15-781 Midterm, Fall 2003

YOUR ANDREW USERID IN CAPITAL LETTERS:

YOUR NAME:

- There are 9 questions. The ninth may be more time-consuming and is worth only three points, so do not attempt 9 unless you are completely happy with the rest of your answers.
- The maximum possible total score is 100.
- Unless otherwise stated there is no need to show your working.

1 Decision Trees (16 points)

The following dataset will be used to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, color and odor.

Shape	Color	Odor	Edible
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

(a) *(4 points)* What is entropy $H(Edible|Order = 1 \text{ or } Odor = 3)$?

$H(Edible|Order = 1 \texttt{ or } Odor = 3) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$

(b) *(4 points)* Which attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)?

`Odor`

(c) *(4 points)* Draw the full decision tree that would be learned for this data (no pruning).

```
          Odor
          / | \
        1/ 2|  \3
        /   |   \
     Yes  Color  No
          / | | \
         /  | |  \
     W /  B|  \G \U
      /   /    \  \
   Yes  Shape  No  No
         / \
       C/   \D
       /     \
     Yes     No
```

(d) *(4 points)* Suppose we have a validation set as follows. What will be the training set error and validation set error of the tree? Express your answer as the number of examples that would be misclassified.

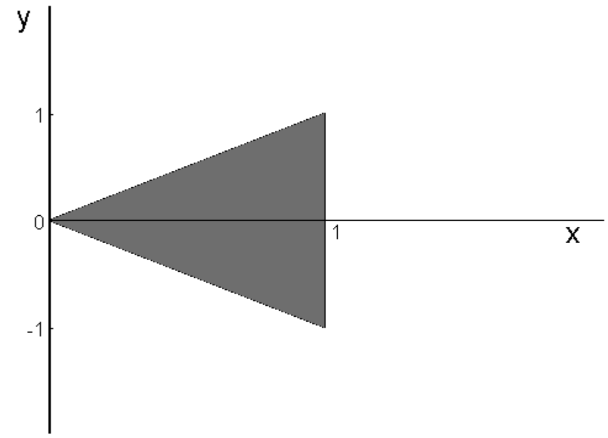
Shape	Color	Odor	Edible
C	B	2	No
D	B	2	No
C	W	2	Yes

```
training set error: 0
validation set error: 1
```

2 Probability Density Functions (8 points)

Suppose the joint Probability Density Function of a pair of random variables (x, y) is given by,

$$p(x,y) = 1 \quad |y| < x, 0 < x < 1$$
$$p(x,y) = 0 \qquad \text{otherwise}$$



(a) *(4 points)* What is $p(y|x = 0.5)$?

$$p(y|x = 0.5) = 1 \quad -0.5 < y < 0.5$$
$$p(y|x = 0.5) = 0 \qquad \texttt{otherwise}$$

(b) *(4 points)* Is x independent of y? (no explanation needed)

`Answer:  No`

3 Bayes Classifiers (12 points)

Suppose you have the following training set with three boolean input x, y and z, and a boolean output U.

x	y	z	U
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	1
0	0	1	1
0	1	0	1
1	1	0	1

Supppose you have to predict U using a naive Bayes classifier,

(a) *(3 points)* After learning is complete what would be the predicted probability

$P(U = 0|x = 0, y = 1, z = 0)$?

$$
\begin{aligned}
& P(U = 0|x = 0, y = 1, z = 0) \\
= \;& \frac{P(U = 0)P(X = 0|U = 0)P(Y = 1|U = 0)P(Z = 0|U = 0)}{P(X = 0, Y = 1, Z = 0)} \\
= \;& \frac{P(U = 0)P(X = 0|U = 0)P(Y = 1|U = 0)P(Z = 0|U = 0)}{P(U = 0)P(X = 0, Y = 1, Z = 0|U = 0) + P(U = 1)P(X = 0, Y = 1, Z = 0|U = 1)} \\
= \;& \frac{8}{35} \\
= \;& 0.229
\end{aligned}
$$

(b) *(3 points)* Using the probabilities obtained during the Bayes Classifier training, what would be the predicted probability $P(U = 0|x = 0)$?

