{x_1, \dots, x_N\}$, which would be a function of u . Set the derivative of this function w.r.t. u to zero, and solve the equation for u .)

b. **[15 pts]** Now suppose u has a prior exponential distribution

$$p(u) = \lambda e^{-\lambda u}, u > 0$$

where λ is a given parameter. Derive a closed form for the maximum a posteriori estimate, i.e.,

$$\hat{u} = \arg \max_u p(x_1, \dots, x_N | u)p(u), u > 0$$

(Hint: refer to [this Wikipedia page](http://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation#Computation)

(http://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation#Computation) and look for the Example section.)