**Assignment 2 – Due: Feb. 23**

**Problem #1 – show your work. If you write code to implement this, show your calculations and compare it with the code solution.**

**1 (40)** Consider the training examples shown in **Table 1** for a binary classification problem.

|  |  |  |  |
| --- | --- | --- | --- |
| a1 | a2 | a3 | Class |
| T | T | 1 | + |
| T | T | 5 | + |
| T | F | 4 | - |
| F | F | 3 | + |
| F | T | 8 | - |
| F | F | 8 | - |
| T | F | 7 | + |
| F | T | 4 | - |
| F | F | 12 | + |
| T | T | 10 | - |

Table 1. Sample training data for a binary classification problem for question #1.

(a) What is the entropy (information content) of the entire training data?

(b) What are the information gains of a1 and a2 and report which one gives a better split.

(c) Compute Gini index for a1 and a2 and report which one gives a better split.

(d) Compute information gain for a3 for each potential split point and report the best split point for a3.

**2 (60)** Implement a decision-tree algorithm for breast cancer diagnosis. The breast cancer data set is available from: <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>

1. If you found missing values, how did you deal with them in your code?
2. Show the accuracy of the classifier with two different measure: entropy and ginixx`

Submit your code.