Resampling Methods

They involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.

For example, cross validation can be used to estimate the test error associated with any given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility.

The Validation Set Approach

Randomly divide the set of observations to training and validation sets

Cons:

- The validation estimate can be highly variable
- Since the training set has only a few observations, the validation estimate tends to overestimate the test error.

Leave-one-out Cross Validation

1. Leave on observation out for validation
2. Get an estimate
3. Choose different observation
4. Repeat
5. Get the average estimate error

Advantages:

- Less bias since more observations
- Always the same result!

BUT, it requires a lot of computational power

k-fold Cross Validation

1. divide observations set in k, approximately equal, groups
2. use one on validation
3. repeat for all k
4. get average estimated error

Usually k=5 or k=10

Advantages:

- Much faster computationally
- Smaller variability than leave-one-out CV
- Much smaller bias AND variance

CV on Classification Problems

- Same idea, but the error is calculated as mismatched classifications
- Can be used to find K for KNN or order of LogReg

Bootstrap

Can be used to estimate the uncertainty associated with a given estimator or statistical learning method

It resamples the original data to create test data