15-781 Final Exam Solutions, Fall 2002

1. Write your name and your andrew email address below.

   Name:

   Andrew ID:

2. There should be 17 pages in this exam (excluding this cover sheet).
3. If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
4. You should attempt to answer all of the questions.
5. You may use any and all notes, as well as the class textbook.
6. All questions are worth an equal amount. They are not all equally difficult.
7. You have 3 hours.
8. Good luck!

1 Computational Learning Theory

1.1 PAC learning for Decison Lists

A decision list is a list of if-then rules where each condition is a literal (a variable or its negation). It can be thought of as a decision tree with just one path. For example, say that I like to go for a walk if it's warm or if it's snowing and I have a jacket, as long as it's not raining. We could describe this as the following decision list:

```
if rainy then no
else if warm then yes
else if not(have-jacket) then no
else if snowy then yes
else no.
```

(a) Describe an algorithm to learn DLs given a data set, for example

a	b	c	class
1	0	0	+
0	1	1	-
1	1	1	+
0	0	0	-
1	1	0	+

Your algorithm should have the characteristic that it should always classify examples that it has already seen correctly (ie, it should be consistent with the data).If it's not possible to continue to produce a decision list that's consistent with the data, your algorithm should terminate and announce that it has failed.

A. Find a rule consistent with the current set of examples that explains at least one of them. If no such rule exists, halt with failure.

B. Add rule to end of decision list

C. Remove examples classified by the decision list so far

D. Repeat until no examples are left

(b) Find the size of the hypothesis space, $|H|$, for decisions lists of k attributes of depth d.

$(4k)^d$ or $\binom{2k}{d}d!2^d$ depending on whether you allowed attributes to be reused.

(c) Find an expression for the number of examples needed to learn a decision list of k attributes with error at most .10 with probability 90%.

$$m \geq \frac{1}{.10}(ln|H| + ln(\frac{1}{1-.90})) \tag{1}$$

(d) What if the learner is trying to learn a decision list, but the representation that it is using is a conjunction of k literals? Find the expression for the number of examples needed to learn the decision list with error at most .10 with 90% probability.

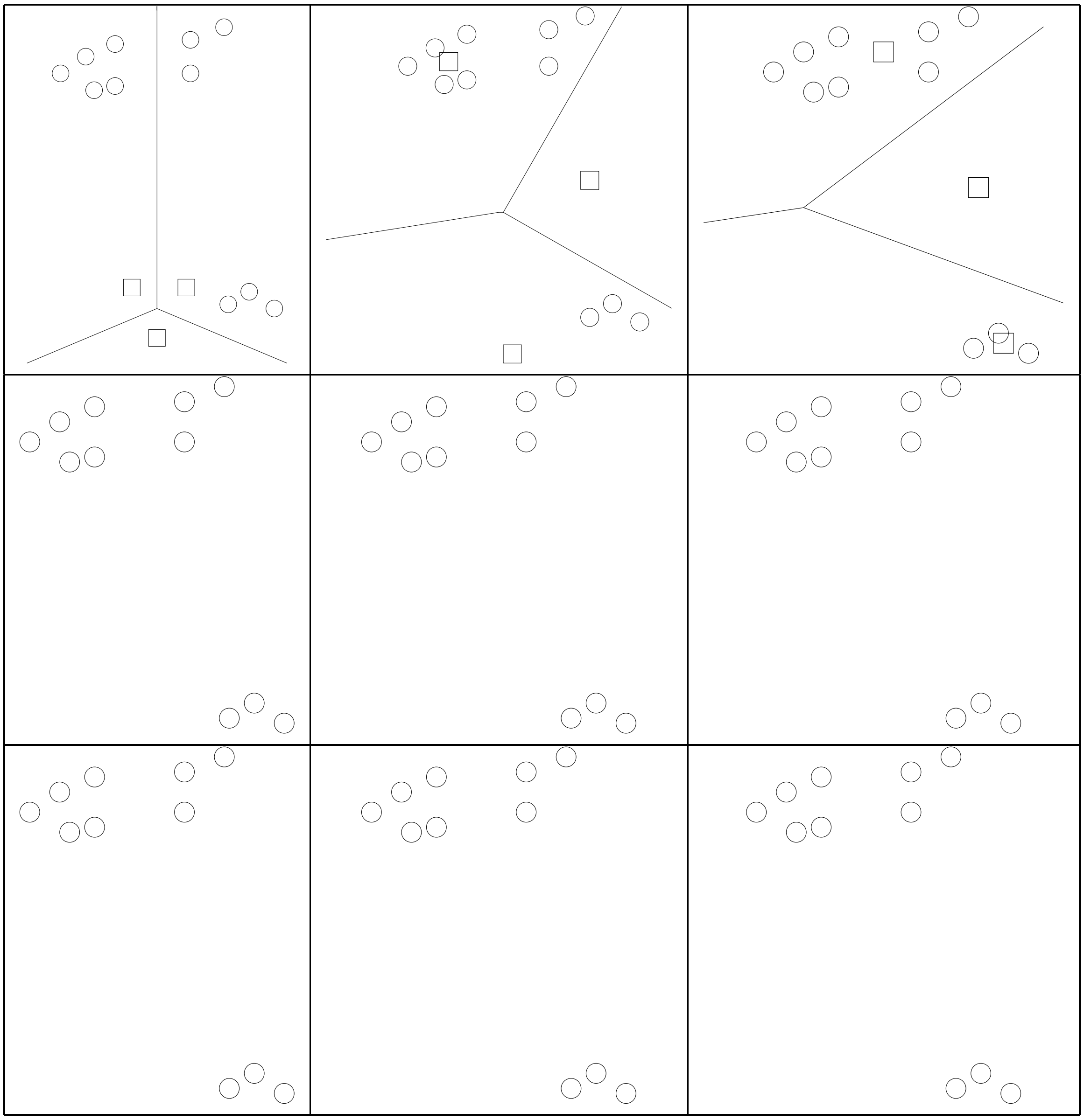
I had intended the answer for this question to require you to use the PAC formula for agnostic learning. However, it seems that people interpreted this qustion in a variety of different ways. As long as the size of the hypothesis space you gave made sense and you used either PAC bound correctly, you received full credit.

