# SEEM2460 Introduction to Data Science
## Lab Assignment

Lab Assignment is due by **5:00 pm, March ~~22$^{ed}$~~ 29** (Fri) 2024.
You need to submit your work (in PDF format) to both the Blackboard and Veriguide.
-----------------------------------------------------------------------------------------------------------

### **Question 1: Decision Tree** (20 marks)

Table 1 presents a simplified animal classification dataset with 12 animal samples. In this table, "Toothed" and "Legs" are two attributes, and each animal sample is categorized as either "Mammal" or "Reptile" species. In this question, you are asked to develop a binary decision tree (i.e., each parent node has two splits) based on this dataset. Please round your calculation results to two decimal places.

| ID | Toothed | Legs | Species |
|---|---|---|---|
| 1 | True | True | Mammal |
| 2 | False | True | Reptile |
| 3 | True | True | Mammal |
| 4 | True | False | Reptile |
| 5 | False | True | Mammal |
| 6 | True | False | Reptile |
| 7 | True | True | Mammal |
| 8 | False | True | Reptile |
| 9 | True | False | Reptile |
| 10 | True | True | Mammal |
| 11 | False | True | Reptile |
| 12 | False | False | Reptile |

Table 1: Animal Classification Dataset.

(a) Based on the **maximize Information Gain** criterion, choose an attribute as root node. Show your calculation details to prove the chosen attribute is preferred. (10 points)
hint: $\text{GAIN}_{\text{split 'toothed'}}$ vs. $\text{GAIN}_{\text{split 'legs'}}$

(b) Based on your chosen root node, draw a **two-layer** decision tree to predict the "Species" column using these two attributes. (5 marks)

(c) **WEKA Tool Practice**: Use the WEKA tool to classify the data in Table 1 with decision tree (**J48**) under the test option "**Use training set**". Copy the results in the 'classifier output' window (refer to Week 8A slides page 49) to your submission and show the visualization of the resulting tree. (5 marks)

## **Question 2: SVM** (20 marks)

Considering the following data points for two classes:

**Class 0**: (1, 0), (2, 2), (3, 1), (4, 3)
**Class 1**: (-5, 5), (-5, 2), (-2, 5), (0, 0)

Please complete the following tasks:

(a) Determine the centroids for both Class 0 and Class 1. (5 marks)

(b) Compute the decision boundary that separates the two classes and create a diagram illustrating the boundary and the points for each class. Ensure that the points for different classes are represented by distinct colors. (10 marks)

(c) Determine which class the new point (0,1) belongs to, and indicate the location of point (0, 1) on the previously created diagram. (5 marks)

## **Question 3: K-means Practice** (20 marks)

Table 3 presents the calorie content of 5 different beverages. Please use K-means clustering to divide these beverages into two clusters based on the initial centroids $C_1$ = 150 and $C_2$ = 250. Illustrate your clustering process by outlining **each step** of the clustering.

| Beverage | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| Calories (kcal) | 160 | 215 | 95 | 45 | 195 |

Table 3: Caloric values of five different beverages.

# Solutions

## **Question 1: Decision Tree** (20 marks)

### <u>Sol (a):</u>

Information gain of splitting an attribute is calculated by:

$$GAIN_{split} = Ent(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Ent(i)\right)$$

$$where$$

$$Ent(p) = -\sum_{j=1}^{k} P(j|p) \log_2 P(j|p) = -\frac{7}{12}\log_2\frac{7}{12} - \frac{5}{12}\log_2\frac{5}{12} \approx 0.98$$

- Splitting “Toothed” attribute:

$$\begin{aligned} GAIN_{sp"toothed"} &= Ent(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Ent(i)\right) \\ &\approx 0.98 - \frac{7}{12}\left(-\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}\right) - \frac{5}{12}\left(-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}\right) \\ &\approx \mathbf{0.11} \end{aligned}$$

**(5 marks, -1 marks for calculation error)**

- Splitting “Legs” attribute:

$$\begin{aligned} GAIN_{sp"legs"} &= Ent(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Ent(i)\right) \\ &\approx 0.98 - \frac{8}{12}\left(-\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8}\right) - \frac{4}{12}\left(-\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4}\right) \approx \mathbf{0.34} \end{aligned}$$

