AI Academy: Introduction to Data Mining
Week 2 Workshop

Workshop 2 contains 2 questions.

1 Sampling (10 points) [Chengyuan]

1. State the sampling method used in the following scenarios and give a reason for your answer. Choose from the following options: simple random sample with replacement, simple random sample without replacement, stratified sampling, progressive/adaptive sampling.

   (a) From the following population, {1,6,8,9,2}, a sample {1,1,6,6,2} was collected. Random

   (b) To learn the average income of software engineers working at Amazon, the population was divided into the following groups: Software Engineer, Senior Engineer, Staff Engineer, Principal Engineer, Distinguished Engineer. 10% of the staff from each group were selected for the study. Stratisfied

   (c) The data is collected in an experiment until the predictive model reaches 95% accuracy. Progressive

2. The U.S. Congress is made up of 2 chambers: 1) a Senate of 100 members, with 2 members from each state, and 2) a House of Representatives of 435 members, with members from each state proportional to that state's population. For example, Alaska has 2 Senators and 1 House representative, while California has 2 Senators and 53 House representatives. Both the Senate and the House are conducting surveys of their constituents, which they want to reflect the makeup of each chamber. You suggest that they use stratified sampling for this survey, sending surveys to a certain number of people from each state. Each survey will be sent to 1200 participants.

   (a) Why is stratified sampling appropriate here? More accurate for population

   (b) For the Senate survey, how many surveys would you recommend sending to people in Alaska? 1/435 *1200 = 2.7

   (c) For the House survey, how many surveys would you recommend sending to people in California? 53/435 * 1200 = 146.2

   (d) What are some advantages of the "Senate" approach and the "House" approach to stratified sampling?

2 Discretization (12 points) [Chengyuan]

Consider the following dataset:

No	HUMIDITY	TEMPERATURE	WINDY	PLAY TENNIS
1	67	47	FALSE	no
2	72	61	TRUE	no
3	67	58	TRUE	yes
4	73	55	FALSE	yes
5	78	61	TRUE	no
6	61	80	TRUE	yes
7	60	70	FALSE	yes
8	79	66	TRUE	no
9	69	52	FALSE	no
10	68	76	FALSE	yes
11	67	67	TRUE	yes
12	65	85	FALSE	yes
13	81	57	TRUE	no
14	65	59	FALSE	no
15	75	58	TRUE	no

1. Discretize the attribute HUMIDITY by binning it into 5 equal-width intervals (the range of each interval should be the same, and they should collectively span from the minimum to maximum values). Show your work by writing intervals for each bin.

2. Discretize the attribute TEMPERATURE by binning it into 5 equal-depth intervals (the number of items in each interval should be the same). Show your work.

3. Consider the following new approach to discretizing a numeric attribute: Given the mean ($\bar{x}$) and the standard deviation (σ) of the attribute values, bin the attribute values into the following intervals: $[\bar{x} + (k-1)\sigma, \bar{x} + k\sigma)$,
for all integer values k, i.e. $k = \ldots -4, -3, -2, -1, 0, 1, 2 \ldots$
Assume that the mean of the attribute HUMIDITY above is $\bar{x} = 70$ and that the standard deviation $\sigma = 5$. Discretize HUMIDITY using this new approach. Show your work.

4. Give an example of a situation where you would want to use equal-width binning, rather than equal-frequency.

60 63
64 67
68 72
73 77
78 81

60, 61
65, 65, 67, 67, 67
68, 69, 72,
73, 75
78, 79, 81

47 52 55,
57 58 58
59 61 61
47, 52, 55
76, 80, 85