Practice Final Exam
DI 101, Fall 2012

Paul Anderson
Department of Computer Science
College of Charleston

December 3, 2012

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

Name: ____________________________________________

1. (5 points) What is the difference between data, information, and knowledge? Use specific examples in your answer.
2. (5 points) Describe the Knowledge Discovery in Databases (KDD) process. Specifically describe the steps of the KDD process and its goals. Your steps may be organized differently than the notes, but the order must be preserved.

3. (5 points) What is the difference between unsupervised and supervised learning?

4. (5 points) What are the four scales of measurement? If you don’t remember the correct name, I’ll give you partial credit for clear examples.

5. (5 points) In clustering, do we want to maximize or minimize within cluster variation?
6. (5 points) In clustering, do we want to maximize or minimize between cluster variation?

7. (5 points) Compare and contrast exclusive clustering and overlapping clustering? General ideas are fine.

8. (5 points) For the following data, draw the first and second principal component vectors.
9. (15 points) Cluster the smartphone dataset using k-means cluster (k = 2). The initial centroids are 10% and 20%.

**Smartphone Penetration Across Global Markets**

Source: The Nielsen Company
10. (15 points) Using the naive Bayes classifier, compute the probability that house number 3 is acceptable. You must first construct the histograms necessary for this calculation (5 points). Then you must compute the probabilities (10 points, showing all of your work).

<table>
<thead>
<tr>
<th>House</th>
<th>Furniture/F</th>
<th># Rooms/R</th>
<th>Kitchen/K</th>
<th>Acceptable/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Not included</td>
<td>3</td>
<td>New</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Included</td>
<td>3</td>
<td>Old</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Not included</td>
<td>4</td>
<td>Old</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Not included</td>
<td>3</td>
<td>Old</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Included</td>
<td>4</td>
<td>Old</td>
<td>Yes</td>
</tr>
</tbody>
</table>
11. (5 points) Using the k-nearest-neighbors algorithms and $k = 3$, what is the class label of the ? mark.

12. (5 points) What is the difference between k-fold and leave-one-out cross-validation?

13. (5 points) When would you argue that leave-one-out is more advantageous than k-fold and hold out (or test set)?

14. (5 points) Why do we perform cross-validation at all?

15. (5 points) Describe leave-one-out cross-validation.

16. (15 points) Construct the decision tree using the table above using 1 - 7.

17. (5 points) What is misclassification rate using samples 8 - 14?

18. (5 points) When would you use linear regression (ordinary least squares)? Give a specific example.
### Table 7.1

Training data tuples from the *AllElectronics* customer database.

<table>
<thead>
<tr>
<th>RID</th>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>Class: buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>31...40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>31...40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>&lt;=30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>31...40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>31...40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>
19. (5 points) When would you use partial least squares regression (PLS)? Give a specific example.

20. (5 points) Describe a problem that you would solve with classification and not regression.

21. (5 points) What are the conditions that linear regression (ordinary least squares) is prone to fail?

22. (5 points) Describe the steps needed to identify lane lines in an image using the video in class as a guide.
23. (15 points) Trace the results of using the Apriori algorithm on the grocery shop with support threshold 1/6 and confidence threshold 50%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence.

T1: Hot Dogs, Buns, Ketchup
T2: Hot Dogs, Bun
T3: Hot Dogs, Coke, Chips
T4: Chips, Coke
T5: Chips, Ketchup
T6: Hot Dogs, Coke, Chips
24. (15 points) Use the hill climbing strategy discussed in class to solve the following bin packing problem. Use the heuristic of selecting the smallest item next.

Capacity of each bin is 12.

The height of the items are 2, 4, 3, 7, 5, 6, 8, 3, 4
25. (15 points) Using the simulated annealing strategy described in class to see if you can find a better solution than the hill climbing strategy using 5 iterations. Here is what you’ll need to know:

a. $T = 5$ initially. It decreased by 1 after each iteration
b. You decide (randomly) whether to merge or split
c. You decide (randomly) which bin to split (and how) or which two bins to merge
d. Here are the random numbers that you’ll need to decide if you need to reject:

0.51 0.95 0.1 0.4 0.3

You reject if $\langle$ insert random number $\rangle$ is greater than $e^{-B/T}$

Clearly explain your random choices.
Clearly indicate your steps and the 5 iterations.

26. (5 points) Describe at least 2 applications of ethics to DI. Explain how these ethical applications can be approached (i.e., how are people working to solve them?).

27. (5 points) What is some of the motivation behind the semantic web? How does RDF relate to the semantic web?

28. (10 points) Definition of two tables and their attributes:

STUDENT(SID, LastName, FirstName, Major, Rank, GPA, AdvisorID)

FACULTY(FID, LastName, FirstName, Rank, DepartmentID)

There are two students defined:
1234, Anderson, Paul, CS, Sophomore, 3.2, 6543 5678, Johnson, Mike, MATH, Freshman, 3.5, 6543

There is one advisor:
6543, Olmsted, Aspen, Professor, 2

Represent the STUDENT and FACULTY tables listed above in an RDF graph.