2 K-means and Gaussian Mixture Models

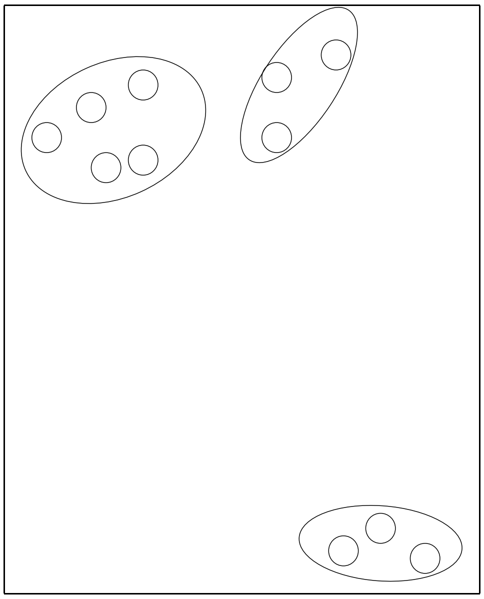
(a) What is the effect on the means found by k-means (as opposed to the true means) of overlapping clusters?

The means found will be farther apart than the true means.

(b) Run k-means manually for the following dataset. Circles are data points and squares are the initial cluster centers. Draw the cluster centers and the decision boundaries that define each cluster. Use as many pictures as you need until convergence.
Note: Execute the algorithm such that if a mean has no points assigned to it, it stays where it is for that iteration.



(c) Now draw (approximately) what a Gaussian mixture model of three gaussians with the same initial centers as for the k-means problem would converge to. Assume that the model puts no restrictions on the form of the covariance matrices and that EM updates both the means and covariance matrices.

