15-781 Midterm, Fall 2002

YOUR ANDREW USERID IN CAPITAL LETTERS:

YOUR NAME:

- There are 5 questions.
- Questions 1-5 are worth 20 points each.
- The maximum possible total score is 100.
- Unless otherwise stated there is no need to show your working.

1 Decision Trees (20 points)

Master Yoda is concerned about the number of Jedi apprentices that have turned to the Dark Side, so he's decided to train a decision tree on some historical data to help identify problem cases in the future. The following table summarizes whether or not each of 12 initiates turned to the Dark Side based on their age when their Jedi training began, whether or not they completed their training, their general disposition, and their species.

Dark Side	Age Started Training	Completed Training	Disposition	Species
0	5	1	Happy	Human
0	9	1	Happy	Gungan
0	6	0	Happy	Wookiee
0	6	1	Sad	Mon Calamari
0	7	0	Sad	Human
0	8	1	Angry	Human
0	5	1	Angry	Ewok
1	9	0	Happy	Ewok
1	8	0	Sad	Human
1	8	0	Sad	Human
1	6	0	Angry	Wookiee
1	7	0	Angry	Mon Calamari

(a) *(3 points)* What is the initial entropy of *Dark Side*?

(b) *(3 points)* Which attribute would the decision-tree building algorithm choose to use for the root of the tree?

(c) *(3 points)* What is the information gain of the attribute you chose to split on in the previous question?

(d) *(3 points)* Draw the full decision tree that would be learned for this data (with no pruning). *(Hint: The tree will have no more than three splits. The correct split at each point should be clear from just the groups it splits the data into, without having to actually compute the information gain for each possible split.)*

(e) *(2 points)* Consider the possibility that the input data above is noisy and not completely accurate, so that the decision tree you learned may not accurately reflect the function you want to learn. If you were to evaluate the three initiates represented by the data points below, on which one would you be most confident of your prediction, and why?

Name	Age Started Training	Completed Training	Disposition	Species
Ardath	5	0	Angry	Human
Barbar	8	0	Angry	Gungan
Caldar	8	0	Happy	Mon Calamari

(f) *(3 points)* Assume we train a decision tree to predict Z from A, B, and C using the following data (with no pruning):

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

What would be the training set error for this dataset? Express your answer as the number of records out of 12 that would be misclassified.

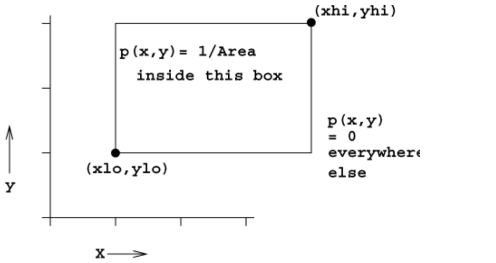
(g) *(3 points)* Consider a decision tree built from an arbitrary set of data. If the output is discrete-valued and can take on k different possible values, what is the maximum training set error (expressed as a fraction) that any data set could possibly have?

2 Probability and Bayes Classifiers (20 points)

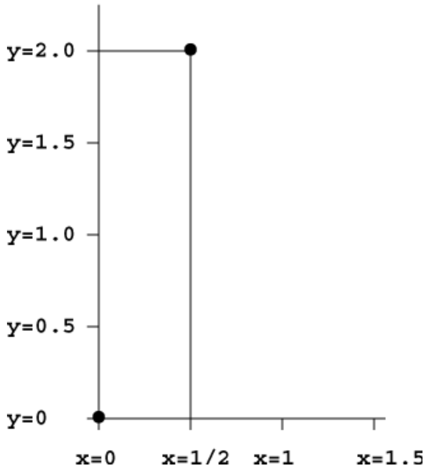
This figure illustrates a simple class of probability density functions over pairs of real-valued variables. We call it the Rectangle PDF.

$(x, y) \sim Rect(x_{lo}, y_{lo}, x_{hi}, y_{hi})$
means

$$p(x,y) \quad = \frac{1}{(x_{hi}-x_{lo})(y_{hi}-y_{lo})} \text{ if } x_{lo} \leq x \leq x_{hi} \text{ and } y_{lo} \leq y \leq y_{hi}$$
$$= 0 \text{ otherwise}$$



(a) *(2 points)* Assuming $(x, y) \sim Rect(0, 0, 0.5, 2)$ (as shown in the diagram to the right), compute the value of the density $p(x = \frac{1}{4}, y = \frac{1}{4})$



(b) *(3 points)* Under the same assumptions, compute the density $p(y = \frac{1}{4})$

(c) *(3 points)* Under the same assumptions, compute the density $p(x = \frac{1}{4})$

(d) *(3 points)* Under the same assumptions, compute the density $p(x = \frac{1}{4} | y = \frac{1}{4})$

