Data Mining, CSCI 347, Fall 2019
Rudimentary Rules, Sept. 4

Always try simple methods before complex ones
Simple methods are often quite effective

0R

- Predict the mean for a numeric class and the mode (most frequent) for a nominal class
- Considered "based line" accuracy
- When the attributes don't correlate with the class, some classifiers may do worse that this baseline

Confusion matrix - assigns letters a and b to the class values and provides expected class values in rows and predicted class values ("classified as") for each column.

1R

- Predict based on one attribute
- Use the attribute that produces the smallest error
- One-level decision tree
- Rule: If attribute = value1 then xxx else if attribute =value2 them xxx …
- Missing values handled by treating missing as another value
- Numerical values can be handled by discretizing them – list the values and the class values in order, and split when the class changes

How can simple algorithms work so well?

- some datasets really are simple
- some are so small/noisy/complex that nothing can be learned from them