**(5 marks, -1 marks for calculation error)**

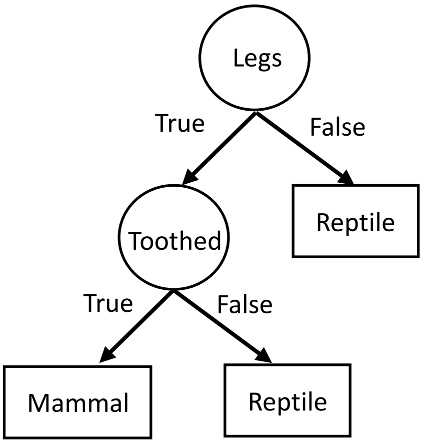
As 0.34 > 0.11, we should choose “Legs” as the root node.

**Sol (b): (5 marks, where 2 marks for Stopping Criteria)**

According to question (a), “Legs” is split as the root node.

For the right partition, stop expending as it is a pure leaf (i.e., all samples belong to the Reptile class).

For the left partition, expand it by splitting “Toothed”. For each resulting leaf, choose the majority class as the prediction result.



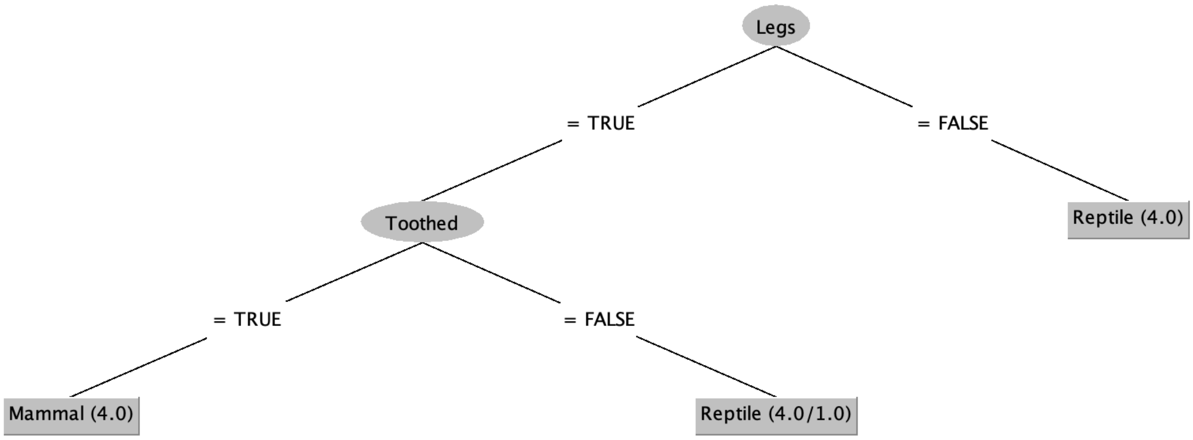
**Sol (c):**

Classification outputs (3 marks):

```
=== Run information ===
Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    data_for_WEKA
Instances:    12
Attributes:   3
             Toothed
             Legs
             Species
Test mode:    evaluate on training data

=== Classifier model (full training set) ===
J48 pruned tree
------------------


Legs = TRUE
|   Toothed = TRUE: Mammal (4.0)
|   Toothed = FALSE: Reptile (4.0/1.0)
Legs = FALSE: Reptile (4.0)

Number of Leaves  :          3

Size of the tree :   5

Time taken to build model: 0 seconds

=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds

=== Summary ===
Correctly Classified Instances          11               91.6667 %
Incorrectly Classified Instances         1                8.3333 %
Kappa statistic                          0.8235
Mean absolute error                      0.125
Root mean squared error                  0.25
Relative absolute error                 25.6098 %
Root relative squared error             50.6945 %
Total Number of Instances               12
```

```
=== Detailed Accuracy By Class ===
             TP Rate  FP Rate  Precision  Recall  F-Measure  MCC     ROC Area  PRC Area  Class
             0.800    0.000    1.000      0.800   0.889      0.837   0.957     0.925     Mammal
             1.000    0.200    0.875      1.000   0.933      0.837   0.957     0.946     Reptile
Weighted Avg.   0.917    0.117    0.927      0.917   0.915      0.837   0.957     0.938

=== Confusion Matrix ===
 a b   <-- classified as
 4 1 | a = Mammal
 0 7 | b = Reptile
```

## Visualization (2 marks):



## Question 2: SVM (20 marks)

Note: The answer is not unique.

