10-601A Machine Learning
Professor Roni Rosenfeld
Spring 2015
Mid-term Exam, 4 March 2015

Instructions:

• Make sure your exam is not missing any sheets, than write your full name and andrew userid on the top sheet, and your andrew userid on all other sheets, as sheets may become separated.

• This exam is closed books, closed notes. You may make use of one page of notes you may have brought with you.

• You may not use a calculator, and must put away, out of site, all electronic devices. Expressions like $\log_2(3), \log_2(5), \log_2(7)$ etc., do not need to be simplified further.

• You must answer all questions.

• Note that some questions ask you to explain your answer. Write your explanations, clearly and concisely, in the space provided. If an explanation was not explicitly requested, it is not required.

Please print your full name and userid here:
1. (14 points) Consider a credit card fraud detection problem, where the goal is to learn
the concept "likely to be fraudulent" for any given credit card transaction.
The input instances are credit card transactions, represented by six fields (attributes),
each with a small number of possible values, as follows:

- cardholder AGE: { <30 , 30+ }
- cardholder annual household INCOME: { <$20K , $20K+ }
- cardholder credit LINE: { <$1000 , $1000+ }
- transaction AMOUNT: { <$100 , $100+ }
- transaction TIME: {day (6AM-6PM), night (6PM-6AM)}
- transaction LOCATION: { urban, suburban }

(a) What is the size of the input space, "X"?

(b) What is the size of the concept space (ie the space of all possible concepts)?

(c) Consider the set of hypotheses that can be expressed by specifying the values of ex-
actly three different attributes (e.g. "TIME=night AND AGE=30+ AND AMOUNT<$100").
What is the size of this hypothesis space, "H"? Does it have a bias? If so, is it hard
(restriction bias) or soft (search bias)? If not, why not?

(d) what is the size of the version space VS(H,D) for the H in part (c) above and the
empty training set D (i.e. a set containing no instances)

Please print your userid here:
2. (18 points)
   (a) True or False? Assuming you are not concerned with the training time, when using an artificial neural network it is best to include enough hidden units so the training error can be reduced as much as possible. Explain in 1 sentence.

   (b) True or False? A single perceptron can only compute linear variations of AND, OR and XOR. Explain in 1 sentence.

   (c) When it is possible to run a Gradient Descent algorithm, what is guaranteed by the algorithm (1 sentence) and what is not guaranteed (1 sentence)?

   (d) What is minimized by the BackPropagation algorithm? (1 sentence or less).

   (e) True or False? Convergence of the BackPropagation algorithm is generally not guaranteed, unless the error surface is convex. Explain in 1 sentence.

   (f) True or False? The necessary and sufficient condition for running a gradient descent algorithm on a neural network is that each neuron’s transfer function be differentiable. Explain in 1 sentence.

Please print your userid here:
3. (12 points) An experiment consists of flipping simultaneously a fair dime and a fair penny. You are asked to predict whether the penny ends up heads or tails.
(a) What is the entropy of your prediction?

(b) You are now told the total number of “tail”s in the experiment (this could be 0, 1 or 2). Does this help your prediction, on average? What is the average entropy of your prediction now? Show your work, or explain.

(c) As in (b), but you are only told whether the total number of “tail”s in the experiment was odd or even. Does this help your prediction, on average? What is the average entropy of your prediction now? Show your work, or explain.
4. (16 points)

Given the real-valued function $F_1(x_1, x_2, \ldots, x_7) = \text{AVERAGE}(x_1, x_2, \ldots, x_7)$ of the real valued parameters $x_1, x_2, \ldots, x_7$:

(a) Can $F_1$ be expressed by a linear unit (an artificial neuron with a linear transfer function)? If yes, briefly show how. If not, briefly explain why not. Is there a known algorithm for automatically learning the appropriate weights from labeled data? If yes, what is it called?

(b) Can $F_1$ be expressed by a perceptron? If yes, briefly show how. If not, briefly explain why not. Is there a known algorithm for automatically learning the appropriate weights from labeled data? If yes, what is it called?

Given the real-valued function $F_2(x_1, x_2, \ldots, x_7) = \sin\left(\frac{1}{(x_1 + x_2 + \ldots + x_7)}\right)$ of the real valued parameters $x_1, x_2, \ldots, x_7$:

(c) Can $F_2$ be expressed by a suitably large network of linear units? Briefly explain. Is there a known algorithm for automatically learning the weights of such a network from labeled data? If yes, what is it called?

(d) Can $F_2$ be expressed by a suitably large network of sigmoid units? Briefly explain. Is there a known algorithm for automatically learning the weights of such a network from labeled data? If yes, what is it called?
5. (14 points) Let $X, Y$ be some jointly distributed numerical random variables. In each of the following, fill in the blank with exactly one of: \{$<, \leq, =, \geq, >, ?\}$, where ‘?’ means that none of the other relations necessarily holds:

- $H(X) + H(Y) \quad ? \quad H(X,Y) + I(X; Y)$
- $H(X) + H(Y) \quad ? \quad H(X + Y)$ (Notice this is $H(X + Y)$, not $H(X, Y)$)
- $H(\cos(Y)) \quad ? \quad H(Y)$
- $H(X^3) \quad ? \quad H(X)$
- $H(Y) \quad ? \quad H(Y | X = x)$ for some given $x$
- $H(Y) \quad ? \quad H(Y | X)$
- $H(Y) \quad ? \quad p(X = x) \cdot H(Y | X = x)$ for some given $x$

6. (8 points) Let $X, Y$ be jointly distributed as follows:

<table>
<thead>
<tr>
<th>$X$ / $Y$</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Calculate:
- Correlation coefficient $\rho_{X,Y} =$
- Mutual Information (in bits) $I(X; Y) =$

Please print your userid here:
7. (18 points) You desperately want to nail the job at that hot startup. Nevertheless, you know that half the candidates that interview are not hired, as evidenced by what happened to the last eight you’ve heard of. In order to better understand your chances, you look at what happened during these eight interviews. You have information about the candidates’ college major, area of experience, and the tie they wore during the interview. You use ID3 (the basic tree-construction algorithm we discussed in class, without any pruning) to build a decision tree that will explain this training data:

<table>
<thead>
<tr>
<th>major</th>
<th>experience</th>
<th>tie</th>
<th>hired?</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>programming</td>
<td>pretty</td>
<td>NO</td>
</tr>
<tr>
<td>CS</td>
<td>programming</td>
<td>pretty</td>
<td>NO</td>
</tr>
<tr>
<td>CS</td>
<td>management</td>
<td>pretty</td>
<td>YES</td>
</tr>
<tr>
<td>CS</td>
<td>management</td>
<td>ugly</td>
<td>YES</td>
</tr>
<tr>
<td>business</td>
<td>programming</td>
<td>pretty</td>
<td>YES</td>
</tr>
<tr>
<td>business</td>
<td>programming</td>
<td>ugly</td>
<td>YES</td>
</tr>
<tr>
<td>business</td>
<td>management</td>
<td>pretty</td>
<td>NO</td>
</tr>
<tr>
<td>business</td>
<td>management</td>
<td>pretty</td>
<td>NO</td>
</tr>
</tbody>
</table>

(a) At the root of the tree, what is the entropy?

(b) At the root of the tree, what is the mutual information offered about the label by each of the three attributes?

(c) Which attribute will ID3 place at the root?

(d) Will ID3 come up with the smallest (i.e. fewest nodes) possible tree that is consistent with this training data? If so, explain why. If not, show a smaller tree.