CS 150AIH EXAM Fall 2011

This Exam is being given under the guidelines of the **Honor Code**. You are expected to respect those guidelines. This exam is due at the beginning of class on **Wednesday, November 2**. There are 6 questions for a total of 120 points. (Each question is worth 20 points total).

Name: _____

1. This year (2011) marks the 30-year anniversary of the introduction of what would eventually be referred to as *acquired immunodeficiency syndrome* (AIDS) to the medical community. However, in the early 80's, significantly less was known about this disease. In the appendix are two early case studies concering early reports regarding AIDS, if you would like some background information.

   Obviously, early on there was uncertainty regarding even a definition of the disease. We will model this with a Bayesian network. For simplicity we will model 5 Boolean variables believed to be potentially relevant at the time: Haitian origin (H), Kaposi's sarcoma (K), homosexual male (M), *Pneumocystis carinii* pneumonia (P), and the latent "mystery" disease (A).

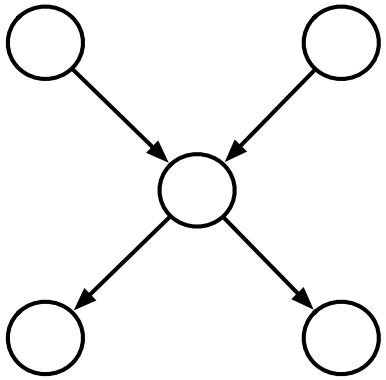

Figure 1: Structure for AIDS Bayesian Network

   (a) (5 points) Given the structure in Figure 1 and the assumption of causal edge semantics, assign the relevant variables to appropriate vertices and generate corresponding conditional probability tables. Note that we do not intend for the estimated parameters to be accurate in the absolute, but just rough estimates, based on your intuition.

   (b) (5 points) Provide the formula for and calculate $p(k|a)$ and $p(k|a,p)$. Briefly interpret this result.

   (c) (5 points) Provide the formula for and calculate $p(h)$, $p(h|a,m)$, and $p(h|a,\neg m)$. Briefly discuss the result of these computations – is there something you can say generally about the subgraph defined by these vertices.

   (d) (5 points) Calculate $p(k)$ and $p(k|m)$ (note that you do not have to write down the formula for this last part if you are using a program). Briefly interpret these results.

2. Within the following years, the human immunodeficiency virus (HIV) was isolated and determined to be the cause of AIDS. By 1985, an antibody screening test was approved. In 2001 it

was mandated that all donated blood in the United States be screened with polymerase chain reaction (PCR) tests. Table 1 represents a hypothetical study of a specific test performed on a population of male intravenous drug users.

PCR result	Gold Standard (+)	Gold Standard (−)	Total
+	72	12	84
−	3	71	74
Total	75	83	158

Table 1: PCR study for intravenous drug users

(a) (5 points) Calculate the sensitivity, specificity, disease prevalence, positive predictive value, and negative predictive value.

(b) (3 points) An asymptomatic male has generated a positive test when donating blood. He has no discernible elevated risk factors and you know that the prevalence of HIV in male intravenous drug users is 25 times as high as in the male community at large. Estimate the pretest probability that this man is infected with HIV.

(c) (4 points) Calculate the post-test probability of the patient having HIV after a positive PCR test.

(d) (4 points) Calculate the post-test probability of the patient having HIV after a negative PCR test.

(e) (4 points) Upon observing a surge of positive tests which were later determined to be negative cases, you decide you wish to develop a test with increased post-test probability of the disease given a positive test results. Should you focus on improving the TPR or TNR of the test – and why?

3. Recall that the cosine similarity between two vectors (e.g., representing documents) d_1 and d_2 is defined thusly:

$$cos(\theta) = \frac{d_1 \cdot d_2}{||d_1||||d_2||} \tag{1}$$

This similarity is defined over 'flat' representations. Suppose we are interested in the similarity of biomedical texts. Further suppose that these have been manually tagged with Medical Subject Headings (MeSH),[1] as is often the case. MeSH terms are hierarchical; *Men* are *Persons* as are, e.g., *Women*, *Students* and *Alcoholics* (amongst others). Assume you are given the *nodes* (lowest-level, e.g., *Men* rather than *Persons*) entry for each relevant term characterizing a text. Denote this set of terms for d_i by $MeSH(d_i)$. Assume you have a look-up dictionary, $\mathcal{D}$, that returns the relevant path for a given leaf-node,[2] with which to construct the feature representation.

(a) (10 points) Using $MeSH$ and $\mathcal{D}$, define a feature-mapping function $\mathcal{F}$ that creates a representation of a given document d_i such that the cosine similarity (Equation 1) between $\mathcal{F}(d_i)$ and $\mathcal{F}(d_l)$ is defined over the $MeSH$ space. What are the benefits of this approach, versus the distance over flat, unigram space? What, if any, are the drawbacks?

