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).

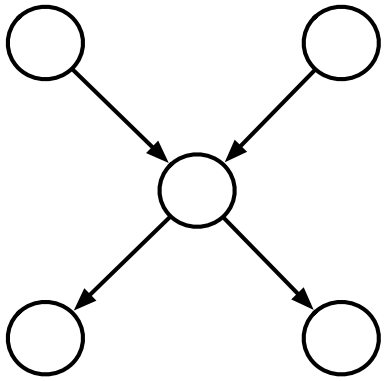


Figure 1: Structure for AIDS Bayesian Network

   (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.

   Solution:

   Figure 2 is the most natural structure for the described problem (although, strictly speaking, there are other possible structures). The estimated parameters below are based on published data when available, but any reasonable estimates were accepted.

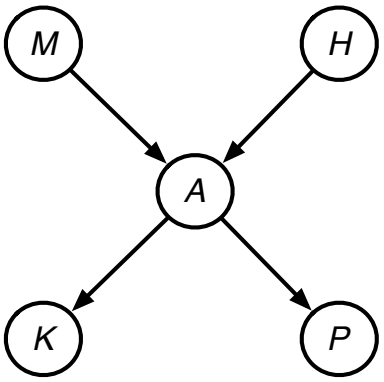


Figure 2: Structure for AIDS Bayesian Network

$p(M)$
0.05

$p(H)$
0.003

m	h	$p(A\|M=m,H=h)$
false	false	0.0001
false	true	0.0002
true	false	0.01
true	true	0.011

a	$p(K\|A=a)$
false	0.00005
true	0.3

a	$p(P\|A=a)$
false	0.001
true	0.65

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

Solution:

$$\begin{aligned} p(k|a,p) &= p(k|a) \\ &= 0.3 \end{aligned}$$

K and P are conditionally independent given A.

(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.

Solution:

$$\begin{aligned} p(h|a,m) &= \frac{p(h)p(a|m,h)}{p(h)p(a|m,h)+p(\neg h)p(a|m,\neg h)} \\ &= \frac{0.003 \cdot 0.011}{0.003 \cdot 0.011 + 0.997 \cdot 0.01} \\ &= 0.003299 \end{aligned}$$

$$\begin{aligned} p(h|a,\neg m) &= \frac{p(h)p(a|\neg m,h)}{p(h)p(a|\neg m,h)+p(\neg h)p(a|\neg m,\neg h)} \\ &= \frac{0.003\cdot 0.0002}{0.003\cdot 0.0002+0.997\cdot 0.0001} \\ &= 0.005982 \end{aligned}$$

This phenomenon is sometimes referred to as *explaining away* the result. The implication here is that if the patient has AIDS but is not a homosexual male, he is likely to be Haitian.

(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.

Solution:

$$p(k) = 0.000678$$

$$p(k|m) = 0.003496$$

As the number of believed risks associated with the latent disease increases, we also expect incidence of symptoms to increase.

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.

Solution:

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.

Solution:

$$sensitivity = \frac{TP}{TP+FN} = \frac{72}{75} = 0.960$$

$$specificity = \frac{TN}{TN+FP} = \frac{71}{83} = 0.855$$

$$prevalence = \frac{diseased}{population} = \frac{75}{158} = 0.475$$

$$PV^{+} = \frac{TP}{TP + FP} = \frac{72}{84} = 0.857$$

$$PV^{-} = \frac{TN}{TN + FN} = \frac{71}{74} = 0.959$$

(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.

Solution:

$$p(A) = \frac{prevalence_{intravenous}}{25} = 0.019$$

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

Solution:

Where R indicates the PCR test random variable,

$$p(A|R) = \frac{p(A) \cdot sensitivity}{p(A) \cdot sensitivity + (1 - p(A)) \cdot (1 - specificity)} = 0.114 \quad (1)$$

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

Solution:

$$p(A|\neg R) = \frac{p(A) \cdot (1 - sensitivity)}{p(A) \cdot (1 - sensitivity) + (1 - p(A)) \cdot specificity} = 9.04 \times 10^{-4}$$

(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?

Solution:

Using equation 1, to the plotting machine!