(i) Centroid points $m_1$ and $m_0$ for both class 1 and class 0:

$$m_1 = \frac{\sum_{x\in\{1\}} x}{4} = (-3,3)^T$$
$$m_0 = \frac{\sum_{x\in\{0\}} x}{4} = (2.5,1.5)^T$$

**(5 marks)**

(ii) Since $m_1 - m_0 = (-5.5, 1.5)^T$ and $m_1 + m_0 = (-0.5, 4.5)^T$, we compute equation of decision boundary:

$$(-5.5,1.5)(x,y)^T - \frac{(-5.5,1.5)(-0.5,4.5)^T}{2} = 0$$
$$-5.5x + 1.5y - 4.75 = 0$$
$$-11x + 3y - 9.5 = 0$$

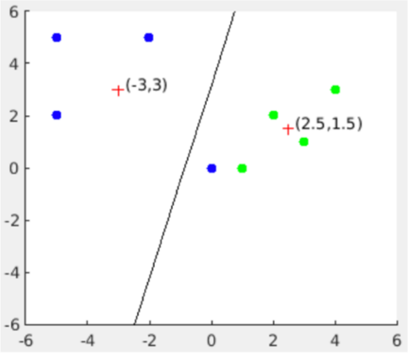


Figure 1: Decision boundary

**(10 marks): 5 marks for calculation, 5 marks for figure.**

(iii)For point $[0,1]$, we compute:

$$-11 \times 0 + 3 \times 1 - 9.5 = -6.5 < 0$$

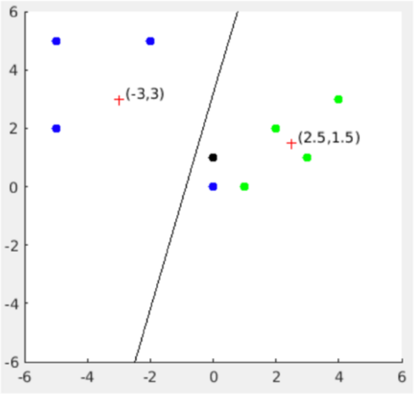


Figure 2: New point [0,1]

(0,1) belongs to class 0.

**(5 marks)**

## Question 3: K-means Practice (20 marks)

**Sol (20 marks):**
**(15 marks for correct iterations, 5 marks for stop criterion, -2 marks for calculation error.)**

Step 1:

Initially,

$$C_1 = 150 \quad , C_2 = 250,$$

hence the ages can be clustered as:

$$Cluster_1 : \{160, 95, 45, 195\}, \;\; Cluster_2 : \{215\}$$

Step 2:

Update centroids:

$$C_1 = \frac{160 + 95 + 45 + 195}{4} = 123.75, \;\; C_2 = 215$$

New membership based on new centroids:

$$C_1 : \{160, 95, 45\}, \;\; C_2 : \{195, 215\}$$

Step 3:

Update centroids:

$$C_1 = \frac{160 + 95 + 45}{3} = 100, \;\; C_2 = \frac{195 + 215}{2} = 205$$

Update membership:

$$C_1 : \{95, 45\}, \;\; C_2 : \{\, 160, 195, 215\}$$

Step 4:

Update centroids:

$$C_1 = \frac{95 + 45}{2} = 70, \;\; C_2 = \frac{160 + 195 + 215}{3} = 190$$

Update membership:

$$C_1 : \{95, 45\}, \;\; C_2 : \{\, 160, 195, 215\}$$

Then stop clustering since the membership does not change anymore. So, the resulting clusters are:

$$\boldsymbol{C_1 : \{B3, B4\}, \;\; C_2 : \{\, B1, B2, B5\}}$$