**CSCI 4525/5525: Written Assignment for Unit 3: Learning**

**Assignment**

Complete the following exercises – they are inspired from *Artificial Intelligence: A Modern Approach* though I have changed them slightly. They are reprinted here for your convenience (and because the digital “global edition” has different problems than the physical copy, this should clarify things—do the problems here!) The numbers I use here are from the physical edition of the book.

**There are 3 total questions for undergrads, totaling 17 points, and 4 total questions for graduate students, totaling 27 points. NOTE: Part c of Question 3 is only required for graduate students.**

**Question 1** *(inspired by 14.1 from AI: A Modern Approach, page 558 but do this one) –* ***2 parts****,* ***5 points total***

We have a bag of three biased coins a, b, and c with probabilities of coming up heads of 30%, 70%, and 90% respectively. One coin is drawn randomly from the bag (with equal likelihood of drawing each of the three coins), and then the coin is flipped three times to generate the outcomes X1, X2, and X3.

1. Draw the Bayesian network corresponding to this setup and define the necessary CPTs (conditional probability tables). Hint: think of the coin and the three outcomes as random variables. (3 points)
2. Calculate which coin was most likely to have been drawn from the bag if the observed flips come out tails twice and heads once. (2 points). Note: you must show your probability calculations, i.e., show your work “along the way”, not just the final answer.

**Question 2** (inspired by *18.6 from AI: A Modern Approach, page 764 but do this one*) – **5 points**

Consider the following data set comprised of three binary input attributes (i.e., the features: A1, A2, and A3) and one binary output *y* (i.e., the class label):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Example | A1 | A2 | A3 | Output *y* |
| x1 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1 | 1 | 0 |
| x3 | 1 | 1 | 0 | 1 |
| x4 | 0 | 0 | 1 | 1 |
| x5 | 1 | 0 | 0 | 1 |

Use the algorithm we covered in class to learn a decision tree for these data. Show **both the final tree that is learned** as well as the computations made to determine the attribute to split at each node, and make clear what each of them are along the way. This includes such calculations as:

**\*Entropy of the parent.**

**\*Entropy of a child node.**

**\*Weighted entropy of all the children nodes.**

**\*Information gained if this node was to be selected.**

You must show information gain calculations for each attribute each time you choose an attribute; it is not enough to show only the attribute that is chosen.

In the event that there is more than one node that would produce the same information gain if it were to be split upon, note this and then split on the node that comes “before” the other one (e.g., if splitting on A1 and A2 would both have the same information gain, split on A1 instead of A2).

**Question 3** (*Custom Clustering Problem!*) – **2 parts for undergrads (7 points total), 3 parts for graduate students (12 points total)**

**At the end of this assignment document is part of a famous collection of observations**, which you’ll need to use to answer the questions in this problem. The data is about the sizes of organs in three different species of Iris flowers (Setosa, Versicolor, and Virginica). 10 examples of each species of Iris are given. You are welcome to write a program or to use weka to help you answer these questions if you would like.

1. **4 points**: Classify these two unknown flowers using 3 Nearest Neighbor (3NN) and Euclidean distance. For both flowers, you must list their 3 nearest neighbors and the distance between that flower and each neighbor.

Flower X:

Sepal Length: 6.5cm

Sepal Width: 2.9cm

Petal Length: 1.5cm

Petal Width: 0.6cm

Flower Y:

Sepal Length: 5.9cm

Sepal Width: 3.0cm

Petal Length: 5.1cm

Petal Width: 1.8cm

1. **3 points**: Use k-Means clustering with k= 3 to group all 30 observations into three clusters. Initially, each cluster should contain every third element. So, initially cluster 1 will contain observations 1,4,7,10,13,16,19,22,25,28, Cluster 2 will contain observations 2,5,8,11,14,17,20,23,26,29, and Cluster 3 will contain observations 3,6,9,12,15,18,21,24,27,30. Repeat the k-means algorithm until it converges. What are the coordinates of the three centroids after convergence? (Note that this is an unsupervised learning problem, so you should ignore the class label when doing this clustering.)
2. **FOR GRADUATE STUDENTS ONLY (5 points):** Use DBSCAN clustering with and *minPoints* = 2 to group all 30 observations. How many clusters were discovered, and which observations are in which clusters? Which observations are classified as noise? (Note that this is an unsupervised learning problem, so you should ignore the class label when doing this clustering).

**Students enrolled in CSCI 5525 must also complete:**

**Question 4** (inspired by *18.15 from AI: A Modern Approach, page 766 but do this one*) – **5 points**

In class when we discussed KNN we discussed it as a classification problem, where you look at the “labels” of neighbors. But KNN can also be a ‘regression’ problem as well, where instead of a class label, each example has a “number,” and our goal is to figure out what ‘number’ we should make our new example. One way to consider this is to pick a number for this new example that “minimizes loss” – however, there are several different loss functions that we might use; the one that we choose might influence what number we assign our new value.

Suppose we run our new example, ‘x’, through a 7-nearest-neighbors regression search and it returns {3,4,8,12,8,99,11} as its 7 nearest *y* values. Given that…

1. What is the value of *ŷ* (i.e., the value of y to assign to this new x example) that minimizes the L1 loss function on this data? That is, what number should our new example have, given the values of its neighbors using L1
2. There is a common name in statistics for this value as a function of the *y* values; what is it?
3. Answer the same above two questions for the L2 loss function.

The hard copy of the book and the global edition has the two loss functions, L1 and L2, on page 711. They are reprinted here for your convenience:

L1 (y, ŷ) = |y - ŷ|

L2 (y, ŷ) = (y - ŷ)2

Note: you must provide some evidence that the values you give for *ŷ* are correct. This can be either a mathematical proof or a sufficiently long list of other possible values, demonstrating that the error is lowest for the values of *ŷ*  you selected.