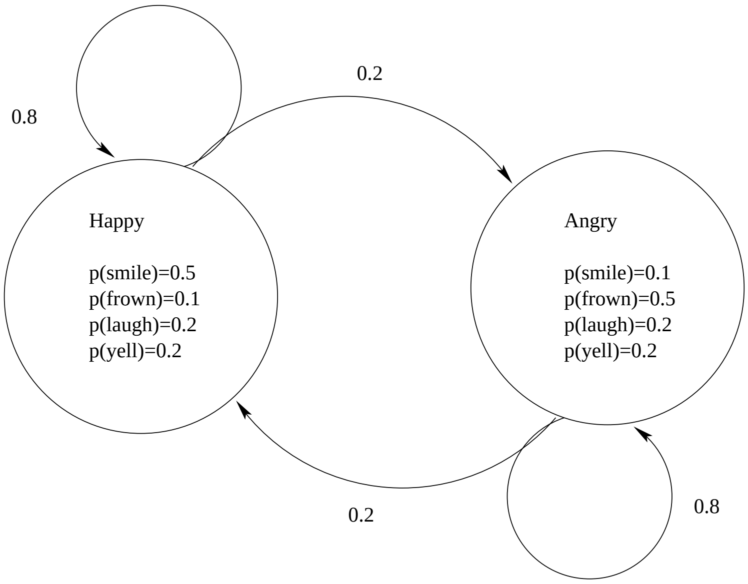


(d) Is the classification given by the mixture model the same as the classification given by k-means? Why or why not?

No. In the mixture model, soft associations (through the weights) are made with every data point by every gaussian, so it can't happen that the cluster center isn't associated with any data point. It was also ok to point out that the algorithms use different distance metrics, or that mixture models with full covariance matrices allow more flexibility in fitting a cluster. It wasn't enough to state the result of each algorithm on the example data (that doesn't say anything about why the result was like that).

3 HMMs

Andrew lives a simple life. Some days he's Angry and some days he's Happy. But he hides his emotional state, and so all you can observe is whether he smiles, frowns, laughs, or yells. We start on day 1 in the Happy state, and there's one transition per day.



Definitions:
q_t = state on day t.
O_t = observation on day t.

(a) What is $P(q_2 = Happy)$?

.8

(b) What is $P(O_2 = frown)$?

.18

(c) What is $P(q_2 = Happy | O_2 = frown)$?

.44

(d) What is $P(O_{100} = yell)$?

.2

(e) Assume that $O_1 = frown$, $O_2 = frown$, $O_3 = frown$, $O_4 = frown$, and $O_5 = frown$. What is the most likely sequence of states?

HAAAA

4 Bayesian Inference

(a) Consider a dataset over 3 boolean attributes, X, Y, and Z.

Of these sets of information, which are sufficient to specify the joint distribution? Circle all that apply.

A. $P(\sim X|Z)$ $P(\sim X|\sim Z)$ $P(\sim Y|X \wedge Z)$ $P(\sim Y|X\wedge \sim Z)$

$P(\sim Y|\sim X \wedge Z)$ $P(\sim Y|\sim X\wedge \sim Z)$ $P(Z)$

B. $P(\sim X|\sim Z)$ $P(X|\sim Z)$ $P(Y|X \wedge Z)$ $P(Y|X\wedge \sim Z)$

$P(Y|\sim X \wedge Z)$ $P(Y|\sim X\wedge \sim Z)$ $P(Z)$

C. $P(X|Z)$ $P(X|\sim Z)$ $P(Y|X \wedge Z)$ $P(Y|X\wedge \sim Z)$

$P(Y|\sim X \wedge Z)$ $P(\sim Y|\sim X\wedge \sim Z)$ $P(\sim Z)$

D. $P(X|Z)$ $P(X|\sim Z)$ $P(Y|X \wedge Z)$ $P(Y|X\wedge \sim Z)$

$P(\sim Y|\sim X\wedge \sim Z)$ $P(Y|\sim X\wedge \sim Z)$ $P(Z)$

A and C

Given this dataset of 16 records:

A	B	C
0	0	1
0	0	1
0	0	1
0	1	0
0	1	1
0	1	1
0	1	1
0	1	1
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	1	0
1	1	0
1	1	1

(b) Write down the probabilities needed to make a joint density bayes classifier

$$\begin{array}{cc}
P(C) = \frac{1}{2} & P(\sim C) = \frac{1}{2} \\
P(A \wedge B|C) = \frac{1}{8} & P(A \wedge B| \sim C) = \frac{2}{8} \\
P(\sim A \wedge B|C) = \frac{4}{8} & P(\sim A \wedge B| \sim C) = \frac{1}{8} \\
P(A \wedge \sim B|C) = 0 & P(A \wedge \sim B| \sim C) = \frac{5}{8} \\
P(\sim A \wedge \sim B|C) = \frac{3}{8} & P(\sim A \wedge \sim B| \sim C) = 0
\end{array}$$

(c) Write down the probabilities needed to make a naive bayes classifier.

$$\begin{array}{cccc}
P(C) = frac12 & P(\sim C) = \frac{1}{2} & & \\
P(A|C) = \frac{1}{8} & P(A| \sim C) = \frac{7}{8} & P(B|C) = \frac{5}{8} & P(B| \sim C) = \frac{3}{8} \\
P(\sim A|C) = \frac{7}{8} & P(\sim A| \sim C) = \frac{1}{8} & P(\sim B|C) = \frac{3}{8} & P(\sim B| \sim C) = \frac{5}{8}
\end{array}$$

(d) Write the classification that the joint density bayes classifier would make for C given A=0,B=1.

