

## 1. Summary:

This week, we focused on resampling methods to assist with model selection and evaluating the best and most accurate model for our situation. We started with learning about cross-validation, which entails splitting the dataset into partitions and training the model on all samples except one. Then, the last partition is the validation set, and our model is validated or tested on. Then we went into bootstrapping, which can be effective in understanding a population's parameters by resampling over a sample dataset.

## 2. Concepts:

### Cross-validation

- This is used to pick the best model without having access to your test set (which is how real life situations are typically)
- Validation set is distinct from train and test sets
- Hold out validation: split training dataset into two portions (train set is longer) with splits around 80-20 or 70-30 (the most common ones)
- Steps:
  - First create and train model using training data
  - Compute MSE on your validation set
- You have different validation MSEs based on how data is split and a lot of variation and also you could lose some information since not everything is used in training this way
- K-fold cross validation: split the data into k folds (positive integer) and all folds except one are used for training and the other is used to validate the model after training
- Go through all possible combinations using the average validation MSE (you can test this on different model complexities as well to find the best one)
- You aggregate the results by taking the average of your validation MSEs
- Leave one out CV is validation on only one data point
  - Not used practically as much
  - $K = n$
  - $CV = 1/n \sum_{i=1}^n ((y_i - \hat{y}_i)/(1 - h_i))^2$  where  $h_i$  is some leverage function
- Choosing a model: k fold for a bunch of models then to get the best one, fully train a model with that complexity on the entire training data this time
- Choosing k:
  - Increased training set leads to lower bias and higher variance
  - Higher k leads to training sets looking more similar (likely more similar results)

- Best  $k$ 's are typically 5, 10
- Validation generally has a minimum error with more complex models compared to the fully trained models because during validation, some of the training data is not used

Bootstrap Resampling:

- Something I've seen before
- Used to quantify uncertainty
- In practice we only have one dataset from the true population
- Bootstrap samples observations from our original dataset with replacement
- Then we train models based on these new datasets and then the model is tested on the data points
- We can use this to get the estimate for different parameters like the population mean, SD, variance, etc.
- This can also be used for confidence intervals as well as hypothesis testing
- When is bootstrap good for confidence intervals
  - Large  $n$  (sample size)
  - The distribution of the estimator needs to be regular

### 3. Uncertainties:

I feel like the topic of cross validation specifically makes a lot of sense intuitively in helping pick the ideal model and also leveraging the data points we have if we have a lot. I am curious more about how bootstrap exactly is used to figure out the best models. I understand that you can create and test models based on your bootstrap sampled datasets, but what exactly are we testing the model on? Is it the original dataset? Also, do we need to use our bootstrapped datasets to create models on every type of complexity that we want to test. Also, I know bootstrap is useful in the sense that we can estimate the parameters of the data, but I need to look into more how this works. Is it just that we get different datasets and then we take the mean of every data set, variance of every dataset, etc. or is there some other method I am missing? I will look into that to have a better understanding.