As an example, let’s say you wanted to compute the error for L1 when *ŷ = 3.* You could do that like:

|  |  |
| --- | --- |
| *ŷ* | L1 |
| 3 | ++++++ = 0+1+5+9+5+96+8=  124 |

So that’s the error for L1 when the algorithm returns 3! What’s the value that minimizes the error?

**Submission**

Please observe these requirements in your submission:

* Submission must include your name and which section of the class you are enrolled in (i.e. 4525 or 5525).
* Submissions must be typed.
* Submissions must be submitted as PDF files.
* Submissions must be uploaded to Moodle on time

**Grading**

Problems will be graded using the following simple grading procedure, applied to each part (e.g. part a, part b, etc.) of the problem:

* Problem not attempted or does not demonstrate significant effort: no credit.
* Problem thoroughly attempted, but answer is incorrect or incomplete: half credit.
* Problem thoroughly attempted, and answer is correct and complete: full credit.

**You must show your work on all problems to receive credit**. Simply giving the answer (correct or incorrect) will earn no credit for that problem. The problems themselves have some pointers as to what I am looking for on each problem.

**Iris Data Set**

This data is a subset of the famous Iris dataset, originally published in:

Fisher, R.A. "The use of multiple measurements in taxonomic problems," Annual Eugenics, 7, Part II, 179-188 (1936).

It was taken, in its current form, from the University of California, Irvine Machine Learning Repository:

Lichman, M. UCI Machine Learning Repository <http://archive.ics.uci.edu/ml>. University of California, Irvine School of Information and Computer Science.

It is given below as both as table and as text suitable for copying and pasting into a .csv file.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Sepal Length (cm)** | **Sepal Width (cm)** | **Petal Length (cm)** | **Petal Width (cm)** | **Species** |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 11 | 7.0 | 3.2 | 4.7 | 1.4 | Iris-versicolor |
| 12 | 6.4 | 3.2 | 4.5 | 1.5 | Iris-versicolor |
| 13 | 6.9 | 3.1 | 4.9 | 1.5 | Iris-versicolor |
| 14 | 5.5 | 2.3 | 4.0 | 1.3 | Iris-versicolor |
| 15 | 6.5 | 2.8 | 4.6 | 1.5 | Iris-versicolor |
| 16 | 5.7 | 2.8 | 4.5 | 1.3 | Iris-versicolor |
| 17 | 6.3 | 3.3 | 4.7 | 1.6 | Iris-versicolor |
| 18 | 4.9 | 2.4 | 3.3 | 1.0 | Iris-versicolor |
| 19 | 6.6 | 2.9 | 4.6 | 1.3 | Iris-versicolor |
| 20 | 5.2 | 2.7 | 3.9 | 1.4 | Iris-versicolor |
| 21 | 6.3 | 3.3 | 6.0 | 2.5 | Iris-virginica |
| 22 | 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |
| 23 | 7.1 | 3.0 | 5.9 | 2.1 | Iris-virginica |
| 24 | 6.3 | 2.9 | 5.6 | 1.8 | Iris-virginica |
| 25 | 6.5 | 3.0 | 5.8 | 2.2 | Iris-virginica |
| 26 | 7.6 | 3.0 | 6.6 | 2.1 | Iris-virginica |
| 27 | 4.9 | 2.5 | 4.5 | 1.7 | Iris-virginica |
| 28 | 7.3 | 2.9 | 6.3 | 1.8 | Iris-virginica |
| 29 | 6.7 | 2.5 | 5.8 | 1.8 | Iris-virginica |
| 30 | 7.2 | 3.6 | 6.1 | 2.5 | Iris-virginica |

5.1,3.5,1.4,0.2,Iris-setosa

4.9,3.0,1.4,0.2,Iris-setosa

4.7,3.2,1.3,0.2,Iris-setosa

4.6,3.1,1.5,0.2,Iris-setosa

5.0,3.6,1.4,0.2,Iris-setosa

5.4,3.9,1.7,0.4,Iris-setosa

4.6,3.4,1.4,0.3,Iris-setosa

5.0,3.4,1.5,0.2,Iris-setosa

4.4,2.9,1.4,0.2,Iris-setosa

4.9,3.1,1.5,0.1,Iris-setosa

7.0,3.2,4.7,1.4,Iris-versicolor

6.4,3.2,4.5,1.5,Iris-versicolor

6.9,3.1,4.9,1.5,Iris-versicolor

5.5,2.3,4.0,1.3,Iris-versicolor

6.5,2.8,4.6,1.5,Iris-versicolor

5.7,2.8,4.5,1.3,Iris-versicolor

6.3,3.3,4.7,1.6,Iris-versicolor

4.9,2.4,3.3,1.0,Iris-versicolor

6.6,2.9,4.6,1.3,Iris-versicolor

5.2,2.7,3.9,1.4,Iris-versicolor

6.3,3.3,6.0,2.5,Iris-virginica

5.8,2.7,5.1,1.9,Iris-virginica

7.1,3.0,5.9,2.1,Iris-virginica

6.3,2.9,5.6,1.8,Iris-virginica

6.5,3.0,5.8,2.2,Iris-virginica

7.6,3.0,6.6,2.1,Iris-virginica

4.9,2.5,4.5,1.7,Iris-virginica

7.3,2.9,6.3,1.8,Iris-virginica

6.7,2.5,5.8,1.8,Iris-virginica

7.2,3.6,6.1,2.5,Iris-virginica