[1] This is an ontology. See: http://www.nlm.nih.gov/mesh/.

[2] Further assume that you have a mapping from feature index j to the word/token.

(b) (10 points) Now design a similarity function defined over document $MeSH$ terms that explicitly takes into account the hierarchal distance between terms in documents. For example, if $MeSH(d_1) = \{Man\}$ and $MeSH(d_2) = \{Woman\}$, we want the similarity function to reflect that these terms are both one-level below their common ancestor of *PERSON*. Your function should be defined generally, so that it is easy to drop in different functions mapping ontological distances to scalars (e.g., linear vs. exponential functions).

4. This question is about Markovian modeling of clinical processes. Specifically, suppose you are tasked with modeling the progression of the deadly disease Smallitus. There are four clinically relevant states in this disease: **1** healthy; **2** initial infection/partial Smallitus; **3** full-blown infection/total Smallitus and **4** death. The progression is not always unidirectional, i.e., the disease sometimes goes into remission (patients get better). The disease status is directly observable via simple, readily available, infallible clinical tests.

   Medical researchers have conducted a trial involving 1000 patients. They kept records of the disease onset and progression in these patients, measured at fixed intervals of D days. The data is provided in the table below.

origin state \ destination state	1	2	3	4
1	500	300	0	0
2	200	80	200	0
3	0	100	150	0
4	0	0	0	200

   (a) (2 points) Is using a Markov model appropriate in this case? Why or why not?

   (b) (3 points) Do we need to include hidden states here? If so, what do the hidden states represent? If not, why?

   (c) (5 points) Calculate the state-transition probability matrix, $\mathcal{A}$.

   (d) (5 points) Draw the state transition diagram (i.e., the graphical representation of the Markov model); include arrows between states only when the corresponding transition probability is non-zero. Annotate these arrows with their transition probabilities (calculated above).

   (e) (5 points) Tragically, your friend has come down with Smallitus (i.e., she enters state **2**). She wants to know how she'll be in $5D$ days time. Calculate the distribution over states **1** through **4** after this many intervals. (You almost certainly want to code this; please hand in the code you use). Is there reason to hope?

5. An insidious new variant of Smallitus has appeared. The disease is still (potentially) fatal, but no longer manifests itself via the aforementioned simple clinical tests. Fortunately, doctors have invented a new test that partially corresponds to the disease status. In particular, this test produces one of the following readings: $\{a, b, c\}$, corresponding to the respective states of

progression (taking this measurement is of course unnecessary for the deceased); thus the total alphabet is $\{a, b, c, DEATH\}$. We are now given a small amount of data from a new study investigating this variant of Smallitus as a sequence of observations for patients.[3] The data is below.

(a) (5 points) How would you modify the above model to accommodate the uncertainty inherent to the measurements? Draw and annotate your model.

(b) (10 points) Using a software package of your choice,[4] estimate the model parameters from the data. Print out the (estimated) start state distribution, as well as the transition and emission probability matrices. Do these seem reasonable? Please hand in any code you used to solve this.

(c) (5 points) How might you improve these results (i.e., make the parameter estimation more accurate) by incorporating domain knowledge? For example, we might believe the deadly variant is in fact similar to the original Smallitus – thus it seems natural to somehow use the information collected for this disease (i.e., the data in Question 4) here. How might you accomplish this? You do not have to implement your proposed solution.

Here is the data.

a, a, b, b, c, c,DEATH,DEATH

b, b, c, c, b, b, b, b

a, a, a, a, a, a, a, a

c, c, c,DEATH,DEATH,DEATH,DEATH,DEATH

b, b, c, c, b, b, b, b

a, a, a, b, b, b, b, c

a, a, b, b, b, a, a, a

b, b, b, c, c, c, b, b

a, a, a, a, b, a, a, a

c, c, c, c, b, b, b, b

a, a, a, b, b, a, a, b

a, a, b, b, b, b, b, a

b, b, c, c, c,DEATH,DEATH,DEATH

a, a, a, a, a, a, a, a

c, c, c,DEATH,DEATH,DEATH,DEATH,DEATH

a, a, b, b, b, b, a, a

b, b, b, b, c, c, c,DEATH

a, a, a, a, a, a, a, a

b, b, b, a, a, a, b, b

c,DEATH,DEATH,DEATH,DEATH,DEATH,DEATH,DEATH

a, b, b, b, b, a, a, a

6. Early cases of Smallitus seemed to affect only short people who scored high on a test where the patient would attempt to recall the US state capitals (given a list of states) within a ten

[3] Assume for the moment that we can no longer rely at all on the previously provided data, because we are uncertain if the disease progression is similar.

