Homework 4
Semi-supervised Learning, Graphical Models

CMU 10-601: Machine Learning (Fall 2015)
http://www.cs.cmu.edu/~10601b/
OUT: Oct 13, 2015
DUE: Oct 27, 2015, 10:30 AM

START HERE: Instructions

- The homework is due at 10:30 am on Thursday October 27, 2015. Each student will given two late days that can be spent on any homeworks but not on projects. Once you have used up your late days for the term, late homework submissions will receive 50% of the grade if they are one day late, and 0% if they are late by more than one day.
- ALL answers will be submitted electronically through the submission website: https://autolab.cs.cmu.edu/10601-f15. You can sign in using your Andrew credentials. You should make sure to edit your account information and choose a nickname/handle. This handle will be used to display your results for any competition questions (such as the class project) on the class leaderboard.
- There are no autograded questions on this homework.
- Collaboration on solving the homework is allowed (after you have thought about the problems on your own). When you do collaborate, you should list your collaborators! You might also have gotten some inspiration from resources (books or online etc...). This might be OK only after you have tried to solve the problem, and couldn't. In such a case, you should cite your resources.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution. You should also state your collaborations in your short-answer writeup. Specifically, please write down the following:

  1. Did you receive any help whatsoever from anyone in solving this assignment? Yes / No. If you answered yes, give full details: (e.g., "Jane explained to me what is asked in Question 3.4").
  2. Did you give any help whatsoever to anyone in solving this assignment? Yes / No. If you answered yes, give full details: (e.g., "I pointed Joe to section 2.3 to help him with Question 2").

  Collaboration without full disclosure will be handled severely, in compliance with CMU's Policy on Cheating and Plagiarism.

1 Probability [Zhenzhen; 16 points]

Given the following joint probability distribution table. Note that calculation is not needed for some of the questions.

	Y=0	Y=1
X=0	3/8	1/8
X=1	3/8	1/8

Table 1: Joint probability distribution of X and Y

1. Are X and Y independent? Show your reasoning. [4pt]
2. Suppose random variable $Z = X + Y$. Are X and Y independent if conditioned on Z? Show your reasoning. [4pt]

3. Suppose random variable $W = XY$. Are X and Y independent if conditioned on W? Show your reasoning. [4pt]

4. Suppose C is the result of an independent coin flip. Are X and Y independent if conditioned on C? Show your reasoning. [4pt]

2 Semi-Supervised Learning [Zhenzhen; 16 points]

We would like to build a naive Bayes classifier for classifying an email as 'spam' or 'not spam'. Our naive Bayes classifier uses a multinomial distribution to model class-conditional word probabilities. Given a dataset of four emails $\{\text{Email}_1, ..., \text{Email}_4\}$ and their corresponding labels $\{\text{Label}_1, \ldots, \text{Label}_4\}$ as well as two emails $\{\text{Email}_5, \text{Email}_6\}$ without labels, we would like to perform a semi-supervised learning to train our classifier. The data are given in Table 2, where 'T' and 'F' correspond to 'spam' and 'not spam', respectively. Assume a bag-of-words representation and consider a vocabulary of 11 words { 'simply', 'reply', 'to', 'this', 'customer', 'placement', 'linux', 'privacy', 'fast', 'credit', 'sale'}.

No.	Email Content	Label
1	simply, reply, to, this	T
2	reply, customer, placement	F
3	sale, reply, to, fast	T
4	privacy, credit, linux	F
5	linux, credit	?
6	reply, to, sale	?

Table 2: Emails

w_1 simply	w_2 reply	w_3 to	w_4 this	w_5 customer	w_6 placement	w_7 linux	w_8 privacy	w_9 fast	w_{10} credit	w_{11} sale	Label
1	1	1	1	0	0	0	0	0	0	0	T
0	1	0	0	1	1	0	0	0	0	0	F
0	1	1	0	0	0	0	0	1	0	1	T
0	0	0	0	0	0	1	1	0	1	0	F
0	0	0	0	0	0	1	0	0	1	0	?
0	1	1	0	0	0	0	0	0	0	1	?

