Data Mining, CSCI 347, Fall 2019
Evaluation – Holdout Methods and Confidence, Nov. 6

Recommendation – when have enough data (and don't need validation data for setting parameters), holdout 1/3 of data for testing (use 2/3rd for training)

Stratified holdout –represent each class in approximately equal proportions in the testing dataset as it was in the overall dataset

Repeated holdout method – Use multiple iterations, in each iteration randomly select a certain proportion of the dataset for training (possibly with stratification). Average the error rates on the different iterations to yield an overall error rate

Cross-validation – decide a fixed number of "folds" or partitions of the dataset. For each of the n folds, train with (n-1)/n of the dataset, test with 1/n of the dataset. Average the resulting n error estimations.

For stratified n-folds cross validation, make each split have instances with the class variable represented proportionally

Recommendation: 10-fold cross validation with stratification

Leave One Out Cross-Validation – another method for when data is limited. It is n-fold cross-validation where n is the number of instances in the dataset.
Advantages:

- Maximize the size of the training data set
- Deterministic, since no random sampling

Disadvantage:

- Guarantees a nonstratified sample

Bootstrap – uses sampling with replacement to form the training set

- Sample a dataset of n instances n times with replacement to form a new dataset of n instances and make this the training set
- Since sampling was done with replacement, this training set will have some repeated instances. Since some instances were repeated, some instances in the original set were not use. Make these unused elements be the testing dataset
- Commonly called 0.632 bootstrap because the training data will contain approximately 63.2% of the instances

For Bootstrapping the error estimate on the test data will be very pessimistic since the training only occurred on ~63% of the instances.

Therefore, combine it with weighted resubstitution error:

$$\text{Error} = 0.632 * \text{error}_{\text{test\_instances}} + 0.368 * \text{error}_{\text{training\_instances}}$$

Repeat the process several times with different replacement samples; average the results

Recommendation: Bootstrapping is the best way of estimating performance for small datasets

Confidence Levels

Vocabulary:

- Mean - average, sum/count
  - μ (Greek letter mu) – average of a population
  - $\bar{x}$ (x bar) – average of a sample drawn from a population
- Variance – the average of the squared differences from the mean
- Standard deviation, σ (Greek letter sigma) - square root of the variance
- Bernoulli event – an experiment with two possible outcomes
- Bernoulli process –a succession of Bernoulli events
- Sample of the data population – a subset of the data for a populate (often, collect a sample and extrapolate the results to the whole population)
- Confidence interval – an estimated value range with a given probability of covering the true population value

Normal distribution – continuous probability distribution shaped like a bell and symmetric around the mean

The normal distribution is actually a family of many different bell-shaped distributions. Each describable with only two parameters:

- Mean, μ
- Standard deviation, σ

Standardize a normal distribution by shifting it so that the mean is 0 (subtract the actual mean), and the standard deviation is 1 (divide by the actual standard deviation)

Properties of standardized normal distributions:

- mean = median = mode
- total area under the curve is 1

Confidence limits for the normal distribution:

$\Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25

Given a desired confidence level, c, the mean, μ, the standard deviation, σ, and the sample population, N, calculate the interval via:

1. Do a table look-up to find the z score where the Pr[X≥z] is (100%-c%)/2.
2. Interval is μ +/- z*(σ/sqrt(N))