Maximum Likelihood Estimation of Rectangles

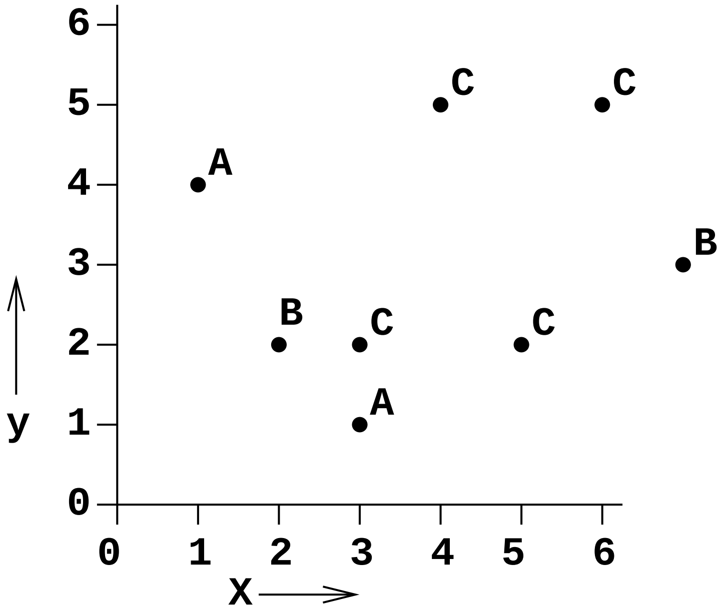
Assume we have R datapoints $(x_1, y_1), (x_2, y_2)...(x_R, y_R)$ where each datapoint is drawn independently from $Rect(x_{lo}, y_{lo}, x_{hi}, y_{hi})$

Suppose we want to find the MLE parameters $(x_{lo}, y_{lo}, x_{hi}, y_{hi})$ that maximize the likelihood of the datapoints. It turns out (no proof given or required) that these MLE values define the bounding box of the datapoints:

$x_{lo}^{MLE} = min_k x_k$
$y_{lo}^{MLE} = min_k y_k$
$x_{hi}^{MLE} = max_k x_k$
$y_{hi}^{MLE} = max_k y_k$

Now, suppose that we use the rectangle distribution as the density estimator for each class of a Bayes Classifier that we're about to learn. The data is

x	y	Class
1	4	A
3	1	A
2	2	B
7	3	B
3	2	C
4	5	C
5	2	C
6	5	C



Assuming we use the Rectangle Bayes Classifier learned from the data, what value will the classifier give for:

(e) *(3 points)* $P(Class = A|x = 1.5, y = 3)$

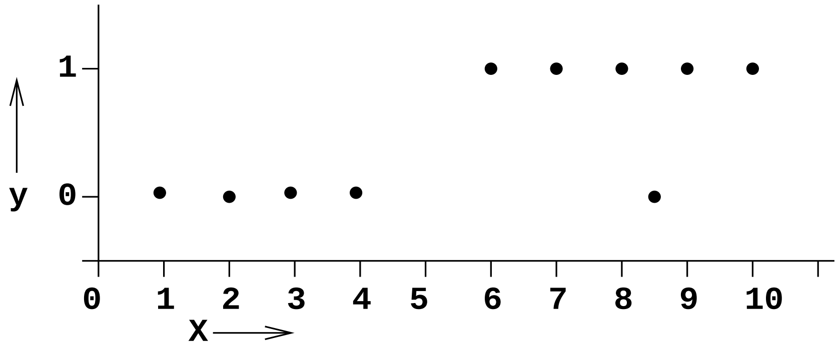
(f) *(3 points)* $P(Class = A|x = 2.5, y = 2.5)$

(g) *(3 points)* $P(Class = A|y = 5)$

3 Cross Validation (20 points)

Suppose we are learning a classifier with binary output values Y=0 and Y=1. There is one real-valued input X. Here is our data:

X	Y
1	0
2	0
3	0
4	0
6	1
7	1
8	1
8.5	0
9	1
10	1



Assume we will learn a decision tree on this data. Assume that when the decision tree splits on the real valued attribute x, it puts the split threshold halfway between the attributes that surround the split. For example, using information gain as the splitting criterion, the decision tree would initially choose to split at x = 5, which is halfway between the x = 4 and x = 6 datapoints.

Let Algorithm DT2 be the method of learning a decision tree with only two leaf nodes (i.e. only one split).

Let Algorithm DT* be the method of learning a decision tree fully with no pruning.

(a) *(5 points)* What will be the training set error of DT2 on our data? In this part, and all future parts, you can express your answer as the number of misclassifications out of 10.

(b) *(5 points)* What will be the leave-one-out-cross-validation error of DT2 on our data?

(c) *(5 points)* What will be the training set error of DT* on our data?

(d) *(5 points)* What will be the leave-one-out-cross-validation error of DT* on our data?

4 Computational learning theory (20 points)

True or false: For a-d, if false, give a counter example. If true, give a 1 sentence justification.

