**CS 422: Data Mining**

**Homework2**

**Problem 1:**

**Question 1.1:**

**8)** Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?

**Answer:**

* If the median of the data lies in the middle of the box in the box plot, then we can say that the values in the data are symmetrically distributed.
* From the figure 3.11 we can see that Sepal length and Sepal width attributes are relatively symmetrically distributed but Petal width and Petal length attributes are slightly Skewed.

**9)** Compare sepal length, sepal width, petal length, and petal width, using Figure 3.12.

**Answer:**

* From the figure 3.12 a) --- For the species type Setosa, Sepal Length>Sepal Width>Petal Length>Petal Width.
* From the figure 3.12 b) --- For the species type Versicolour, Sepal Length >Petal Length>Sepal Width >Petal Width.
* From the figure 3.12 c) --- For the species type Verginica, Sepal Length >Petal Length>Sepal Width >Petal Width.
* For the specious Versicolour and Verginica Sepal Length> Sepal Width, Petal Length>Petal Width and Petal Length > Sepal Width.

**10)** Comment on the use of a box plot to explore a data set with four attributes: age, weight, height, and income.

* From the box Plot, for each attribute we can infer the distribution of the values (i.e.) Minimum, Maximum, Median, Inter Quartile Range and Symmetric or asymmetric Distribution of the data.
* We can compare two attributes across different categories. For instance, data with lower age tend to have lower weight, height and income compared to data with higher age range.

**Question 1.3:**

**20)** Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, “+” and “−.” Half of the data set is used for training while the remaining half is used for testing.

a) Suppose there are an equal number of positive and negative records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

* Since the data set is equally divided among test and train data and the predictor predicts every test record to be positive the expected error rate of the classifier on the test data is **50%.**

b) Repeat the previous analysis if the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.

* The expected error rate of the classifier on the test data is **50%.**

c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?

* The expected error rate of the classifier on the test data is **33.3%.**

d) Repeat the previous analysis if the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.

* The expected error rate of the classifier on the test data is **44.4%.**