$P(U = 0|x = 0) = \frac{1}{2}$

In the next two parts, assume we learned a Joint Bayes Classifier. In that case...

(c) *(3 points)* What is $P(U = 0|x = 0, y = 1, z = 0)$?

$P(U = 0|x = 0, y = 1, z = 0) = 0$

(d) *(3 points)* What is $P(U = 0|x = 0)$?

$P(U = 0|x = 0) = \frac{1}{2}$

4 Regression (9 points)

I have a dataset with R records in which the i^{th} record has one real-valued input attribute x_i and one real-valued output attribute y_i.

We have the following model with one unknown parameter w which we want to learn from data.

$$y_i \sim N(\exp(wx_i), 1)$$

Note that the variance is known and equal to one.

(a) *(3 points) (no explanation required)* Is the task of estimating w

A. a linear regression problem?

B. a non-linear regression problem?

`Answer: B`

(b) *(6 points) (no explanation required)* Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?

A. $\sum_i x_i \exp(wx_i) = \sum_i x_i y_i \exp(wx_i)$

B. $\sum_i x_i \exp(2wx_i) = \sum_i x_i y_i \exp(wx_i)$

C. $\sum_i x_i^2 \exp(wx_i) = \sum_i x_i y_i \exp(wx_i)$

D. $\sum_i x_i^2 \exp(wx_i) = \sum_i x_i y_i \exp(\frac{wx_i}{2})$

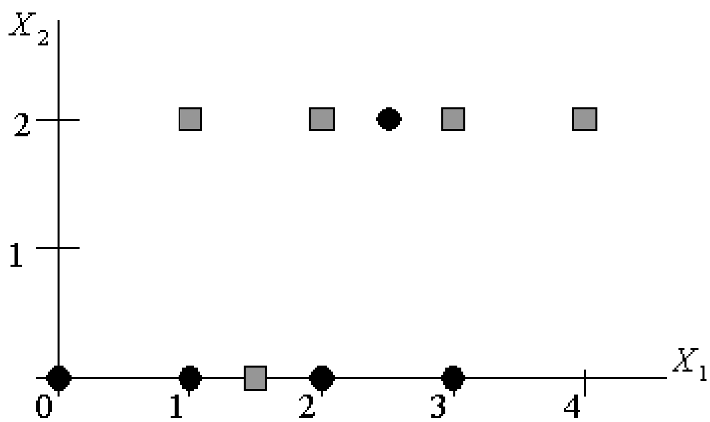
E. $\sum_i \exp(wx_i) = \sum_i y_i \exp(wx_i)$

`Answer: B`

5 Cross Validation (12 points)

Suppose we are learning a classifier with binary output values $Y = 0$ and $Y = 1$. There are two real-valued input attributes X_1 and X_2. Here is our data:

X_1	X_2	Y
0	0	0
1	0	0
2	0	0
2.5	2	0
3	0	0
1	2	1
1.5	0	1
2	2	1
3	2	1
4	2	1



Assume we will learn a decision tree using ID3 algorithm on this data. Assume that when the decision tree splits on the real-valued attributes, it puts the split threshold halfway between the values that surround the highest-scoring split location. For example, if X_2 is selected as the root attribute, the decision tree would choose to split at $X_2 = 1$, which is halfway between $X_2 = 0$ and $X_2 = 2$.

Let Algorithm DT2 be the method of learning a decision tree with only *two* leaf nodes (i.e. only one split), and Algorithm DT* be the method of learning a decision tree *fully* with no pruning.