Table 3: Bag of words representation

Recall the EM algorithm for learning a Naive Bayes classifier from both labeled and unlabeled data. In order to initialize the EM algorithm, we first use the labeled data to obtain a maximum likelihood estimate of the parameters and then iterate between E step and M step to train with both labeled and unlabeled data. Note that when you estimate the class-conditional word probabilities during initialization and M steps, you should use a pseudocount of two, one for observing the word and another for not observing the word for the given class, since many of the word counts are 0. After the initialization, perform your first E-step and compute the following:

1. $P(\text{Label}_5|\text{Email}_5)$ [5pt]

2. $P(\text{Label}_6|\text{Email}_6)$ [5pt]

3. Now we perform an M step to retrain the classifier using both labeled and unlabeled emails. Using the new classifier, how would you classify a new email that contains two words 'to credit'? Show your calculation. [6pt]

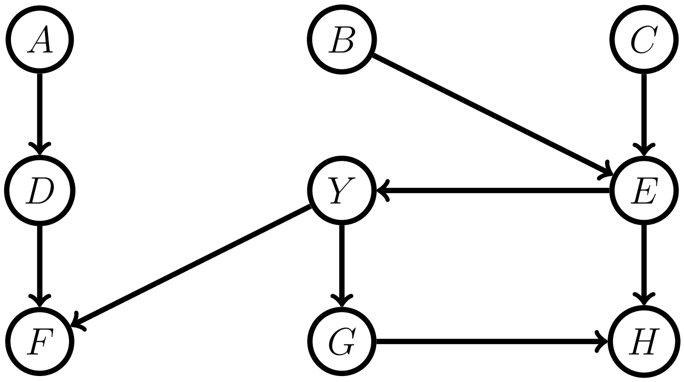

Figure 1: Bayesian networks

3 Graphical Models Representation [Pengcheng Xu; 16 points]

1. Suppose we are going to classify documents for a newspaper agency. Y is the type of the document. A is the length of the document. B is most frequent word in the document. C is publish date of the document. We use Naive Bayes to do the classification. Write down the factorization of the joint probability P(Y,A,B,C), and draw the corresponding Bayesian Network. [4pt]

2. Given the Bayesian network from the last question, assume the length of the document A is also dependent on the publish date C. Write down the factorization of the joint probability P(Y,A,B,C), and draw the Bayesian Network. [4pt]

3. Suppose we have the variables related to the document type Y and the dependency relationships in Figure 1. List all the variables we must consider when we try to classify the document. [4pt]

4. Based on last question, Write down the factorization of the joint probability when we classify the document. [4pt]

4 Independence Assumption [Pengcheng Xu; 24 points]

Consider the Bayesian network in Figure 2. Answer the following questions

1. Conditional on D, list all variables that are independent of G. [4pt]

2. List all variables that are independent of A. [4pt]

3. Conditional on D, list all variables that are independent of I. [4pt]

4. Conditional on E, F, G, write down the set of variables whose joint probability is independent of D, and the set has to be as large as possible. [4pt]

5. Conditional on D, write down the set of variables, whose joint probability is independent of E, F, G, and the set has to be as large as possible. [4pt]

6. Conditional on E, write down the set of variables whose joint probability is independent of H, and the set has to be as large as possible. [4pt]

