36-350: Data Mining

Homework 3
Date: September 14, 2001 **Due: start of class September 21, 2001**

As usual for all problems, you must submit your work and your computer code. Showing your reasoning is also useful for partial credit.

1. In a 40-word document, the word "and" appears 3 times. Assuming the words were independent draws from a word distribution, give a 68% confidence interval on the true probability of the word "and" in the word distribution.

2. A cellular phone network wants to profile a customer according to where the customer is when calls are made. The city is initially divided into many small cells. Suppose a customer has made 20 calls in region A, 5 calls in region B, and 25 calls in region C. Region A is adjacent to B and B is adjacent to C. The areas of the regions are 3, 1, and 5, respectively, in square miles. The company is considering whether to merge regions A/B or B/C.

   (a) Which merge produces the most balance?

   (b) Which merge preserves density?

   (c) Which merge should the company use if they are interested in analyzing this particular set of calls at an abstract level?

   (d) Which merge should the company use if they are interested in computing the probability of a new call?

3. The data in `hw3.dat` represents a customer profile that we want to simplify.

   (a) Make a histogram of this data, with error bars, using 20 equally-spaced bins.

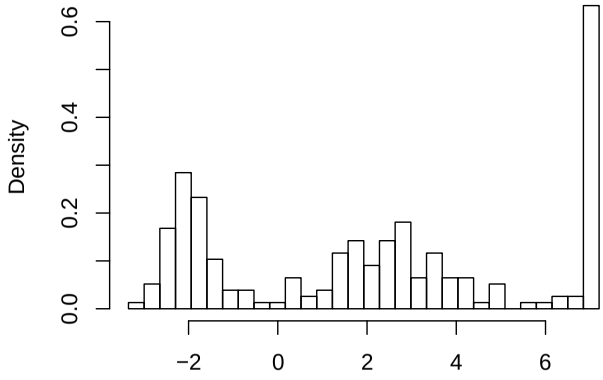
   (b) Use `bhist.merge` to reduce the number of bins to 8 and to 4.

   (c) Using the trace of chi-square differences, explain why the 8-bin and 4-bin solutions are interesting.

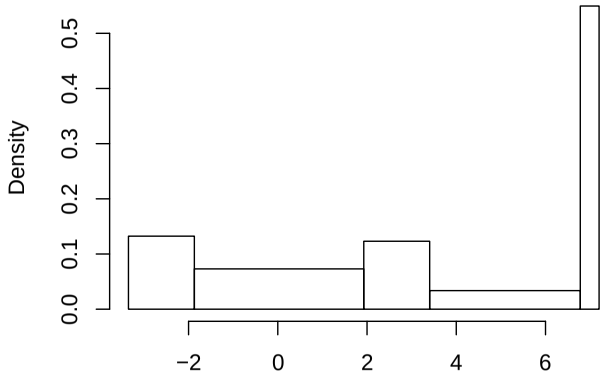
4. On the next page is shown the histogram of a dataset along with two different binnings. One used `bhist.merge` and the other used algorithm B (from the last problem set) to repeatedly merge adjacent bins.

   (a) Which binning used `bhist.merge` and which used algorithm B? (Look carefully at the bin breaks.)

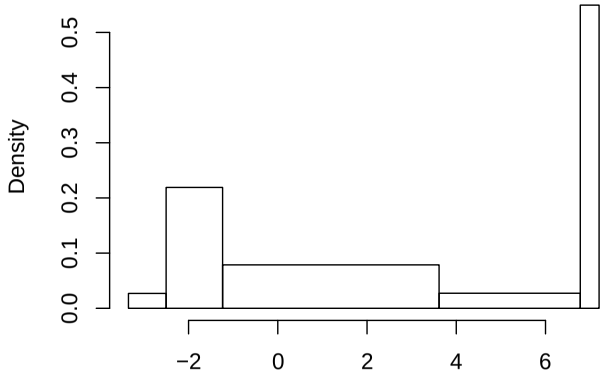
   (b) Describe one specific advantage that the binning from `bhist.merge` has over the binning from algorithm B for this data.



Histogram



Binning 1



Binning 2