RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Tutorial Problems Week 4

1. (a) Sketch the graph of a normal distribution with mean 20 and standard deviation 3. Remember that about 66% of the observations are within 1 standard deviation of the mean, about 95% are within 2 standard deviations and about 99% are within 3 standard deviations.

   (b) Sketch the graph of a normal distribution with mean 75 and standard deviation 15.

   (c) Suppose you selected 100 RMIT students at random. What do you think that the age distribution would be?

   (d) Suppose you selected 100 old age pensioners at random. What do you think that the age distribution would be?

   (e) If you took the RMIT students and old age pensioners together, computed the mean and standard deviation and plotted the corresponding distribution, what would it look like? Sketch the distribution.

   (f) Explain how the above relates to clustering.

2. Consider the following EM output:

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:     iris-weka.filters.unsupervised.attribute.Remove-R5
Instances:    150
Attributes:   4
              sepallength
              sepalwidth
              petallength
              petalwidth
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


EM
==

Number of clusters selected by cross validation: 5
```

```
              Cluster
Attribute           0       1       2       3       4
               (0.18)  (0.23)  (0.28)  (0.15)  (0.15)
======================================================
sepallength
  mean         4.7748  6.8585  6.1613  5.2823  5.5432
  std. dev.    0.2405  0.5228  0.4138  0.2407  0.3159

sepalwidth
  mean         3.1789  3.0862  2.8547  3.7037  2.5786
  std. dev.    0.2599  0.2891  0.2687  0.2857  0.2512

petallength
  mean         1.4194  5.7859  4.7484  1.5173   3.863
  std. dev.    0.1692  0.4745  0.3193  0.1592  0.3516

petalwidth
  mean         0.1948  2.1327  1.5757  0.3028  1.1696
  std. dev.    0.0557  0.2359  0.2196  0.1212  0.1351

Clustered Instances
0       28 ( 19%)
1       35 ( 23%)
2       42 ( 28%)
3       22 ( 15%)
4       23 ( 15%)
Log likelihood: -1.60803
```

(a) Give an iterpretation of this output.

(b) There are three classes in the iris data. Why has the algorithm generated 5 clusters?

3. The following heights(cm) of a number of basketball players and jockeys were measured:

   Basketball: 213, 210, 214, 200, 195
   Jockeys: 145, 140, 150, 146, 141

   (a) What is the average and standard deviation for the basket ball players?

   (b) Sketch the distribution.

   (c) What is the average and standard deviation for the jockeys?

   (d) Sketch the distribution on the same graph as the basket ballplayers.

   (e) Suppose that the basketball players and jockeys were all in the same data file (sports.arff), for example,

```
@relation sports
@attribute height real
@data
213
210
214
200
195
145
140
150
146
141
```

What would be the average and standard deviation of all 10 individuals. Sketch the distribution on the same graph as above.

(f) Suppose that the EM algorithm was run on sports.arff. What would you expect the output to be.

4. Consider the following data:

```
@relation sports1
@attribute height real
@attribute age real
@data
213 22
141 35
210 23
146 31
214 21
200 24
195 23
150 30
145 32
140 31
```

(a) Starting with centroids (100 10) and (200, 20) work through three iterations of the K-means algorithm for K=2.

(b) What happens if the starting centroids are (177,28) and (178,29).

(c) Is there a pair of starting seeds for which the algorithm wont converge as expected?

5. Consider the following EM output:

```
=== Run information ===
Scheme:       weka.clusterers.EM -I 100 -N 3 -M 1.0E-6 -S 100
Relation:     studentdata
Instances:    300
Attributes:   4
              sex
              course
              age
              average_mark
Test mode:    evaluate on training data

=== Model and evaluation on training set ===
EM
==
Number of clusters: 3

                Cluster
Attribute             0        1        2
                 (0.34)   (0.33)   (0.33)
==========================================
sex
  m                 104        1      100
  f              1.0013  98.9987        1
  [total]      105.0013  99.9987      101
course
  MBC                 1        1      100
  BAPPSCI      104.0013  98.9987        1
  [total]      105.0013  99.9987      101
age
  mean          20.0427  20.0317   35.097
  std. dev.      0.9744   0.8606   1.4423

average_mark
  mean          75.0758  85.0595  95.0925
  std. dev.      1.6871   1.3594   1.4407

Clustered Instances
0      103 ( 34%)
1       98 ( 33%)
2       99 ( 33%)
Log likelihood: -4.41923
```

(a) Give an interpretation of this data.

(b) Are there any golden nuggets in this data?