CS 422-01
Data Mining

Homework #2

March 4, 2018

Josh Bowden
(A20374650)

Professor Vijay Gurbani

Problem 1 Tan, Chapter 3

1.1 Exercise 8

Describe how a box plot can give you 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?

A box plot shows you the relative distribution of an attribute by quartiles. If the line the mean is in the middle of the box plot and the two adjacent quartiles are the relatively same size, the attribute is symmetrically distributed.

We can say that the *sepal length* and *sepal width* is nearly symmetrically distributed, while the *petal length* and the *petal width* are not symmetrically distributed.

1.2 Exercise 9

Compare sepal length, sepal width, petal length, and petal width, using Figure 3.12.

For all three of the flowers, the *sepal length* retains nearly the same symmetrical distribution with a median of nearly 6 cm.

For the *sepal width*, the mean of the distribution is around symmetrical distributed for the *versicolour* and the *virginica* while the *setosa* is skewed right.

For the *petal length*, the *versicolour* and the *virginica* has a similar sized distribution while the *setosa* has a significantly smaller distribution.

For the *petal width*, all three specifies a have a similarly small distribution except for the *virginica* having a larger distribution. The *setosa* has the smallest distribution with a few higher valued outliers.

1.3 Exercise 10

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

A box plot would be suitable at exploring this kind of dataset.

Depending on the dataset, the distribution of the *age* could tell which age groups the dataset focused on or if younger or older people were preferred.

The *width* and *height* likely would be a normal distribution. Otherwise, any skew shows if more people in the data set were shorter/taller and skinner/wider.

The *income* displayed as a box plot should show the median income and the 25% and 75% of the income. The skew of the box plot would show if there were poorer or richer people. The length of the box plot would show how wide a distribution of incomes would be in the dataset.

Problem 2 Tan, Chapter 4

2.1 Exercise 2

Consider the training examples shown in Table 4.7 for a binary classification problem.

a) Compute the Gini index for the overall collection of training examples.

b) Compute the Gini index for the *Customer ID* attribute.

	ID = 1	…	10	11	…	20
C0	1	…	1	0	…	0
C1	0	…	0	1	…	1

d) Compute the Gini index for the *Gender* attribute.

	M	F
C0	6	4
C1	4	6

$$\left(\frac{10}{20}\right)\left(1-\left(\frac{6}{10}\right)^2-\left(\frac{4}{10}\right)^2\right)+\left(\frac{10}{20}\right)\left(1-\left(\frac{4}{10}\right)^2-\left(\frac{6}{10}\right)^2\right)=0.06$$