[4] Or you can calculate this by hand, if you insist.

minute period. However, as more cases were discovered amongst taller people it was clear that Smallitus also affects people with lower recall on the given diagnostic test. Figure 2 represents a plot for the two seemingly relevant dimensions for diagnosing Smallitus.

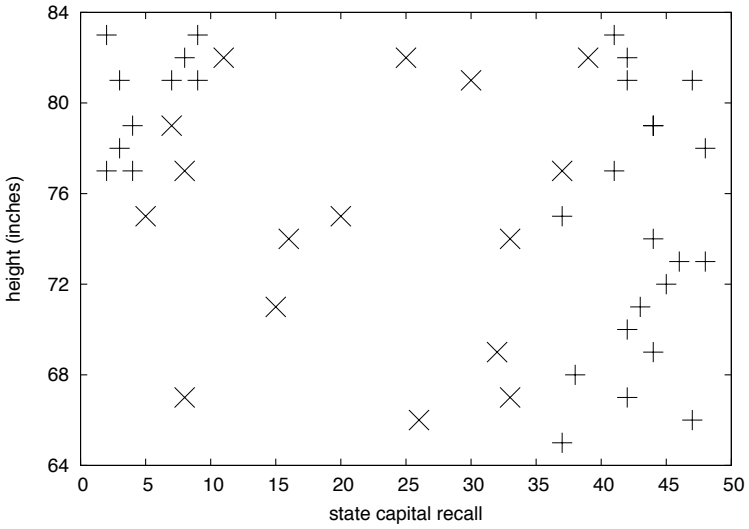

Figure 2: Study to generate diagnostic test for Smallitus

(a) (10 points) Assuming the threshold values align with the tic marks, derive two different decision trees – one which minimizes error and one which emphasizes generalization (clearly labeling each one). Note that we are asking you to do this by inspection and *not* by using ID3 or some other machine learning algorithm.

(b) (4 points) Calculate the sensitivity and specificity for each decision tree.

(c) (3 points) Briefly discuss how an asymmetric utility function (with respect to false positives and false negatives) might changes the desired decision tree (you don't have to actually generate a new decision tree).

(d) (3 points) Is a decision tree an appropriate mechanism for creating this test? Why or why not? What other classifiers might be appropriate and why?

Appendix

Pneumocystis Pneumonia – Los Angeles. MMWR 30(21):1-3, June 5, 1981.

In the period October 1980-May 1981, 5 young men, all active homosexuals, were treated for biopsy-confirmed *Pneumocystis carinii* pneumonia at 3 different hospitals in Los Angeles, California. Two of the patients died. All 5 patients had laboratory-confirmed previous or current cytomegalovirus (CMV) infection and candidal mucosal infection. Case reports of these patients follow.

Patient 1: A previously healthy 33-year-old man developed *P. carinii* pneumonia and oral mucosal candidiasis in March 1981 after a 2-month history of fever associated with elevated liver enzymes, leukopenia, and CMV viruria. The serum complement-fixation CMV titer in October 1980 was 256; in may 1981 it was 32.* The patient's condition deteriorated despite courses of treatment with trimethoprim-sulfamethoxazole (TMP/SMX), pentamidine, and acyclovir. He died May 3, and postmortem examination showed residual *P. carinii* and CMV pneumonia, but no evidence of neoplasia.

Patient 2: A previously healthy 30-year-old man developed *p. carinii* pneumonia in April 1981 after a 5-month history of fever each day and of elevated liver-function tests, CMV viruria, and documented seroconversion to CMV, i.e., an acute-phase titer of 16 and a convalescent-phase titer of 28* in anticomplement immunofluorescence tests. Other features of his illness included leukopenia and mucosal candidiasis. His pneumonia responded to a course of intravenous TMP/.SMX, but, as of the latest reports, he continues to have a fever each day.

Patient 3: A 30-year-old man was well until January 1981 when he developed esophageal and oral candidiasis that responded to Amphotericin B treatment. He was hospitalized in February 1981 for *P. carinii* pneumonia that responded to TMPSMX. His esophageal candidiasis recurred after the pneumonia was diagnosed, and he was again given Amphotericin B. The CMV complement-fixation titer in March 1981 was 8. Material from an esophageal biopsy was positive for CMV.

Patient 4: A 29-year-old man developed *P. carinii* pneumonia in February 1981. He had had Hodgkins disease 3 years earlier, but had been successfully treated with radiation therapy alone. He did not improve after being given intravenous TMP/SMX and corticosteroids and died in March. Postmortem examination showed no evidence of Hodgkins disease, but *P. carinii* and CMV were found in lung tissue.

