COMP 5790/ 6790/ 6796

Special Topics: Information Retrieval

Instructor: Shubhra ("Santu") Karmaker

Assignment #1: Basic Concepts of Information Theory. [200 points]

Notice: This assignment is due **Friday, January 31, 2020 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. Probabilistic Reasoning [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.

2. 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, $\{x_1, \ldots, x_N\}$, 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. **[35 pts]** Derive a closed form formula for this estimate.

(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.)

3. 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 $H(W)$?

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.

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 $H(W)$.

4. Conditional Entropy and Mutual Information [20 pts]

a. **[10 pts]** What is the value of the conditional entropy $H(X \mid X)$?
b. **[10 pts]** What is the value of mutual information $I(X;Y)$ if X and Y are independent? Why?

5. Mutual Information of Words [55 pts]

Mutual information can be used to measure the correlation of two words. Suppose we have a collection of N documents. For a word A in the collection, we use $p(X_A)$, where $X_A \in \{0, 1\}$, to represent the probability that A occurs ($X_A = 1$) in one document or not ($X_A = 0$). If word A appears in N_A documents, then $p(X_A = 1) = \frac{N_A}{N}$ and $p(X_A = 0) = \frac{N - N_A}{N}$. Similarly, we can define the probability $p(X_B)$ for another word B. We also define the joint probability of word A and B as follows:

- $p(X_A = 1, X_B = 1)$: the probability of word A and word B co-occurring in one document. If there are N_{AB} documents containing both word A and B in the collection, then $p(X_A = 1, X_B = 1) = \frac{N_{AB}}{N}$
- $p(X_A = 1, X_B = 0)$: the probability that word A occurs in one document but B does not occur in that document. It can be calculated as $p(X_A = 1, X_B = 0) = \frac{N_A - N_{AB}}{N}$.

a. **[10 pts]** Given the values of N_A, N_B, N_{AB} for two words A and B in a collection of N documents, can you write down the formulas for the rest two joint probabilities of A and B, i.e. $p(X_A = 0, X_B = 1)$ and $p(X_A = 0, X_B = 0)$?

b. **[10 pts]** Next, we will use the following tables to do some real computation of Mutual Information. The tables contain the document counts for different words. There are a total of $N = 26,394$ documents in the collection.

Table 1 contains the document counts for words 'computer' and 'program', derived from the document collection (Hint: If $A = computer$ and $B = program$, then $N_{AB} = 349$. This means there are 349 documents that contain 'computer' AND 'program'):

	$X_{computer} = 1$	$X_{computer} = 0$
$X_{program} = 1$	349	2,021
$X_{program} = 0$	1,041	22,983

Table 2 contains the document counts for words 'computer' and 'baseball', derived from the same document collection:

	$X_{computer} = 1$	$X_{computer} = 0$
$X_{baseball} = 1$	23	2,121
$X_{baseball} = 0$	1,367	22,883

Calculate $I(X_{computer}\ ;\ X_{program})$ and $I(X_{computer}\ ;\ X_{baseball})$ using the document counts from Table 1 and 2.

c. **[5 pts]** Compare the results of $I(X_{computer}\ ;\ X_{program})$ and $I(X_{computer}\ ;\ X_{baseball})$. Do the results conform with your intuition? Explain your intuition.

d. **[10 points]** Next, we will use the CACM test collection to do some real computation. You can download the data from here, in which highly frequent words and very low frequent words have been removed (the vocabulary size of the original data is very large, so we won't use it for this assignment). The CACM collection is a collection of titles and abstracts from the journal CACM. There are about 3,000 documents in the collection. The data set has been processed into lines. Each line contains one document, and the terms of each document are separated by blank space.

Use any programming language you like (You may find it relatively easy if you use perl or python), for each pair of words in the collection, calculate the number of documents that contain both of the two words. Then, rank all the word pairs by their cooccurrence document counts. Print the largest 10 counts (one count number per line) (Hint: you may consider using hash to store the document counts for each word pair)

e. **[20 points]** Calculate the mutual information of all the possible word pairs in the collection. Rank the word pairs by their mutual information and print the results out. How are the top 10 pairs with the highest mutual information different from the top 10 pairs that you've got from problem B (i.e., the 10 pairs with the highest counts of co-occurrences)? Please write down the top 5 words which have the highest mutual information with word "programming" in the collection. Do you think your results reasonable? (Hint: In practice, we need to do some smoothing in our formulas in order to avoid the log0 problem. For joint probability estimation, we assume that each of the four cases (corresponding to four different combinations of values of X_a and X_b) gets 0.25 pseudo count, thus in total we introduced 0.25*4=1 pseudo count. We can then compute marginal probability based on the joint probability, i.e. $p(X_a=1)=p(X_a=1,X_b=0)+p(X_a=1,X_b=1)$. For example, $p(X_A=1, X_B=1) = (N_{AB} + 0.25) / (1 + N)$ and $p(X_A=1) = (N_A + 0.5)/(1+N)$. Please use these smoothing formulas in your code)

6. Kullback-Leibler Divergence (KL Divergence) [20 pts]

a. **[5 pts]** Please answer the following questions: 1) Whats the range of KL Divergence? 2) Under which circumstances is KL Divergence equal to 0?

b. **[10 pts]** From the course we know that KL Divergence is not symmetric. Show that this is true by creating two probability distributions p and q, where $D(p||q) \neq D(q||p)$.

c. **[5 pts]** When calculating $D(p||q)$, what issues do you run into when an event has 0 probability in distribution q? How can you deal with 0 probabilities in this case?