PSG College of Technology, Coimbatore -641 004

Department of Applied Mathematics and Computational Sciences

8th Semester MSc TCS

18XT87 Data Mining Lab

Problem Sheet - 3

1. Consider the following data (in increasing order) for the attribute age:
   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   (a) Use smoothing by bin means to smooth these data, using a bin depth of 3.
   (b) Use smoothing by bin median to smooth these data, using a bin depth of 3.
   (c) Use smoothing by bin boundaries to smooth these data, using a bin depth of 3.

2. Using the data for age and body fat given in Exercise 2.4, (page no. 80)answer the following:

   (a) Normalize the two attributes based on z-score normalization.
   (b) Calculate the correlation coefficient (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

3. Propose an algorithm, in pseudocode or in your favorite programming language, for the following:
   (a) The automatic generation of a concept hierarchy for nominal data based on the number of distinct values of attributes in the given schema.
   (b) The automatic generation of a concept hierarchy for numeric data based on the equal-width partitioning rule.

   Refer: 3.4.6 Histograms (page no. 106 &107)