# AIML Feature ENGINEERING Assignment 2

Submission Date: 20 May 2020 11.55 PM

Weightage: 12%

ChiMerge is a supervised, bottom-up (i.e. merge based) data discretization method. It depends upon ChiSquare analysis: Adjacent intervals with the least ChiSquare values are merged together until the chosen stopping criterion satisfies.

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 |

1. Take the IRIS data set, obtained from the University of California-Irvine Machine Learning Repository (link provided in the reference section), as a data set to be discretized. Perform data discretization for each of four numeric attribute using ChiMerge method. (Let the stopping criteria be: max-interval 6). You need to write a small Python program ro do this to avoid clumsy numerical computations. Submit your simple analysis and your test results: split-points, final intervals and the well documented source program in Python Jupyter notebook. [8]

Notes

* This is a take-home assignment to be carried out by each learner individually and independently.
* This is programming exercises - requiring the datasets to be used – on Jupyter notebook environment.
* You may consult / discuss with other learners peripheral aspects such as the environment but not on solving the specific problems in terms of design or implementation.
* You have to write the appropriate Python code in Jupyter notebook to support you answers and submit with following nomenclature

Part (a) - FE\_Assignment2\_<Student\_ID>.docx

Part (b) - FE\_Assignment2\_<Student\_ID>.ipynb

* Provide appropriate justification when processing the data or arriving at the conclusions.
* In case of any further queries, if those are generic once, learners are encouraged to use discussion forums, otherwise they can reach out to me at [ppawar@wilp.bits-pilani.ac.in](mailto:ppawar@wilp.bits-pilani.ac.in).
* Manage your efforts properly as there is no scope to shift the deadlines announced above.

References:

1. Chapter 2, “Data”, Introduction to Data Mining eBook: Vipin Kumar, Pang-Ning Tan Michael Steinbach
2. [Kerber-ChimErge Original Paper](https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Kerber-ChimErge-AAAI92.pdf)
3. [Explanation about chimerge-discretization-algorithm](https://alitarhini.wordpress.com/2010/11/02/chimerge-discretization-algorithm/)
4. [Iris Data Set](http://archive.ics.uci.edu/ml/datasets/Iris)