BMI/CS 576
Introduction to Bioinformatics
Fall 2007 Final Exam

Name ____

Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, make sure your exam has every page (numbered **1** through **10**).

Problem	Score	Max Score
1.	____________	12
2.	____________	10
3.	____________	14
4.	____________	12
5.	____________	12
6.	____________	14
7.	____________	16
8.	____________	10
Total		100

1. *K*-Means Clustering (12 points): Show how *k*-means would cluster the following genes represented as vectors:

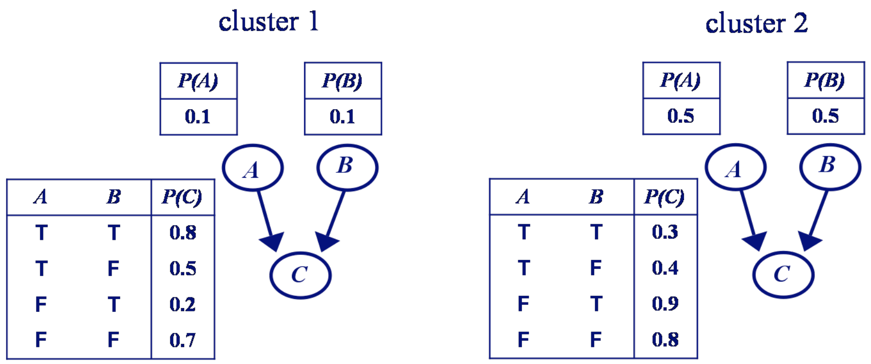
$x_1 = \langle 3, 1 \rangle$
$x_2 = \langle 3, 3 \rangle$
$x_3 = \langle 5, 2 \rangle$
$x_4 = \langle 7, 8 \rangle$

Assume that $k = 2$, the initial coordinates of the cluster centers are $\langle 2, 2 \rangle$ and $\langle 6, 3 \rangle$ and we are using Manhattan distance: $\text{dist}(x_i, x_j) = \sum_e \left| x_i[e] - x_j[e] \right|$

2. EM Clustering (10 points): Consider using EM clustering to cluster instances that are represented three by Boolean values:

$x_1 = \langle \neg a, b, \neg c \rangle$
$x_2 = \langle \neg a, b, c \rangle$
$x_3 = \langle a, \neg b, \neg c \rangle$
…

Suppose that we have two clusters, and each represented by a Bayes net as shown below.



cluster 1

P(A)
0.1

P(B)
0.1

A	*B*	*P(C)*
T	T	0.8
T	F	0.5
F	T	0.2
F	F	0.7

cluster 2

P(A)
0.5

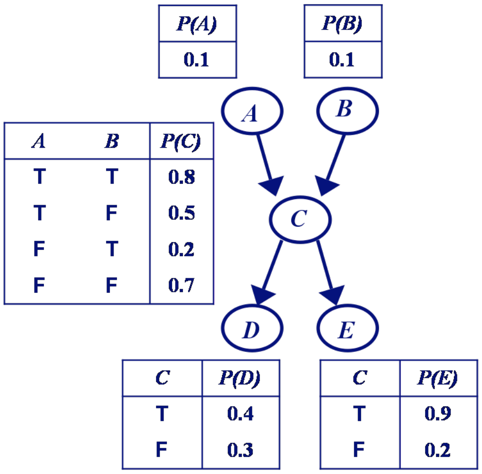
P(B)
0.5

A	*B*	*P(C)*
T	T	0.3
T	F	0.4
F	T	0.9
F	F	0.8

(a) Show the calculations that would be done in the E-step for the three instances shown.

(b) Briefly describe how you would do the M step.

3. Bayes Net Inference (14 points): For the Bayes network below, show how you would answer the query P(*a* | *b, e*) (i.e., the probability that *A* is True, given that *B* is True and *E* is True) using *Variable Elimination* with the variable ordering *E, D, C, B*.



P(A)
0.1

P(B)
0.1

A	*B*	*P(C)*
T	T	0.8
T	F	0.5
F	T	0.2
F	F	0.7

C	*P(D)*
T	0.4
F	0.3

C	*P(E)*
T	0.9
F	0.2

4. Gene Expression Analysis (12 points):

One challenge in gene-expression analysis is the high-dimensionality of the data (i.e. measurements are made for thousands of genes). Two methods that we discussed for dealing with this issue, in different contexts, are (i) calculating a *false discovery rate* and (ii) using *bootstrapping to characterize features* in a Bayes net. Briefly compare and contrast these two methods in terms of:

(a) what the methods take as input,
(b) what they calculate and return as output.

5. Gene Expression Analysis (12 points):

Suppose you are working with a biologist to study the gene-expression responses in a bacterium that is treated with various antibiotics. Each experiment involves treating a batch of the bacteria culture with a particular antibiotic and then using a microarray to measure changes in gene expression. Your role in the project is to use Bayes net methods to induce a model of the relationships among affected genes.

(a) How would your analysis benefit by having the expression response for each antibiotic be measured at *several* time points?

(b) Suppose the biologist can knock-out (i.e. disable) any gene before running one of these experiments. How might you use your current Bayes net model to select the genes that would be most informative to knock out? Why would this particular type of experiment be useful?

(c) How would you assess the accuracy of your Bayes net model?

6. Threading (14 points): Consider a simple threading problem in which we have a template with three segments (*i, j, k*). We are given a sequence for which there are two possible starting positions for each segment.

(a) Given the following values for the scores of the individual segments and the scores for segment interactions, show how the branch-and-bound method would find the optimal threading.

$g_1(i, 2) = 4$	$g_1(j, 8) = 2$	$g_1(k, 13) = 1$
$g_1(i, 3) = 3$	$g_1(j, 9) = 5$	$g_1(k, 14) = 10$
$g_2(i, j, 2, 8) = 6$	$g_2(i, k, 2, 13) = 1$	$g_2(j, k, 8, 13) = 3$
$g_2(i, j, 2, 9) = 0$	$g_2(i, k, 2, 14) = 0$	$g_2(j, k, 8, 14) = 12$
$g_2(i, j, 3, 8) = 1$	$g_2(i, k, 3, 13) = 9$	$g_2(j, k, 9, 13) = 5$
$g_2(i, j, 3, 9) = 0$	$g_2(i, k, 3, 14) = 0$	$g_2(j, k, 9, 14) = 11$

Use the "simple lower bound" presented in class. When splitting a threading, split the segment having the minimal g_1 value for some position (e.g. split on k first since $g_1(k, 13) = 1$). To split a selected segment, divide it into two intervals of length one.

(b) Is there a case in the previous solution where the value calculated by the lower bound does not correspond to a legal threading? If so, show such a case.

(c) Could we use dynamic programming instead of branch-and-bound to solve this threading problem? Briefly justify your answer.

7. Branch and Bound Search (16 points): We considered branch-and-bound search in two contexts: (i) maximum parsimony inference of phylogenetic trees, and (ii) protein threading. For both of these tasks, briefly describe the following:

(a) What do states represent?
(b) How are states expanded into other states (i.e. what do the operators do)?
(c) How are lower bounds calculated (i.e. what do they take into account)?
(d) What are the stopping criteria (i.e. how do we know when the search is done)?

8. Short Answer (10 points): Briefly define each of the following terms.

multiple testing problem

cross validation

Markov blanket

module network

homology modeling