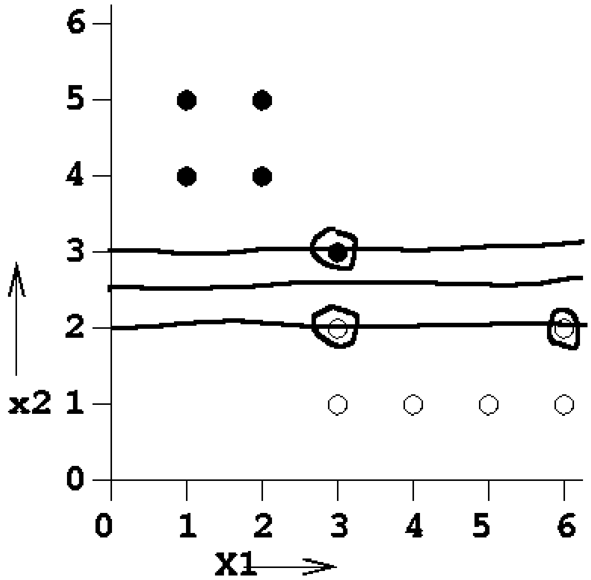
$P(C| \sim A \wedge B) = \frac{4}{5}$
$C = 1$

(e) Write the classification that the naive bayes classifier would make for C given A=0,B=1.

$P(C| \sim A \wedge B) = \frac{35}{38}$
$C = 1$

5 Support Vector Machines

This picture shows a dataset with two real-valued inputs (x_1 and x_2) and one categorical output class. The positive points are shown as solid dots and the negative points are small circles.



(a) Suppose you are using a linear SVM with no provision for noise (i.e. a Linear SVM that is trying to maximize its margin while ensuring all datapoints are on their correct sides of the margin). Draw three lines on the above diagram, showing the classification boundary and the two sides of the margin. Circle the support vector(s).

(b) Using the familiar LSVM classifier notation of class = sign($\mathbf{w}.\mathbf{x} + b$), calculate the values of $\mathbf{w}$ and b learned for part (a)

$w_1 = 0$
w_2 and b such that

$$\begin{aligned} 2w_2 + b &= -1 \\ 3w_2 + b &= 1 \\ &\cdots \\ w_2 &= 2 \\ b &= -5 \end{aligned}$$

(c) Assume you are using a noise-tolerant LSVM which tries to minimize

$$\frac{1}{2}\mathbf{w}.\mathbf{w} + C\sum_{k=1}^{R} \epsilon_k \tag{2}$$

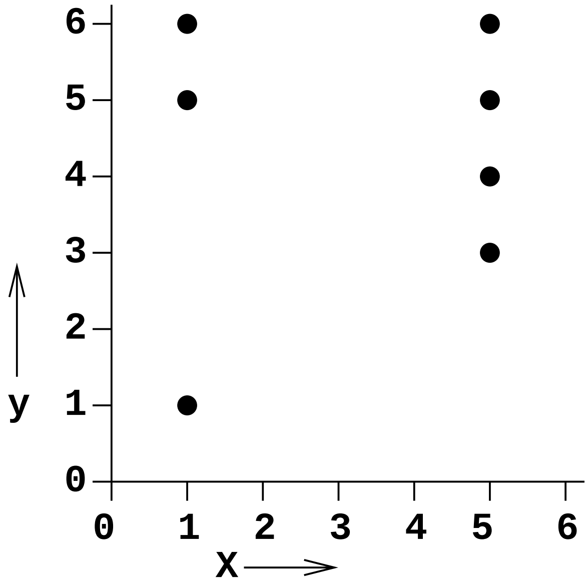
using the notation of your notes and the Burges paper.

