

ASSIGNMENT 7
SPECIAL TOPICS IN EDUCATIONAL DATA MINING
PROFESSOR RYAN S.J.d. BAKER
CLUSTERING
DUE NOON, WEDNESDAY APRIL 15

The goal of this assignment is to cluster the data, using a clustering algorithm of your choice, using the same data set used in Assignment 5.

You should cluster the data using all variables except for P(SLIP | TRIO).

You must build the detector using an automated algorithm. You cannot simulate the algorithm in Excel. You can use any data mining package (e.g. SAS, R, Weka, KEEL, Stata, Matlab) you want.

Please turn in:

- The data set you input into the data mining package, if different than the original data set
- The model which assigns data points to clusters based on the features
- The assignment (probabilistic or non-probabilistic) of data points to clusters
- All data mining code you used to generate the outputs
- A document explaining how you completed the assignment
- An interpretation of what the clusters mean, based on analysis of either the variables used in the clustering, or the data points assigned to the clusters

You will be graded on completeness and comprehensibility of your hand-in, whether you correctly and validly apply the method you choose to this data, and whether the methods you chose fit the requirements of this assignment.

BONUS: The student who succeeds in producing the “most interesting” single cluster which is correctly (or reasonably) interpreted from the data will receive the bonus. Please identify which cluster you would like me to consider as most interesting for the bonus, and explain why it is interesting. (Hand-ins which do not do so will not be eligible for the bonus)