(a) *(6 points)* What will be the training set error of DT2 and DT* on our data? Express your answer as the number of misclassifications out of 10.

```
Training set error
    DT2: 2 out of 10
    DT*: 0 out of 10
```

(b) *(6 points)* What will be the *leave-one-out* cross-validation error of DT2 and DT* on our data?

```
Leave-one-out cross validation error
    DT2: 2 out of 10
    DT*: 6 out of 10
```

6 Neural Nets (15 points)

(a) *(5 points)* Consider a single sigmoid threshold unit with three inputs, x_1, x_2, and x_3.

$$y = g(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3) \quad \text{where} \quad g(z) = \frac{1}{1 + \exp(-z)}$$

We input values of either 0 or 1 for each of these inputs. Assign values to weights w_0, w_1, w_2 and w_3 so that the output of the sigmoid unit is greater than 0.5 if an only if (x_1 AND x_2) OR x_3.

`There are many solutions. One of them is:`
$w_0 = -0.75$
$w_1 = w_2 = 0.5$
$w_3 = 1$

(b) *(10 points)* Answer the following true or false. (No explanation required).

A. One can perform linear regression using either matrix algebra or using gradient descent.

`Answer: True`

B. The error surface followed by the gradient descent Backpropagation algorithm changes if we change the training data.

`Answer: True`

C. Incremental gradient descent is always a better idea than batch gradient descent.

`Answer: False`

D. Given a two-input sigmoid unit with weights w_0, w_1, and w_2, we can negate the value of the unit output by negating all three weights.

Answer: This question is ambiguous as the meaning of "negate the value of the unit output" is not clear. On one hand, the value of the unit ouput can never be negative, on the other hand, as $\frac{1}{1+\exp(-(-x))} = 1 - \frac{1}{1+\exp(-x)}$, the value of the unit output can be "negated" according to 1. Thus answering either True or False will be deemed as correct.

E. The gradient descent weight update rule for a unit whose output is $w_0 + w_1(x_1 + 1) + w_2(x_2^2)$ is:

$$\begin{aligned}
\Delta w_0 &= \eta \sum_d (t_d - o_d) \\
\Delta w_1 &= \eta \sum_d [(t_d - o_d) x_{d1} + (t_d - o_d)] \\
\Delta w_2 &= \eta \sum_d [(t_d - o_d) 2 x_{d2}]
\end{aligned}$$

where

- t_d is the target output for the dth training example
- o_d is the unit output for the d^{th} example.
- x_{d1} is the value of x_1 for the dth training example
- x_{d2} is the value of x_2 for the dth training example

`Answer: False.`
`The formula for calculating` Δw_2 `is wrong. It should be`

$$\Delta w_2 = \eta \sum_d [(t_d - o_d) x_{d2}^2]$$

7 PAC Learning of Interval Hypotheses (15 points)

In this question we'll consider learning problems where each instance x is some integer in the set $X = \{1, 2, \ldots, 125, 126, 127\}$, and where each hypothesis $h \in H$ is an interval of the form $a \leq x \leq b$, where a and b can be any integers between 1 and 127 (inclusive), so long as $a \leq b$. A hypothesis $a \leq x \leq b$ labels instance x positive if x falls into the interval defined by a and b, and labels the instance negative otherwise. Assume throughout this question that the teacher is only interested in teaching concepts that can be represented by some hypothesis in H.

(a) *(3 points)* How many distinct hypotheses are there in H? (hint: when $b = 127$ there are exactly 127 possible values for a). (No explanation required)

$$\frac{(1 + 127) \times 127}{2} = 8128$$

(b) *(3 points)* Assume the teacher provides just one training example: x=64, label=+, then allows the student to query the teacher by generating new instances and asking for their label.

Assuming the student uses the optimal querying algorithm for this case, how many queries will they need to make? No explanation is required, you don't need to write down the optimal algorithm, and we will not be concerned if your answer is wrong by a count of one or two.

$\log_2 63 + \log_2(127 - 64) \approx 12$

(c) *(3 points)* Suppose the teacher is trying to teach the specific target concept $32 \leq x \leq 84$. What is the minimum number of training examples the teacher must present to guarantee that any consistent learner will learn this concept exactly?

