## AIML Feature Engineering Assignment 2

QUESTION 1:

1. Briefly describe how ChiMerge works for the following simple dataset attributes.[4]

|  |  |  |
| --- | --- | --- |
| X | Y | Class |
| 1 | 2 | A |
| 3 | 4 | B |
| 5 | 6 | A |
| 7 | 8 | B |
| 9 | 10 | A |
| 11 | 12 | B |
| 13 | 14 | A |

Answer:

ChiMerge

Many classification algorithms require that the training data contain only discrete attributes.

To use such an algorithm when there are numeric attributes, all numeric values must first be converted into discrete values-a process called discretization.

ChiMerge, a general, robust algorithm that uses the x2 statistic to discretize (quantize) numeric attributes.

The ChiMerge algorithm follows the axis of bottom-up. x2 statistic helps to determine if the relative class frequencies of adjacent intervals are distinctly different or if they are similar enough to justify merging them into a single interval, where intervals are continuously merged until a termination condition is met.

Here are the steps we need to follow for any sample data:

1. We need to sort the feature, which helps identifying unique values and intervals.
2. Then it calculates the chi square value between each interval and its immediate next interval
3. If we have n intervals, we will get n-1 chi square values
4. Find the minimum chi square value out of these n-1 values
5. Find the first occurrence of chi square with the minimal value and its associated intervals
6. Merge these two intervals into one interval and remove the second interval as it is merged with first one
7. Go to step 3 and continue the process until the user specified max no. of intervals condition is met. Typically, the no. of max intervals is chosen as 2<=min intervals <=max intervals. In this problem, we are not choosing any min/max intervals and going by just comparing Chi sq value to the critical value.

Table 1: X attribute DataSet

|  |  |
| --- | --- |
| *X* | *Class* |
| *1* | ***A*** |
| *3* | ***B*** |
| *5* | ***A*** |
| *7* | ***B*** |
| *9* | ***A*** |
| *11* | ***B*** |
| *13* | ***A*** |

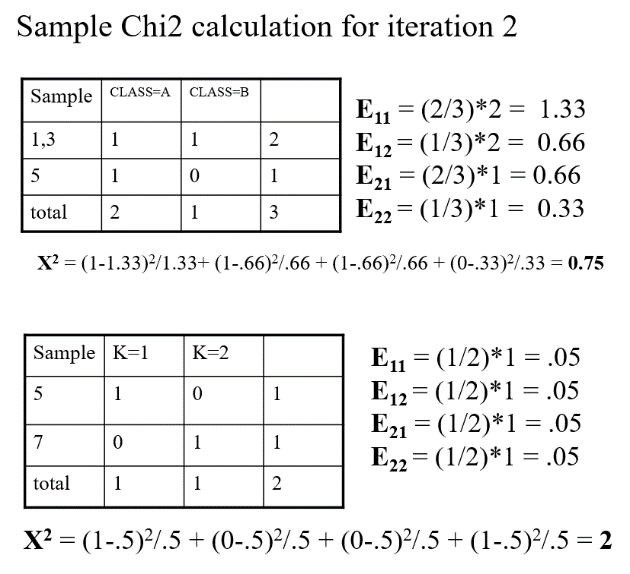
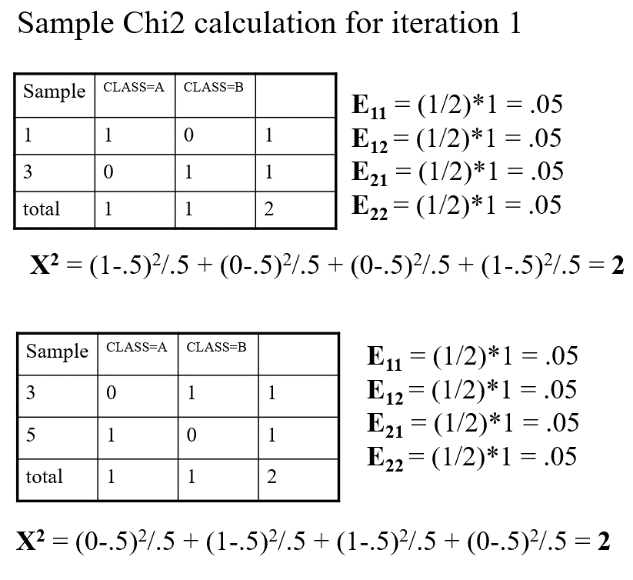
Table 2: Y attribute DataSet

