**Solution 1 –**

Chi-Merge is a simple algorithm that uses the chi-squared statistics to discretize the numeric attributes. It is a bottom-up data discretization method. It checks each pair of adjacent rows in order to determine if the class frequencies of the two intervals are significantly different. It is a univariate approach in which only one attribute is examined at a time without regard to the values of the other attributes.

Discretization is the process of transferring continuous functions, models, variables, and equations into discrete counterparts. This process is usually carried out as a first step toward making them suitable for numerical evaluation and implementation on digital computers.

Chi-Merge algorithm tests the given hypothesis:

***The two adjacent intervals are independent.***

* If the hypothesis is confirmed, then the intervals are merged into a single interval.
* If the hypothesis is not confirmed, then the intervals are separated and remain separated.

Steps of the algorithm -

1. Sort the features or attributes and finds unique values
2. Divide each unique value as an interval
3. Calculate the chi square value between each interval and its immediate next interval
4. If there are “n” intervals, it will give “n-1” chi square values
5. Find the minimum chi square value out of these “n-1” values
6. Find the first occurrence of chi square with the minimal value and its associated intervals
7. Merge these two intervals into one interval and remove the second interval as it is merged with first one
8. 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**

Now we can leverage our knowledge about the Chi-Merge algorithm to solve the problem given above.

We have give N data set:

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

Assumptions/Conditions:

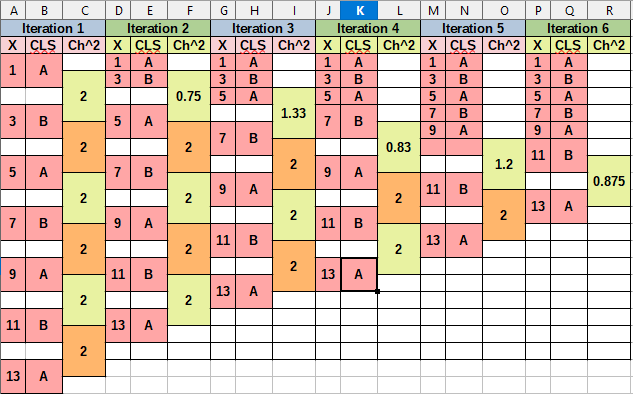
We will choose 2 as the ***maximum number of intervals***.

We will choose the significance level as 0.1 and degrees of freedom is determined as 1 (no. of classes -1). From chi square distribution chart, the critical value is 2.70 i.e. merging of consecutive intervals with minimal chi square value can only happen when the X2 < 2.70.

Determining Intervals for Feature X:

For feature X, the values are already in sorted order and they are unique. Applying Chi-merge algorithm for feature X produces following chi values and intervals after each iteration.

An example calculation of chi square is given below:



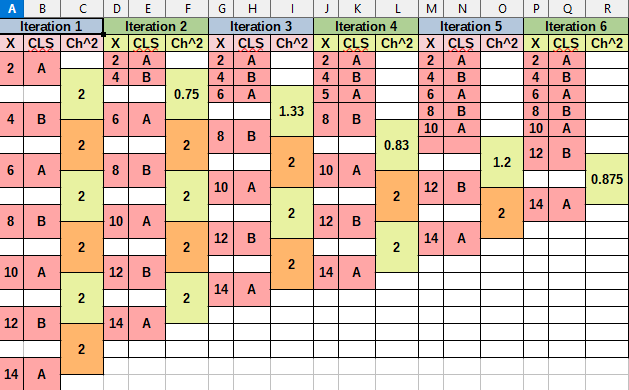
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Chi^2 calculation for iteration 1** | | | | |
| **Sample** | **Class=A** | **Class=B** |  | **E11=(1/2)\*1=0.05**  **E12=(1/2)\*1=0.05**  **E21=(1/2)\*1=0.05**  **E22=(1/2)\*1=0.05** |
| **1** | **1** | **0** | **1** |
| **3** | **0** | **1** | **1** |
| **Total** | **1** | **1** | **2** |
| **X^2 = (1 – 0.5)^2 / 0.5 + (0 – 0.5)^2 / 0.5 + (0 – 0.5)^2 / 0.5 + (1 – 0.5)^2 / 0.5 = 2** | | | | |
| **Sample** | **K = 1** | **K = 2** |  | **E11 = (1 / 2) \* 1 = 0.05**  **E12 = (1 / 2) \* 1 = 0.05**  **E21 = (1 / 2) \* 1 = 0.05**  **E22 = (1 / 2) \* 1 = 0.05** |
| **1** | **1** | **0** | **1** |
| **3** | **0** | **1** | **1** |
| **Total** | **1** | **1** | **2** |
| **X^2 = (0 – 0.5)^2 / 0.5 + (1 – 0.5)^2 / 0.5 + (1 – 0.5)^2 / 0.5 + (0 – 0.5)^2 / 0.5 = 2** | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Chi^2 calculation for iteration 2** | | | | |
| **Sample** | **Class=A** | **Class=B** |  | **E11 = (2 / 3) \* 2 = 1.33**  **E12 = (1 / 3) \* 2 = 0.66**  **E21 = (2 / 3) \* 1 = 0.66**  **E22 = (1 / 3) \* 1 = 0.33** |
| **1, 3** | **1** | **1** | **2** |
| **5** | **1** | **0** | **1** |
| **Total** | **2** | **1** | **3** |
| **X^2 = (1 – 1.33)^ 2 / 1.33 + (1 – 0.66)^ 2 / 0.5 + (1 – 0..66)^ 2 / 0.5 + (0 – 0.33)^ 2 / 0.5 = 0.75** | | | | |
| **Sample** | **K = 1** | **K = 2** |  | **E11 = (1 / 2) \* 1 = 0.05**  **E12 = (1 / 2) \* 1 = 0.05**  **E21 = (1 / 2) \* 1 = 0.05**  **E22 = (1 / 2) \* 1 = 0.05** |
| **5** | **1** | **0** | **1** |
| **7** | **0** | **1** | **1** |
| **Total** | **1** | **1** | **2** |
| **X^2 = (1 – 0.5) ^ 2 / 0.5 + (0 – 0.5) ^ 2 / 0.5 + (0 – 0.5) ^ 2 / 0.5 + (1 – 0.5) / 0.5 = 2** | | | | |