I'm going with true negative rate, although I was very lenient regarding grading if your rationale made sense. Note that the key point is that p(A) is a rare event.

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

$$cos(\theta) = \frac{d_1 \cdot d_2}{||d_1||||d_2||} \quad (2)$$

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

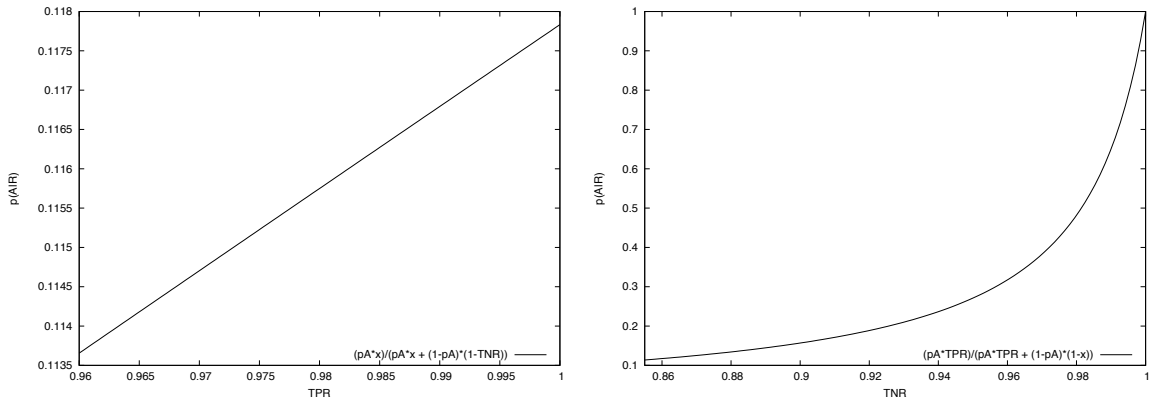


Figure 3: Comparing improving TPR and TNR w.r.t. post-test probability

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 2) 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?

(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).

SOLUTION

a. Define a feature-space F_{MeSH} that contains an entry for every MeSH term. Map each document d_i to this space as follows. Given a node n from d_i first set the value at the index corresponding to n to 1. Next, trace n up to a root, setting the corresponding parent node's indices to 1 along the way. The desired similarity function is computed as the usual cosine similiarity in this space. One drawback to this approach is that it will substantially blow-up the feature space size. A second, more important, drawback is that matching for 'distant' or generic features/terms (e.g., 'Person') carries the same weight as matching for more specific terms (e.g., 'Alcoholic').

b. To exploit the aforementioned hierarchical relationships between terms in documents, we first introduce the notion of MeSH-distance between two terms, t_1 and t_2. Define this to be the minimum path in the ontological tree from t_1 and t_2 (this can be readily calculated from $\mathcal{D}$). If no connecting path exists, then the distance is infinite.

[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.

Next, we introduce a function $\mathcal{R}$ that maps pairs of terms to 'scores' between 0 and 1 as a function of their MeSH-distance (note that we would need to normalize values to fall in the 0-1 range). For example, we may define $\mathcal{R}$ to relate similarity to distance linearly (i.e., linear in MeSH-distance). Alternatively, if we are more concerned about specificity of matches, we might define $\mathcal{R}$ to relate similarity to MeSH distance exponentially, so that terms with a shared parent are considered exponentially more similar than those with only a shared grand-parent (and so on).

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	50
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?

SOLUTION

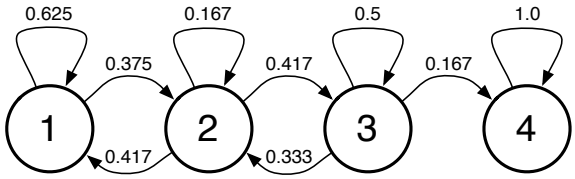
a. Yes – the aim is to model progression over time, and one can see from the observation table that the first-order Markovian property is obeyed, i.e., the state next transitioned into depends only on the state you are currently in.

b. No. The test is infallible and available; there are no latent states.

c.

state	1	2	3	4
1	.625	.375	0	0
2	.417	.167	.417	0
3	0	.333	.5	.167
4	0	0	0	1

d.



e.

```
cur_state = numpy.array([0.0, 1.0, 0.0, 0.0])
T = numpy.array([[.625, .375, 0.0, 0.0],
         [.417, .167, .417, 0.0],
         [0, .3 33, .5, .167],
         [0.0,0.0,0.0,1.0]])


for i in xrange(5):
  cur_state = numpy.dot(cur_state, T)

print cur_state
> array([ 0.31394907,  0.24952383,  0.23504111,  0.20351083])
```

So there's a pretty good chance for survival!

