RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Tutorial Problems Week 9

1. With respect to neural network training for classification and numeric tasks, distinguish the following types of error:
   (a) TSS (Total Sum Squared Error)
   (b) MSE (Mean Squared Error)
   (c) Classification error
   (d) Mean Absolute error
2. Explain the relevance of each of these errors to a classification task.
3. Explain the relevance of each of these errors to a classification task.
4. Explain the relevance of each of these errors to a numeric prediction task.
5. Which data set do they apply to, training, validation, test?
6. If the MSE is very large what would you expect the classification error to be?
7. If the MSE is very small what would you expect the classification error to be?
8. The attributes 'Density' to 'Wrist_Circumference_CM' are to be used in a neural network to predict 'class_Percent_Bodyfat'

```
@relation 'bodyfat'
@attribute Density real
@attribute Age real
@attribute Weight real
@attribute Height real
@attribute Neck_Circumference_CM real
@attribute Chest_Circumference_CM real
@attribute Abdomen_Circumference_CM real
@attribute Hip_Circumference_CM real
@attribute Thigh_Circumference_CM real
@attribute Knee_Circumference_CM real
@attribute Ankle_Circumference_CM real
@attribute Biceps_Circumference_CM real
@attribute Forearm_Circumference_CM real
@attribute Wrist_Circumference_CM real
@attribute class_Percent_Bodyfat real
@data
```

```
1.0708,23,154.25,67.75,36.2,93.1,85.2,94.5,59,37.3,21.9,32,27.4,17.1,12.3
1.0853,22,173.25,72.25,38.5,93.6,83,98.7,58.7,37.3,23.4,30.5,28.9,18.2,6.1
1.0414,22,154,66.25,34,95.8,87.9,99.2,59.6,38.9,24,28.8,25.2,16.6,25.3
1.0751,26,184.75,72.25,37.4,101.8,86.4,101.2,60.1,37.3,22.8,32.4,29.4,18.2,10.4
1.034,24,184.25,71.25,34.4,97.3,100,101.9,63.2,42.2,24,32.2,27.7,17.7,28.7
1.0502,24,210.25,74.75,39,104.5,94.4,107.8,66,42,25.6,35.7,30.6,18.8,20.9
1.0549,26,181,69.75,36.4,105.1,90.7,100.3,58.4,38.3,22.9,31.9,27.8,17.7,19.2
1.0704,25,176,72.5,37.8,99.6,88.5,97.1,60,39.4,23.2,30.5,29,18.8,12.4
1.09,25,191,74,38.1,100.9,82.5,99.9,62.9,38.3,23.8,35.9,31.1,18.2,4.1
1.0722,23,198.25,73.5,42.1,99.6,88.6,104.1,63.1,41.7,25,35.6,30,19.2,11.7
1.083,26,186.25,74.5,38.5,101.5,83.6,98.2,59.7,39.7,25.2,32.8,29.4,18.5,7.1
1.0812,27,216,76,39.4,103.6,90.9,107.7,66.2,39.2,25.9,37.2,30.2,19,7.8
1.0513,32,180.5,69.5,38.4,102,91.6,103.9,63.4,38.3,21.5,32.5,28.6,17.7,20.8
```

(a) How would you prepare the inputs?

(b) Compose a bash command to find the smallest value of class_Percent_Bodyfat using `head tail cut` and `sort`

(c) Modify your command to find the highest value.

(d) What are the minimum and maximum values of the class attribute in the above data?

(e) Using these values, how would you scale the class attribute in constructing the training and test data?

(f) What will be the training target for the value 12.3?

(g) Suppose that the output of the trained network is 0.11. What will be the corresponding value of class_Percent_Bodyfat?

(h) What is the theoretical minimum value of class_Percent_Bodyfat? The theoretical maximum value?

(i) Is it better to use the actual min/max values or the theoretical ones in scaling?

(j) Why is it important to scale the training and test data in the same way?