Question: is it possible to invent a dataset and a positive value of C in which (a) the dataset is linearly separable but (b) the LSVM would nevertheless misclassify at least one training point? If it is possible to invent such an example, please sketch the example and suggest a value for C. If it is not possible, explain why not.

It is possible. Consider the 1D data set with one point at $x = 0$ and the other point at $x = 1$. If $C = 10^{-10}$, the solution will put $w = 0$ (the margin will be infinitely wide).

6 Instance-based learning

This picture shows a dataset with one real-valued input x and one real-valued output y. There are seven training points.



Suppose you are training using kernel regression using some unspecified kernel function. The only thing you know about the kernel function is that it is a monotonically decreasing function of distance that decays to zero at a distance of 3 units (and is strictly greater than zero at a distance of less than 3 units).

(a) What is the predicted value of y when $x = 1$?

4

(b) What is the predicted value of y when $x = 3$?

$\frac{30}{7}$

(c) What is the predicted value of y when $x = 5$?

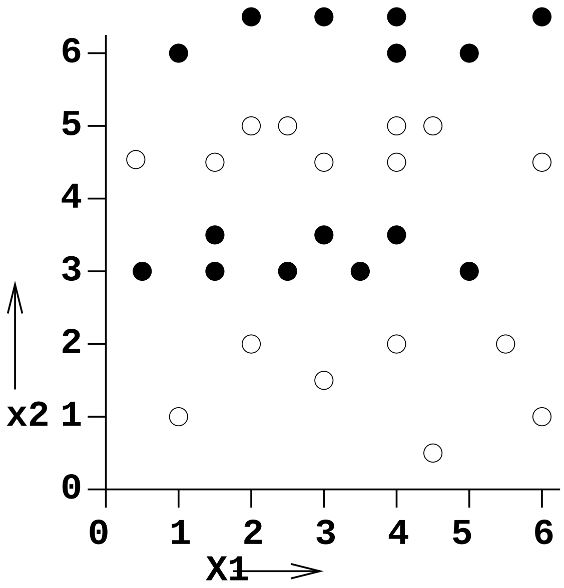
4.5

(d) What is the predicted value of y when $x = 6$?

4.5

The final two parts of this question concern 1-nearest neighbor used as a classifier.

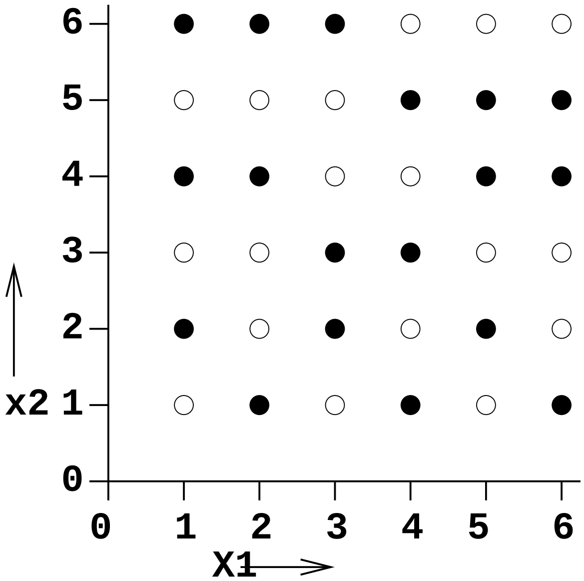
The following dataset has two real valued inputs and one binary categorical output. The class is denoted by the color of the datapoint.



(e) Does there exist a choice of Euclidian distance metric for which 1-nearest-neighbor would achieve zero training set error on the above dataset?