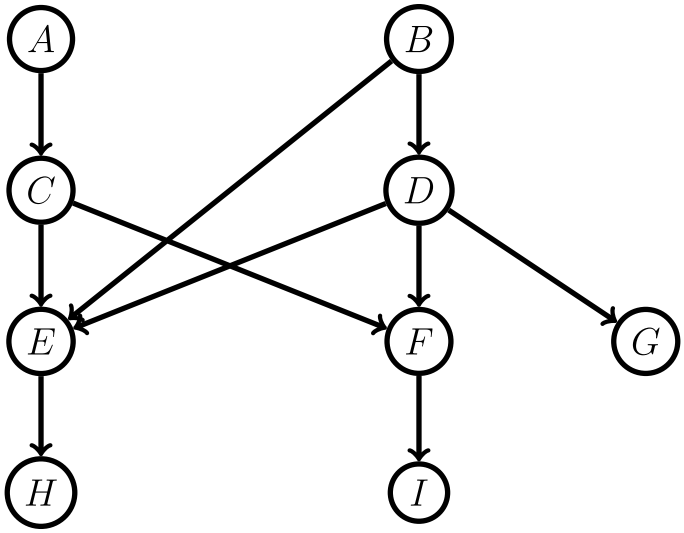

Figure 2: Bayesian network

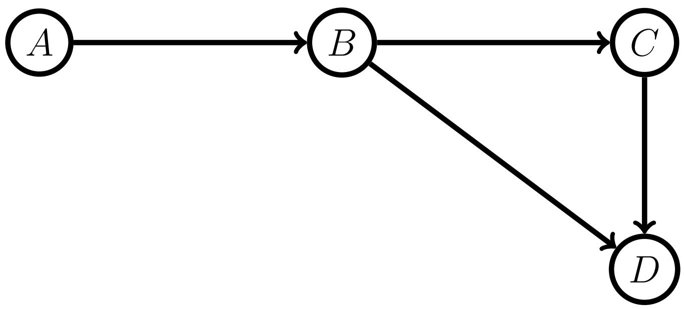

Figure 3: Bayesian networks

5 Bayesian Network Inference [Pengcheng Xu; 10 points]

Consider the Bayesian network in Figure 3 over 4 binary random variables.

$P(A = true) = 0.2$
$P(A = false) = 0.8$

$P(B = true|A = true) = 0.5$
$P(B = false|A = true) = 0.5$
$P(B = true|A = false) = 0.4$
$P(B = false|A = false) = 0.6$

$P(C = true|B = true) = 0.9$
$P(C = false|B = true) = 0.1$
$P(C = true|B = false) = 0.2$
$P(C = false|B = false) = 0.8$

$P(D = true|B = true, C = true) = 0.99$
$P(D = false|B = true, C = true) = 0.01$
$P(D = true|B = true, C = false) = 0.98$
$P(D = false|B = true, C = false) = 0.02$
$P(D = true|B = false, C = true) = 0.95$

$P(D = false|B = false, C = true) = 0.05$
$P(D = true|B = false, C = false) = 0.9$
$P(D = false|B = false, C = false) = 0.1$

Compute the following quantities. Show your work.

1. $P(A = false, B = false, C = false, D = false)$ [1pt]
2. The distribution $P(B|A = true, C = false)$ [2pt]
3. The distribution $P(D|A = false)$ [3pt]
4. The distribution $P(D|B = false)$ [3pt]
5. $P(D = true|A = true, B = true, C = true)$ [1pt]

6 Bayesian Networks Learning [Zhenzhen; 18 points]

Consider the Bayesian network in Figure 3 over four binary random variables, A, B, C, and D. Suppose we want to learn the parameters of the Bayesian network, given the training data in Table 4. Note that the observation for C is missing for training example 17. We would like to use both fully observed and partially observed data in Table 4 during training. Let's denote the value for variable X of the i_{th} training example as X_i. For example, $A_{17} = 0$. For training example 1 through 16, we observe each possible assignment for A,B,C,D exactly once.

No.	A	B	C	D
1	0	0	0	0
...	...	...	...	...
16	1	1	1	1
17	0	1	?	0

Table 4: Training Examples

After initialization of the parameters with MLE from only the fully observed data, for the first E step, compute

1. $P(C_{17} = 1|A_{17}, B_{17}, D_{17})$ [3pt]
2. $P(C_{17} = 0|A_{17}, B_{17}, D_{17})$ [3pt]

After this E step, now for the first M step, update

1. $P(C|B = 1)$ [3pt]
2. $P(C|B = 0)$ [3pt]
3. What is $P(A = 0, B = 1, C = 0, D = 1)$ if we stop here? [3pt]

Make sure you have updated all relevant parameters. For the second E step, use the updated parameters to compute

1. $P(C_{17} = 1|A_{17}, B_{17}, D_{17})$ [3pt]