CIS 521: HW 5 - Probability

1 Instructions

This homework containts only a written portion (no programming). Please submit your answers on blackboard by the **beginning** of class on **Thursday, March 22**.

Let us know if you have any questions, check the discussion board (i.e. Google Groups), and remember that you can always come to Office Hours, even if only to keep us company.

2 Written Portion

1. (20 points) **Real World Probabilistic Modeling**
   Consider the following question: "What is the probability of Obama being re-elected?" What does this question mean to a frequentist? I.e., What is the event space in which something (what thing) could happen 100 times such that a certain fraction the answer would be "true" (or "re-elected") and the rest of the time it would be false? Put differently, how would you estimate the probability of Obama being re-elected? What data would you collect?

2. (20 points) **Belief Net Construction**
   Given the following observed counts for all different combinations of the binary random variables A, B, C and D (each variable can be true (T) or false (F)), construct a belief net using the algorithm described in class, where variables are added sequentially to the network. Note that there are 64 observations in total. Consider the variables in the order A, B, C, D, and make sure to give *both* the graph and the conditional probability tables.

A	B	C	D	Count	A	B	C	D	Count
T	T	T	T	6	F	T	T	T	18
T	T	T	F	6	F	T	T	F	18
T	T	F	T	0	F	T	F	T	0
T	T	F	F	4	F	T	F	F	18
T	F	T	T	6	F	F	T	T	12
T	F	T	F	6	F	F	T	F	18
T	F	F	T	0	F	F	F	T	0
T	F	F	F	4	F	F	F	F	6

3. (18 points) **Naive Bayes**

You have been hired as an expert witness for the defense to help convince the Jury that the defendant did not write the threatening email that the murder victim received. You decide to use the old trick of noting word frequencies used by different authors. You collect emails and tweets by the three suspects, Alice, Bob and Carol, discard the content words that relate to subject of the emails, and keep only the style words. Here are the counts you find

	Alice	Bob	Carol
Moreover	0	0	3
OMG	2	4	2
BBFL	5	0	1
:)	5	4	2

Your task is to calculate the probability that each of the suspects wrote the threats. You will, of course, use a Nave Bayes model as described in class.

(a) (8 points) Before we look at the threats, you need to compute some probabilities that you will need. What are the base rates:

P(word=Moreover) _____
P(word=OMG) _____
P(word=BBFL) _____
P(word=:)) _____

What are the MLE estimates of the probabilities of each word given each author, P(word—author):

	Alice	Bob	Carol
Moreover	_____	_____	_____
OMG	_____	_____	_____
BBFL	_____	_____	_____
:)	_____	_____	_____

What are the smoothed estimates of the probabilities of each word given each author, P(word—author) if you use the simple smoothing of adding 1 to each count?

	Alice	Bob	Carol
Moreover	_____	_____	_____
OMG	_____	_____	_____
BBFL	_____	_____	_____
:)	_____	_____	_____

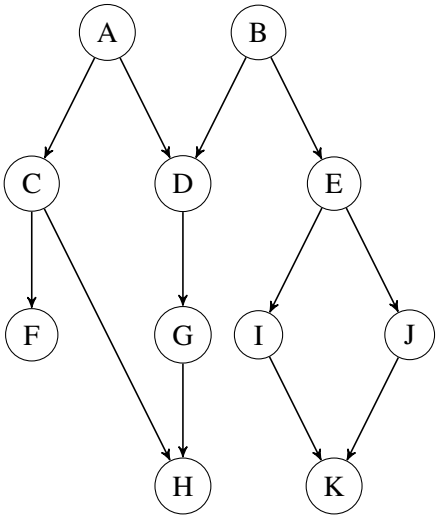
(b) (3 points) What is the Nave Bayes formula that gives probability $p(author|words)$ if all four words were present in the document being identified?
I.e. $p(author = Alice|Moreover, OMG, BBFL, :))$

(c) (1 point) The email in question actually read:
Dear David, you are a real jerk. :) You shouldnt have done that. I hate you. OMG, bad things will happen you. your former BBFL
Which of our target words showed up in the email?

(d) (6 points) Plug in the numbers (use the smoothed probability estimates) into the Nave Bayes formula to determine which of the three suspects is most likely to have written the threat. Remember, you will need to drop from the formula any words that dont show up in the document. Since we dont know otherwise, lets start with priors $p(author = alice) = p(author = bob) = p(author = carol) = 1/3$. Please show your work.

4. (14 points) **Conditional Independence in Bayesian Network**

Consider the following dependency graph, which of the following conditional independence is true according to the graph?



(a) $P(F, K) = P(F)P(K)$

(b) $P(A, B) = P(A)P(B)$

(c) $P(A, B|H) = P(A|H)P(B|H)$

(d) $P(H, E|G) = P(H|G)P(E|G)$

(e) $P(E, K|I, J) = P(E|I, J)P(K|I, J)$

(f) $P(C, D|A, H) = P(C|A, H)P(D|A, H)$

(g) $P(C, G|A) = P(C|A)P(G|A)$