Yes, virtually any metric (eg, "scale axes equally") will work

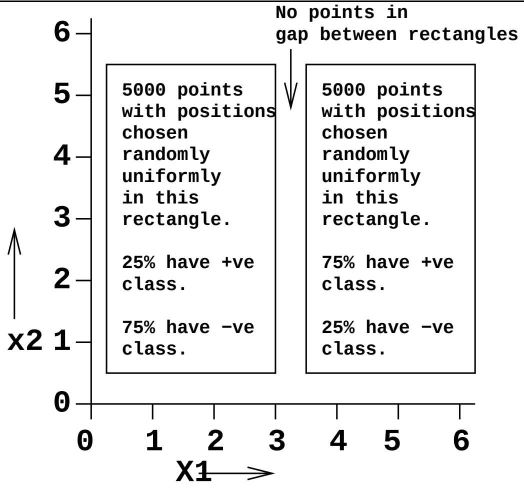
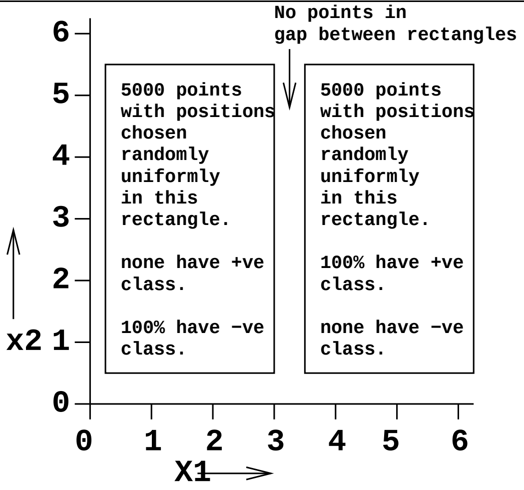
Now let's consider a different dataset:



(f) Does there exist a choice of Euclidian distance metric for which 1-nearest-neighbor would achieve zero training set error on the above dataset?

same as (e)

7 Nearest Neighbor and Cross-Validation

<table>
<tr>
<td>Recipe for making training set of 10,000 datapoints with two real-valued inputs and one binary output class:</td>
<td>Recipe for making test set of 10,000 datapoints with two real-valued inputs and one binary output class:</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
</table>

Using the above recipes for making training and test sets you will see that the training set is noisy: in either region, 25% of the data comes from the minority class. The test set is noise-free.

In each of the following questions, circle the answer that most closely defines the expected error rate, expressed as a fraction.

(a) What is the expected training set error using one-nearest-neighbor?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

0

(b) What is the expected leave-one-out cross-validation error on the training set using one-nearest-neighbor?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

3/8

(c) What is the expected test set error if we train on the training set, test on the test set, and use one-nearest-neighbor?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

1/4

(d) What is the expected training set error using 21-nearest-neighbor?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

1/4

(e) What is the expected leave-one-out cross-validation error on the training set using 21-nearest-neighbor?

`0    1/8    1/4    3/8    1/3    1/2    5/8    2/3    3/4    7/8    1`

1/4

(f) What is the expected test set error if we train on the training set, test on the test set, and use 21-nearest-neighbor?

`0    1/8    1/4    3/8    1/3    1/2    5/8    2/3    3/4    7/8    1`

0

8 Learning Bayes Net Structure

For each of the following training sets, draw the structure and CPTs that a Bayes Net Structure learner should learn, assuming that it tries to account for all the dependencies in the data as well as possible while minimizing the number of unnecessary links. In each case, your Bayes Net will have three nodes, called A B and C. Some or all of these questions have multiple correct answers...you need only supply one answer to each question.

(a)

A	B	C
0	0	0
0	0	0
0	0	1
0	0	1
0	1	0
0	1	0
0	1	1
0	1	1
1	0	0
1	0	0
1	0	1
1	0	1
1	1	0
1	1	0
1	1	1
1	1	1

All independent (nodes aren't connected).

(b)

A	B	C
0	0	0
0	0	0
0	0	0
0	1	1
0	1	1
0	1	1
1	1	0
1	1	0
1	1	0
1	0	1
1	0	1
1	0	1

A and B each have an arrow into C, or
B and C each have an arrow into A, or
A and C each have an arrow into B.

(c)

A	B	C
0	0	0
0	0	0
0	0	0
1	0	1
1	0	1
1	0	1
1	1	0
1	1	0
1	1	0

A has an arrow into B and an arrow into C. B has an arrow into C.

9 Markov Decision Processes

Consider the following MDP, assuming a discount factor of $\gamma = 0.5$. Note that the action "Party" carries an immediate reward of +10. The action "Study" unfortunately carries no immediate reward, except during the senior year, when a reward of +100 is provided upon transition to the terminal state "Employed".

[See file rlans.ps for answers drawn on diagram]

(a) What is the probability that a freshman will fail to graduate to the "Employed" state within four years, even if they study at every opportunity?

.34

(b) Draw the diagram for the Markov Process (not the MDP, the MP) that corresponds to the policy "study whenever possible."

The diagram above, only without the arrows for the "party" action from each state.