`Answer: 4.`
`The training examples are:` $(a-1, -)$, $(a, +)$, $(b, +)$, $(b+1, -)$

(d) *(3 points)* Suppose now that instances are drawn at random according to a particular probability distribution $P(X)$, which is unknown to the learner. Each training example is generated by drawing an instance at random according to $P(X)$ then labeling it.

How many such training examples suffice to assure with probability 0.95 that any consistent learner will output a hypothesis whose true error is at most 0.10?

$$\begin{aligned} m &\geq \frac{1}{\epsilon}(\ln |H| + \ln \frac{1}{\delta}) \\ &= \frac{1}{0.1}(\ln(8128) + \ln(\frac{1}{1-0.95})) \\ &\approx 120 \end{aligned}$$

(e) *(3 points)* True or False (no explanation needed). In the above statement, the phrase "to assure with probability 0.95" means that if we were to run the following experiment a million times, then in roughly 950,000 cases or more, the consistent learner will output a hypothesis whose true error is at most 0.10. Each experiment involves drawing the given number of training instances at random (drawn i.i.d. from $P(X)$) and then running the consistent learner.

`Answer:  True`

8 Mistake Bounds (9 points)

Assume that we have the predictions below of five experts, as well as the correct answer.

(a) *(3 points)* Using the Weighted-Majority algorithm (with $\beta = 0.5$) in order to track the best expert, show how the weight given to each expert changes after each example. Show your work.

Expert	1	2	3	4	5	Correct Answer
	T	T	T	F	F	F
Weights	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	1	
	F	T	F	T	T	T
Weights	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1	1	
	T	F	F	F	T	F
Weights	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	1	$\frac{1}{2}$	

(b) *(3 points)* Suppose we run the Weighted-Majority algorithm using n experts and $\beta = 0.5$. We find out that the average number of mistakes made by each expert is m but the best expert makes no mistake. Circle below the expression for the bound on mistakes made by the Weighted-Majority algorithm.

A) $O(n)$ B) $O(log_2 n + m)$ C) $O(log_2 n)$ D) none of the above

`Answer: C`

(c) *(3 points)* Notice that since there is an expert who made zero mistakes, we could use the Halving Algorithm instead (which of course corresponds to Weighted-Majority algorithm when $\beta = 0$). Circle below the bound on mistakes made by the Halving algorithm when given the same n experts.

A) $O(n)$ B) $O(log_2 n + m)$ C) $O(log_2 n)$ D) none of the above

`Answer: C`

9 Decision Trees - Harder Questions (4 points)

(Only 4 points, so only attempt this if you are happy with all your other answers).

(a) *(2 points)* Suppose we have three binary attributes (A, B and C) and 4 training examples. We are interested in finding a *minimum-depth* decision tree consistent with the training data. Please give a target concept and a *noise-free* training set for which ID3 (no pruning) will not find the decision tree with the minimum depth.

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

`Target concept:` A `XOR` B

```
Minimum-depth tree:
          A
         / \
      0/   \1
      /     \
     B       B
    / \     / \
  0/   \1 0/   \1
  /     \ /     \
  0     1 1     0
```

```
Tree learned by ID3:
          C
         / \
      0/   \1
      /     \
     0       A
            / \
          0/   \1
          /     \
         B       1
        / \
      0/   \1
      /     \
     0       1
```

(b) *(2 points)* Suppose we learned a decision tree from a training set with binary output values (class = 0 or class = 1). We find that for a leaf node l, (1) there are M training examples falling into it; and (2) its entropy is H. Sketch a simple algorithm which takes as input M and H and that outputs the number of training examples misclassified by leaf node l.

`The entropy function` H `can be approximated using`

$$1 - 4 \times (0.5 - p)^2 \ \ (0 \le p \le 1).$$

`So` $p = 0.5 - \sqrt{\frac{1-H}{4}}$ `(here` $0 \le p \le 0.5$`).`
`The number of misclassified training examples is roughly`

$$M \times p = M \times (0.5 - \sqrt{\frac{1-H}{4}}).$$