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

[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.

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

5.

a. We would use an HMM to account for uncertainty.

b.

```
import scikits.learn
from scikits.learn import hmm

###
# here's our data
observations = \
        [['a', 'a', 'b', 'b', 'c', 'c', 'd', 'd'],
         ['b', 'b', 'c', 'c', 'b', 'b', 'b', 'b'],
         ['a', 'a', 'a', 'a', 'a', 'a', 'a', 'a'],
         ['c', 'c', 'c', 'd', 'd', 'd', 'd', 'd'],
         ['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', 'd', 'd', 'd'],
         ['a', 'a', 'a', 'a', 'a', 'a', 'a', 'a'],
```

```
        ['c', 'c', 'c', 'd', 'd', 'd', 'd', 'd'],
        ['a', 'a', 'b', 'b', 'b', 'b', 'a', 'a'],
        ['b', 'b', 'b', 'b', 'c', 'c', 'c', 'd'],
        ['a', 'a', 'a', 'a', 'a', 'a', 'a', 'a'],
        ['b', 'b', 'b', 'a', 'a', 'a', 'b', 'b'],
        ['c', 'd', 'd', 'd', 'd', 'd', 'd', 'd'],
        ['a', 'b', 'b', 'b', 'b', 'a', 'a', 'a']]


# the scikits implementation wants observations to
# be numeric (even in the multinomial/symbolic case)
to_ints = lambda x: {'a':0, 'b':1, 'c':2, 'd':3}[x]
obs_ints = [[to_ints(x_i) for x_i in x] for x in obs]


# initialize our model; recall that we have four states
model= hmm.MultinomialHMM(n_states=4)
# set our alphabet size (bizarrely, the constructor doesnt
# take this parameter. oh well.)
model.n_symbols=4

# fit the model to the observations
fitted_model = model.fit(obs_ints)

# here's the transition matrix
print fitted_model.transmat

### output
'''
    [['0.889779', '0.025944', '0.084277', '0.000000'],
     ['0.000000', '0.705935', '0.196148', '0.097918'],
     ['0.067287', '0.118990', '0.813722', '0.000000'],
     ['0.000000', '0.037437', '0.013948', '0.948615']]
'''

print fitted_model.start_probs
'''
    ['0.176084', '0.529756', '0.294155', '0.000005']
'''
# notice above that it's taken 1 as the start state (i.e., what we labeled
# as state 0). here are the emission probabilities

[['0.000000', '0.000000', '0.439408', '0.560592'], # death state (4)
 ['0.969281', '0.030719', '0.000000', '0.000000'], # start state (0)
 ['0.000011', '0.862241', '0.137748', '0.000000'], # 1
 ['0.972983', '0.027017', '0.000000', '0.000000']] # 2
```

c. We could improve this by putting a prior over state transitions reflecting the estimates calculated from the data above, ie., for the original disease variant.

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 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 4 represents a plot for the two seemingly relevant dimensions for diagnosing 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.

