Data Mining, CSCI 347, Fall 2019
Evaluation, Sept. 4

Testing is necessary because data mining methods do not adhere to the assumptions required by statistical theory. Typically use holdout data for testing.

In many cases 3 data sets are needed:

- Training dataset for selecting the learning algorithm
- Validation dataset for setting parameters on the chosen learning algorithms
- Testing dataset for determining the accuracy

Training, testing and validation datasets need to be:

- independent of each other
- representative of the underlying problem
- balanced with equal numbers of each of the outcomes (i.e. each of the class values)

How much data is enough? Depends on:

- Algorithms being used
- Complexity of the data
- Relative frequency of possible outcomes
- Required success rate – in some cases the cost of misclassification is much more serious than in other cases

From Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Gordon S. Linoff, Michael J. A. Berry

Once error rate is measured, re-bundle all three datasets and train again – but don't re-measure the error rate!

Vocabulary:

- Success rate for classification problems – the proportion of times the classifiers predicted correctly
- Error rate – the proportion of times the classifiers predicted incorrectly
- Resubstitution error rate – error rate resulting from testing on the training data

Know that the resubstitution error rate is likely to be highly optimistic