(c) What is the value associated with the state "Junior" under the "study whenever possible" policy?

40.5/.95

(d) Exactly how rewarding would parties have to be during junior year in order to make it advisable for a junior to party rather than study (assuming, of course, that they wish to optimize their cumulative discounted reward)?

Note we know from (c) the value of state "Junior" under the the "study all the time" policy. Therefore, we can answer this question by finding the value of party reward, R, that makes the value of the state higher under the "party when possible, study otherwise" policy.

$R \geq 9.47$

(e) Answer the following true or false. If true, give a one-sentence argument. If false, give a counterexample.

- **(True or False?)** If partying during junior year an optimal action when it is assigned reward r, then it will also be an optimal action for a freshman when assigned reward r.

  True, because the reward for studying will be discounted further into the future for freshman than for juniors.

- **(True or False?)** If partying during freshman year is an optimal action when it is assigned reward r, then it will also be an optimal action for a junior when assigned reward r.

  False. A party reward of 9 is sufficient to make it optimal during freshman year, but not during senior year.

10 Q Learning

Consider the robot grid world shown below, in which actions have deterministic outcomes, and for which the discount factor $\gamma = 0.5$. The robot receives zero immediate reward upon executing its actions, except for the few actions where an immediate reward has been written in on the diagram. Note the state in the upper corner allows an action in which the robot remains in that same state for one time tick.

IMPORTANT: Notice the immediate reward for the state-action pair $< C, South >$ is -100, not +100.

(a) Write in the Q value for each state-action pair, by writing it next to the corresponding arrow.

(b) Write in the $V^*(s)$ value for each state, by writing its value inside the grid cell representing that state.

(c) Write down an equation that relates the $V^*(s)$ for an arbitrary state s to the $Q(s, a)$ values associated with the same state.

$$V^*(s) = max_i Q(s, a) \tag{3}$$

(d) Describe one optimal policy, by circling only the actions recommended by this policy

(e) Hand execute the deterministic Q learning algorithm, assuming the robot follows the trajectory shown below. Show the sequence of Q estimates (describe which entry in the Q table is being updated at each step):

state	action	next-state	immediate-reward	updated-Q-estimates
A	East	B	0	Q(a,E)=0
B	East	C	10	Q(b,E)=10
C	Loop	C	0	Q(c,L)=0
C	South	F	-100	Q(c,S)=-100
F	West	E	0	Q(f,W)=0
E	North	B	0	Q(e,N)=5
B	East	C	10	Q(b,E)=10

(f) Propose a change to the immediate reward function that results in a change to the Q function, but not to the V function.

change r(b,E) from 10 to 9 or change r(c,S) from -100 to -101

11 Short Questions

(a) Describe the difference between a *maximum likelihood* hypothesis and a *maximum a posteriori* hypothesis.

The maximum likelihood hypothesis maximizes the probability of the data given the hypothesis. The maximum a posteriori hypothesis maximizes the probability of the hypothesis given the data.

(b) Consider a learning problem defined over a set of instances X. Assume the space of possible hypotheses, H, consists of all possible disjunctions over instances in X. I.e., the hypothesis $x_1 \vee x_6$ labels these two instances positive, and no others. What is the VC dimension of H?

$|X|$

(c) Consider a naive Bayes classifier with 2 boolean input variables, X and Y, and one boolean output, Z.

- Draw the equivalent Bayesian network.

  Z has arrows pointing at X and Y.

- How many parameters must be estimated to train such a naive Bayes classifier?

  5

- How many parameters would have to be estimated if the naive Bayes assumption is not made, and we wish to learn the Bayes net for the joint distribution over X, Y, and Z?

  8

True or False? If true, explain why in at most two sentences. If false, explain why or give a brief counterexample.

- **(True or False?)** The error of a hypothesis measured over the training set provides a pessimistically biased estimate of the true error of the hypothesis.

  F. Would be true if it said "optimistic".

- **(True or False?)** Boosting and the Weighted Majority algorithm are both methods for combining the votes of multiple classifiers.

  T

- **(True or False?)** Unlabeled data can be used to detect overfitting.

  T

- **(True or False?)** Gradient descent has the problem of sometimes falling into local minima, whereas EM does not.

  F. EM does have this problem.

- **(True or False?)** HMM's are a special case of MDP's.

  F. MDP's assume fully observable state.