Patient 5: A previously healthy 36-year-old man with clinically diagnosed CMV infection in September 1980 was seen in April 1981 because of a 4-month history of fever, dyspnea, and cough. On admission he was found to have *P. carinii* pneumonia, oral candidiasis, and CMV retinitis. A complement-fixation CMV titer in April 1981 was 128. The patient has been treated with 2 short courses of TMP/SMX that have been limited because of a sulfa-induced neutropenia. He is being treated for candidiasis with topical nystatin.

The diagnosis of *Pneumocystis* pneumonia was confirmed for all 5 patients antemortem by closed or open lung biopsy. The patients did not know each other and had no known common contacts or knowledge of sexual partners who had had similar illnesses. Two of the 5 reported having frequent homosexual contacts with various partners. All 5 reported using inhalant drugs, and 1 reported parenteral drug abuse. Three patients had profoundly depressed *in vitro* proliferative responses to mitogens and antigens. Lymphocyte studies were not performed on the other 2 patients.

Opportunistic Infections and Kaposi's Sarcoma among Haitians in the United States.
MMWR 31(26):353-354,360-361, July 9, 1982.

Reports of opportunistic infections and Kaposi's sarcoma among Haitians residing in the United States have recently been received at CDC. A total of 34 cases in 5 states have been reported to date.

Florida: From April 1, 1980, through June 20, 1982, 19 Haitian patients admitted to Jackson Memorial Hospital, Miami, had culture, biopsy, or autopsy evidence of opportunistic infections, and 1 other patient had biopsy- and autopsy-confirmed Kaposi's sarcoma. The infections identified included *Pneumocystis carinii* pneumonia (6 patients), cryptococcal meningitis or fungemia (4), toxoplasmosis of the central nervous system (CNS) (7), Candida albicans esophagitis (7) and thrush (5), esophageal or disseminated cytomegalovirus infection (3), progressive herpes simplex virus infection (1), disseminated tuberculosis (8), and chronic enteric Isospora belli infection (2). Fourteen patients had multiple opportunistic infections. Three patients had recurring infection. The clinical course has been severe; 10 patients have died. The type of infection was initially recognized at autopsy for 6 patients.

The 20 patients ranged in age from 22 to 43 years (mean 28.4 years); 17 were males. All the patients had been born in Haiti and had resided in the Miami-Dade County area for periods ranging from 1 month to 7 years (median 20.5 months).

When initially seen, 18 of the 20 patients had peripheral lymphopenia (1,000 lymphocytes/mm((3))). Skin tests performed on 17 patients with various combinations of tuberculin, mumps, streptokinase/streptodornase, Candida, and Trichophyton antigens were all negative. Immunologic studies at CDC on specimens from the 11 patients tested showed severe T-cell dysfunction. Monoclonal antibody analysis of peripheral-blood T-cell subsets revealed a marked decrease of the T-helper cell subset with inversion of the normal ratio of T-helper to T-suppressor cells.

Of the 7 patients with histologically confirmed toxoplasmosis of the CNS, 5 have died. Because there was no history of underlying conditions or drugs associated with immunosuppression, CNS toxoplasmosis was not considered in the premortem diagnosis of the first 4 cases. Pathology findings for all these patients were confirmed with an immunoperoxidase method for toxoplasmosis and, in one instance, with electron microscopy as well. Tachyzoites were the predominant form of the parasite observed; encysted forms were rare or absent in many tissue blocks.

In addition to the 20 cases reported from Miami, a Haitian female from Naples, Florida, was reported to have *P. carinii* pneumonia.

New York: From July 1, 1981, through May 31, 1982, 10 Haitian residents of Brooklyn were diagnosed as having the following opportunistic infections: *P. carinii* pneumonia (5 patients), CNS toxoplasmosis (2), disseminated cryptococcosis (1), esophageal candidiasis (1), and disseminated tuberculosis (2). None had any underlying disease or history of therapy known to cause immunosuppression. Five died of their infections.

All 10 patients were males and ranged in age from 22 to 37 years. Eight stated they were heterosexual; the sexual orientation of the other 2 was not known. One patient gave a history of intravenous (IV) drug abuse; 8 denied drug abuse, and for 1, no information was available on drug use. The 10 had resided in the United States for periods ranging from 3 months to 8 years (the majority, for 2 years or less). At least 1 patient had onset of illness before arriving in the United States. Immunologic studies performed at CDC on specimens from 2 patients showed results comparable to those for the 11 patients from Miami.

Other States: Opportunistic infections or Kaposi's sarcoma were also reported for 3 other Haitians located in California, Georgia, and New Jersey. All 3 were heterosexual males who denied IV drug abuse. One patient had *P. carinii* pneumonia, another had Kaposi's sarcoma, and the third had esophageal candidiasis.