|  |  |
| --- | --- |
| *Y* | *Class* |
| *2* | ***A*** |
| *4* | ***B*** |
| *6* | ***A*** |
| *8* | ***B*** |
| *10* | ***A*** |
| *12* | ***B*** |
| *14* | ***A*** |

Assumptions/Conditions:

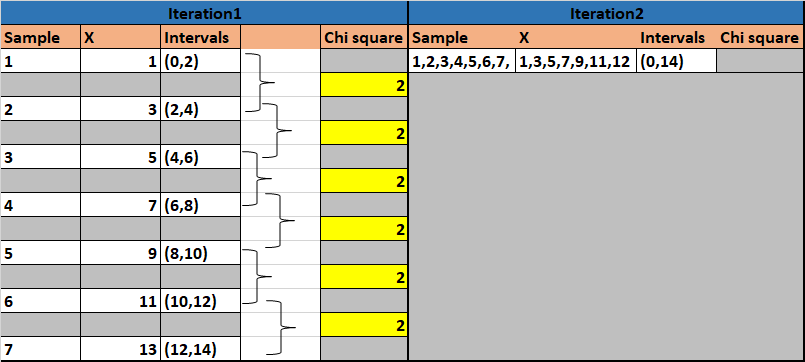
We have chosen the significance level as 0.05 and degrees of freedom is determined as 1 (no. of classes -1). From chi square distribution chart, the critical value is 3.84 i.e., merging of consecutive intervals with minimal chi square value can only happen when the X2 < 3.84.

An example calculation of chi square is given below, we will us the same method for calculating all chi squares X2 for X and Y intervals.



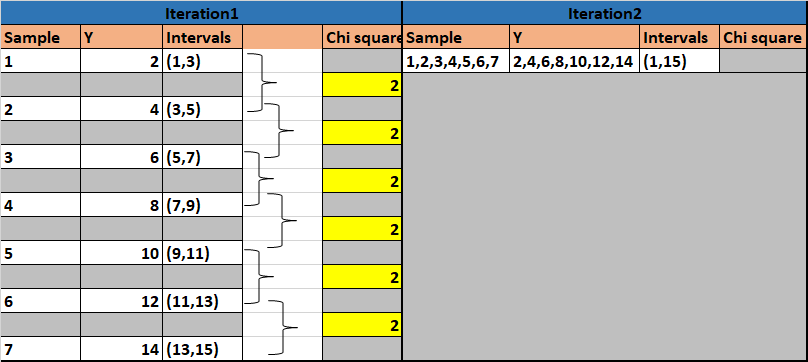
Feature X Intervals calculation:

For feature X, the values are already in sorted order and they are unique. Applying Chimerge algorithm for feature X produces following chi square values in iteration1 and all of them are same and less than 3.84. So merging all intervals after first iteration.



Feature Y Intervals calculation:

For feature Y, the values are already in sorted order and they are unique. Applying Chimerge algorithm for feature Y produces following chi square values in iteration1 and all of them are same and less than 3.84. So merging all intervals after first iteration.



Final Observation:

For the given dataset X and Y, the chi square values are always less than 3.84 critical value and all intervals got same Chi square , so we merged all intervals into a single interval. Since we did not define any min or max intervals, we merged till single interval is obtained as chi square was never more than 3.84.

* The final interval for X is (0 ,14). The Split point for X is 0,14.
* The final interval for Y is (1 ,15). The Split point for Y is 1,15.