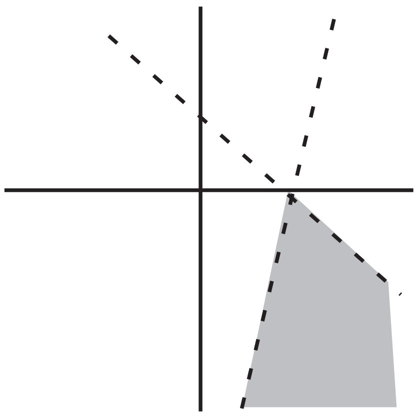
(a) *(3 points)* Within the setting of the PAC model it is impossible to assure with probability 1 that the concept will be learned perfectly (i.e., with true error=0), regardless of how many training examples are provided.

(b) *(3 points)* If the Halving Algorithm has made exactly $\lfloor log_2|H| \rfloor$ mistakes, and H contains the target concept, then it must have learned a hypothesis with true error=0, regardless of what training sequence we presented and what hypothesis space H it considered.

(c) *(3 points)* It is impossible for the Halving Algorithm to learn any concept without it making at least VC(H) mistakes, regardless of what training sequence we present, and what hypothesis space H it considers.

(d) *(3 points)* The PAC bounds make a worst case assumption about the probability distribution over the instances X, but it is possible to learn from fewer examples for some distributions over X.

Consider the class of concepts H2p defined by conjunctions of two arbitrary perceptrons. More precisely, each hypothesis $h(x) : X \rightarrow \{0,1\}$ in H2p is of the form $h(x) = p_1(x)$ AND $p_2(x)$, where $p_1(x)$ and $p_2(x)$ are any **two-input** perceptrons. The figure illustrates one such possible classfier in two dimensions.



(e) *(4 points)* Draw a set of three points in the plane that cannot be shattered by H2p.

(f) *(4 points)* What is the VC dimension of H2p? (Partial credit will be given if you can bound it, so show your reasoning!)

5 Regression and neural networks (20 points)

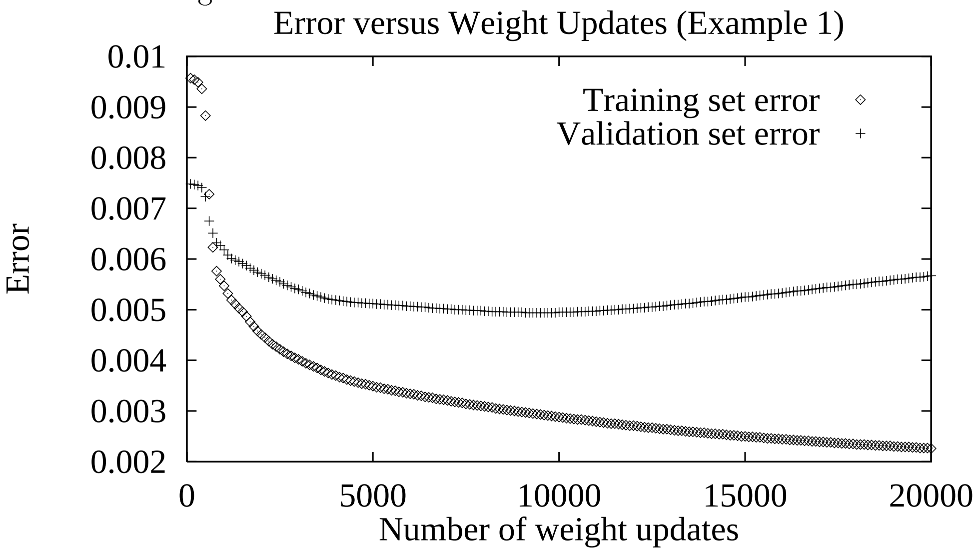
(a) *(8 points)* Derive a gradient descent training algorithm that minimizes the sum of squared errors for a variant of a perceptron where the output o of the unit depends on its inputs x_i as follows:

$o = w_0 + w_1 x_1 + w_1 x_1^3 + w_2 x_2 + w_2 x_2^3 + ... + w_n x_n + w_n x_n^3$

Give your answer in the form $w_i \leftarrow w_i + \ldots$ for $1 \leq i \leq n$. You do not need to give the update rule for w_0.

(Question continues on next page)

Consider the following plot showing training set error and validation set error for the Backpropagation algorithm training a neural network for a particular medical diagnosis problem. Note that the training error decreases monotonically with increasing gradient descent steps, whereas the validation error does not. Suppose now that we were to retrain the same neural network using exactly the same algorithm, but using ten times as much training data.



(b) *(6 points)* Would you expect the training curve to be different? If so, draw what you would expect. In either case, explain your reasoning in **at most** three sentences.

(c) *(6 points)* Would you expect the validation set curve to be different? If so, draw what you would expect. In either case, explain your reasoning in **at most** three sentences.