COMP 5790/ 6790/ 6796 Special Topics: Information Retrieval

Instructor: Shubhra (“Santu”) Karmaker

Assignment #1: Probabilistic Reasoning and Entropy. [100 points]

⚠ **Notice:** This assignment is due **Wednesday, January 27, 2021 at 11:59pm**.

Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a **typeset PDF**. Please do not include scanned or photographed equations as they are difficult for us to grade.

1. Revisiting Probability [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	2/6	3/6	2/6	1/2
1	4/6	5/6	2/6	1/2

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?

$p(V = 1 \mid A = 0, K = 1, L = 0) = p(A = 0, K = 1, L = 0 \mid V = 1) * p(V = 1) / p(A = 0, K = 1, L = 0)$

$p(V = 1) = ½$

$p(A = 0, K = 1, L = 0) = p(A = 0) * p(K = 1) * P(L = 0) = ½ * 2/3 * 2/3 = 2/9$

$p(A = 0, K = 1, L = 0 \mid V) = p(A = 0 \mid V) * p(K = 1 \mid V) * P(L = 0 \mid V)$

$p(A = 0, K = 1, L = 0 \mid V) = 1/3 * 5/6 * 4/6 = 20/108$

$p(V = 1 \mid A = 0, K = 1, L = 0) = 5/27 * ½ / 2/9 = 5/12$

$p(V = 0 \mid A = 0, K = 1, L = 0) = 7/12$

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?

$p(V = 1 \mid A = 0, K = 1, L = 0) = ½$

$p(V = 0 \mid A = 0, K = 1, L = 0) = ½$

No, because the data does not represent a uniform distribution as described in the problem specifications.

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?
The only restriction is with P(V). P(V) must add up to 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?
No, $p(V = 1 | A = 0, K = 1, L = 0) \sim= p(V = 1 | A = 1, K = 1, L = 0)$

f. **[5 pts]** Note that the conditional independence assumption $p(A, L, K | V) = p(A | V)p(L | V)p(K | V)$ helps simplify the computation of $p(A, L, K | V)$. In particular, with this assumption, we can compute $p(A, L, K | V)$ based on $p(A | V)$, $p(L | V)$, and $p(K | V)$. If we were to specify the values for $p(A, L, K | V)$ directly, what is the minimum number of probability values that we would have to specify in order to fully characterize the conditional probability distribution $p(A, L, K | V)$? Why? Note that all the probability values of a distribution must sum to 1.

According to Bayes theory, we would have to specify p(V | A, L, K), p(A,L,K), and p(V).

g. **[5 pts]** Explain why the independence assumption $p(A, L, K | V) = p(A | V)p(L | V)p(K | V)$ does not necessarily hold in reality.
In reality, the values of a variable may have an influence on another variable. (A=1 may affect L or K)

2. Entropy [30 pts]

Consider the random experiment of picking a word from an English text article. Let W be a random variable denoting the word that we might obtain from the article. Thus W can have any value from the set of words in our vocabulary $V = \{w_1, \ldots, w_N\}$, where w_i is a unique word in the vocabulary, and we have a probability distribution over all the words, which we can denote as $\{p(W = w_i)\}$, where $p(W = w_i)$ is the probability that we would obtain word w_i. Now we can compute the entropy of such a variable, i.e., $H(W)$.

a. **[10 pts]** Suppose we have in total N unique words in our vocabulary. What is the theoretical minimum value of $H(W)$? What is the theoretical maximum value of $H(W)$?

Minimum: H(X) = 0

Maximum: H(X) = log(N)

b. **[10 pts]** Suppose we have only 6 words in the vocabulary $\{w_1, w_2, w_3, w_4, w_5, w_6\}$. Give two sample articles using this small vocabulary set for which $H(W)$ reaches the minimum value and maximum value, respectively.
V = {dog, cat, a, fish, bird, horse}
Minimum: “al;ksdjf as;ldkfdjk jksdfljd jlkfjlkfsdjlk lskdjwoij” – H(W) = 0
Maximum: “A dog item is the best and the cat is the worst” H(W) = log(6) = 2.58

c. **[10 pts]** Suppose we have two articles A_1 and A_2 for which $H(W) = 0$. Suppose we concatenate A_1 and A_2 to form a longer article A_3. What is the maximum value can $H(W)$ be for article A_3? Give an example of A_1 and an example of A_2 for which A_3 would have the maximum $H(W)$.
Max H(W) = 1
A1 = “AAA”
A2 = “BBB”
A3 = “AAABBB”

3. Maximum Likelihood Estimation [45 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, $\{x_1, \ldots, x_N\}$, 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 $\{x_1, \ldots, x_N\}$, i.e.,

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

a. **[30 pts]** Derive a closed form formula for this estimate.

$$L(u \mid x1, x2, \ldots, xn) = \prod_{i=1}^{n} P(X = x_i \mid u) = \prod_{i=1}^{n} \frac{u^{x_i} e^{-u}}{x_i!}$$

$$\ln\ L(u \mid x1, x2, \ldots, xn) = -nu + (\sum_{i=1}^{n} x_i) \ln{(u)} + ln \prod_{i=1}^{n} x_i!$$

$$\frac{\partial u}{\partial x} \ln L(u \mid x1, x2, \ldots, xn) = \frac{\sum_{i=1}^{n} x_i}{n}$$

(Hint: Write down the log likelihood of $\{x_1, \ldots, x_N\}$, 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.,

$$\hat{u} = \arg\max_{u} p(x_1, \ldots, x_N \mid u)p(u), u > 0$$

(Hint: refer to this Wikipedia page and look for the Example section.)

$$L(\lambda) = \sum_{i=1}^{n} \log \lambda - \lambda u_i = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial \lambda}{\partial x} L(\lambda) = \frac{n}{\lambda} - \sum_{i=0}^{n} x_i$$