Practice Midterm II

CMPS 140 – Winter 2019
2:40pm - 3:45pm, March 1st, 2019

Your name:

Your id:

Problems	Total points	Your points
Q1		
Q2		
Q3		
Bonus		
Total	100 (+?)	

Your final score is determined by $\min\{\text{Your total points}, 100\}$.

1 Fact checking & short answers [30 pts]

Answer the following questions with short answers. Explain your answers. For true or false questions, simply answering true or false without explanation will **not** receive points.

a) The following is true, without needing any independence assumption

$$P(A|B,C) = \frac{P(A|C)P(B,C)}{P(C)}$$

b) The following is true, assuming $\mathbf{A} \perp \mathbf{B} \mid \mathbf{C}$:

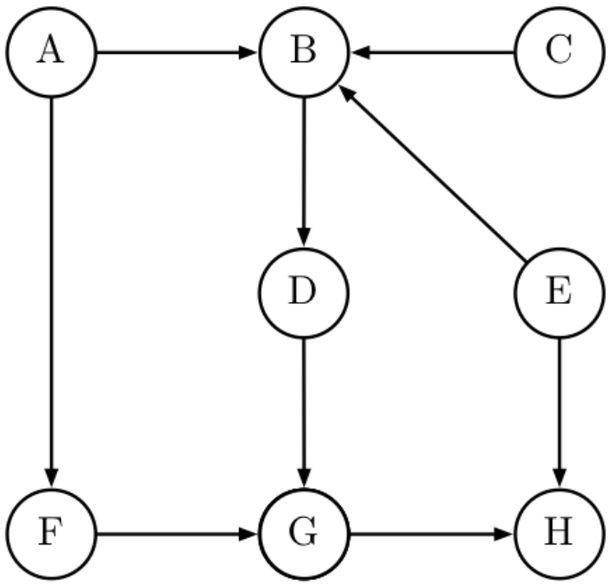
$$P(A,B|C) = \frac{P(B,C|A)P(A)}{P(B,C)}$$

[Well, in order to perform well in this exam, you need to master at least the chain rule and the Bayes rule.]

c) Review lecture slides and your notes

- I'll ask questions like "we mentioned this in the lecture. Why do we need it?"
- also think about what do we do when building a naive bayes classifier? why?
- What do we do when building a decision tree?
- How does a perceptron algorithm work?

2 Bayesian Networks [30 pts]



- What do we need to completely characterize the joint distribution for this BN?
- Which variables are independent and dependent of F, without assuming any observations?
- Which variables are independent and dependent of F, if G is given?

[Structure of BN, independence conditions: the three rules.]
[Work on your assignment 3.]
[You will have a simplier graph in the real exam...]

3 Naïve Bayes

The Naïve Bayes model has been famously used for classifying spam. We will use it in the "bag-of-words" model:

- Each email has binary label Y which takes values in $\{\text{spam}, \text{ham}\}$.
- Each word w of an email, no matter where in the email it occurs, is assumed to have probability $P(W = w \mid Y)$, where W takes on words in a pre-determined dictionary. Punctuation is ignored.
- Take an email with K words $w_1, \ldots, w_K$. For instance: email "`hi hi you`" has $w_1 = \text{hi}, w_2 = \text{hi}, w_3 = \text{you}$. Its label is given by

$$\arg\max_y P(Y = y \mid w_1, \ldots, w_K) = \arg\max_y P(Y = y) \prod_{i=1}^{K} P(W = w_i \mid Y = y).$$

Above is the background. Answer the following questions:

a) You are in possession of a bag of words spam classifier trained on a large corpus of emails. Below is a table of some estimated word probabilities.

W	keyword1	keyword2	keyword3
$P(W \mid Y = \text{spam})$	1/2	1/10	1/10
$P(W \mid Y = \text{ham})$	1/10	1/2	1/3

You are given a new email to classify

`keyword1, keyword3`

What is the probability

$$P(Y = \text{spam} | \texttt{keyword1, keyword3})$$

b) You are given only three emails as a training set:

(Spam) `Whatever amount you send do let us know. Remember, no amount below` $50 `will be accepted.`

(Ham) `hey, lunch at 12?`

(Ham) `sorry I have a midterm tomorrow.`

You are training with the emails with Laplace Smoothing with $k = 2$. There are V words in the dictionary. Write expressions for:

$P(W = \text{amount} \mid Y = \text{spam})$

hint: given an email is spam, randomly see a word, what's the chance of being "amount"

$P(Y = \text{spam})$

Try other combinations yourself and familiarize yourself with the counting and computation.