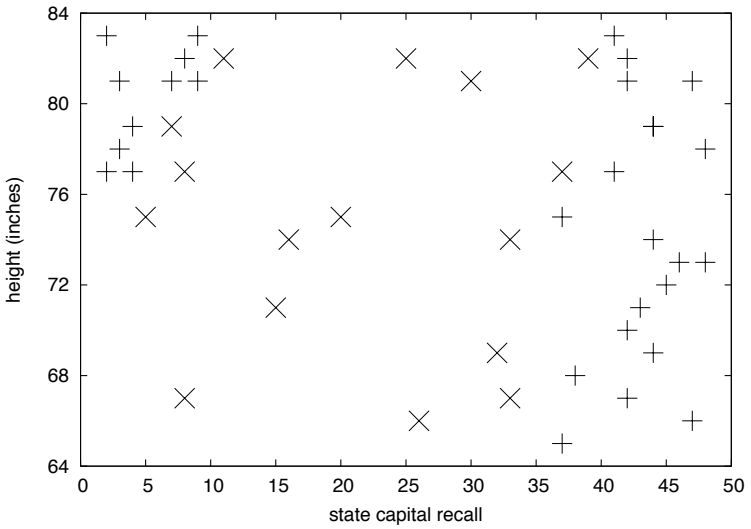


Figure 4: Study to generate diagnostic test for Smallitus

Solution:

Obviously, there are multiple possible solutions here. I am going to present one possible solution. Figure 5(a) minimizes error while Figure 5(b) emphasizes generalization.

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

Solution:

$$sensitivity(a) = \frac{TP}{TP + FN} = \frac{29}{29} = 1.0$$

$$specificity(a) = \frac{TN}{TN + FP} = \frac{15}{15} = 1.0$$

I am going to assume that we label the ambiguous node as completely positive (just by majority vote)

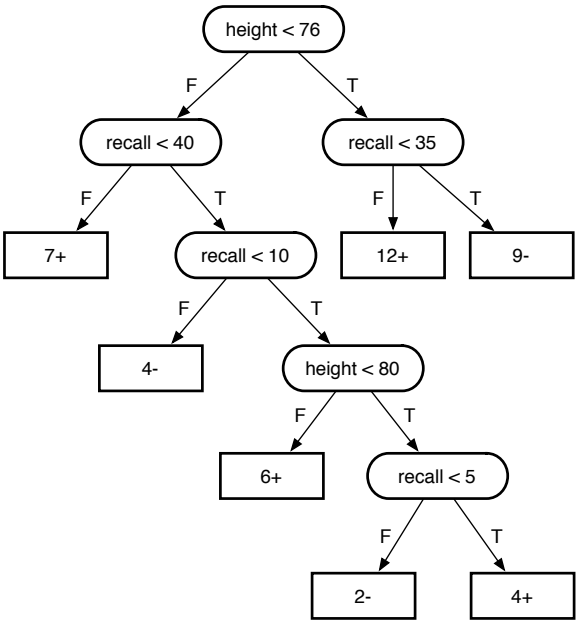
$$sensitivity(b) = \frac{TP}{TP + FN} = \frac{29}{29} = 1.0$$

$$specificity(b) = \frac{TN}{TN + FP} = \frac{13}{15} = 0.867$$

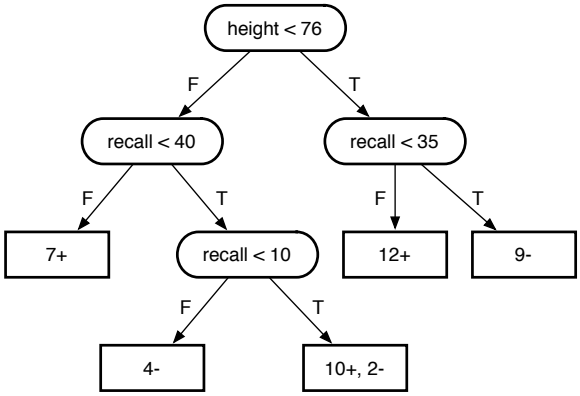
(c) (3 points) Briefly discuss how an asymmetric utility function (with respect to false positives and false negatives) might changes the desired decision tree (you don't have to actually generate a new decision tree).

Solution:

If the classes have sufficiently asymmetric costs, we allow errors of a specific type to occur with greater frequency. I would expect this to particularly be the case in the state capital recall range of 35-40.



(a) Decision tree which minimizes error on training data



(b) Decision tree which emphasizes generalization to new data

(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?

Solution:

Yes, it is a reasonable classifier as a hierarchy of axis-parallel thresholding units describes the data fairly well. Secondly, the results are interpretable. Many other classifiers may be appropriate, but the key is that we require a non-linear function to accurately classify the data (although a linear function learned in a higher dimensional projection may be suitable).