Conclusion: Based on the above result tables, we can see that the split points for X are (1, 13) and the final intervals for X are (1,12) and (13,13)

Determining Intervals for Feature Y:

For feature Y, the values are already in sorted order and they are unique. Applying Chi-merge algorithm for feature Y produces following chi values and intervals after each iteration.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Chi^2 calculation for iteration 1** | | | | |
| **Sample** | **Class=A** | **Class=B** |  | **E11 = (1 / 2) \* 1 = 0.05**  **E12 = (1 / 2) \* 1 = 0.05**  **E21 = (1 / 2) \* 1 = 0.05**  **E22 = (1 / 2) \* 1 = 0.05** |
| **2** | **1** | **0** | **1** |
| **4** | **0** | **1** | **1** |
| **Total** | **1** | **1** | **2** |
| **X^2 = (1 – 0.5)^ 2 / 0.5 + (0 – 0.5)^ 2 / 0.5 + (0 – 0.5)^2 / 0.5 + (1 – 0.5)^2 / 0.5 = 2** | | | | |
| **Sample** | **K = 1** | **K = 2** |  | **E11 = (1 / 2) \* 1 = 0.05**  **E12 = (1 / 2) \* 1 = 0.05**  **E21 = (1 / 2) \* 1 = 0.05**  **E22 = (1 / 2) \* 1 = 0.05** |
| **4** | **0** | **1** | **1** |
| **6** | **1** | **0** | **1** |
| **Total** | **1** | **1** | **2** |
| **X^2 = (0 – 0.5)^2 / 0.5 + (1 – 0.5)^2 / 0.5 + (1 – 0.5)^2 / 0.5 + (0 – 0.5)^2 / 0.5 = 2** | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Chi^2 calculation for iteration 2** | | | | |
| **Sample** | **Class=A** | **Class=B** |  | **E11 = (2 / 3) \* 2 = 1.33**  **E12 = (1 / 3) \* 2 = 0.66**  **E21 = (2 / 3) \* 1 = 0.66**  **E22 = (1 / 3) \* 1 = 0.33** |
| **2, 4** | **1** | **1** | **2** |
| **6** | **1** | **0** | **1** |
| **Total** | **2** | **1** | **3** |
| **X^2 = (1 – 1.33)^ 2 / 1.33 + (1 – 0.66)^ 2 / 0.66 + (1 – 0.66)^2 / 0.66 + (0 – 0.33)^2 / 0.33 = 0.75** | | | | |
| **Sample** | **K = 1** | **K = 2** |  | **E11 = (1 / 2) \* 1 = 0.05**  **E12 = (1 / 2) \* 1 = 0.05**  **E21 = (1 / 2) \* 1 = 0.05**  **E22 = (1 / 2) \* 1 = 0.05** |
| **6** | **1** | **0** | **1** |
| **8** | **0** | **1** | **1** |
| **Total** | **1** | **1** | **2** |
| **X^2 = (1 – 0.5)^2 / 0.5 + (0 – 0.5)^2 / 0.5 + (0 – 0.5)^2 / 0.5 + (1 – 0.5)^2 / 0.5 = 2** | | | | |

Conclusion: As we can see in above results table, the split points for Y are (2, 14) and final intervals for Y are (2,13) and (14,14)

**Observation:**

After applying the Chi-merge algorithm, we did multiple iterations, and the algorithm is applied for features X and Y, there was never occurred a chi-square with value greater than (>) 2.70. Based on the calculated values, we were able to merge the intervals until we reach the max number of intervals condition.