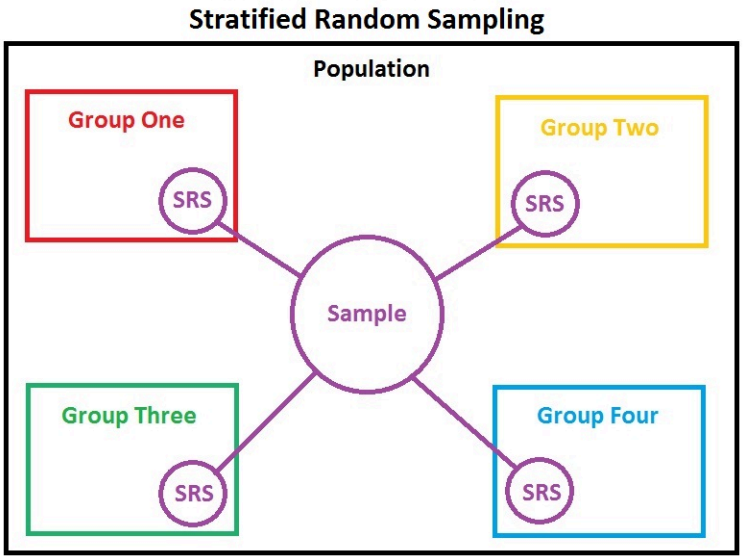
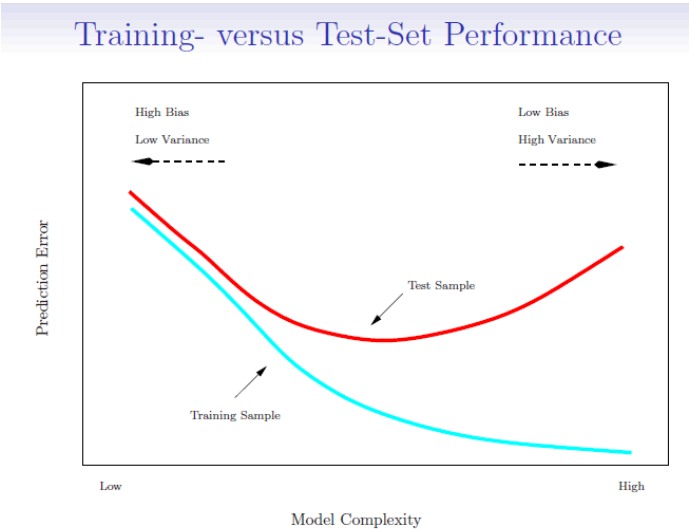
Statistical Learning - Ch 5 (Nov 30, 2018 10 am)
Resampling Methods
https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about

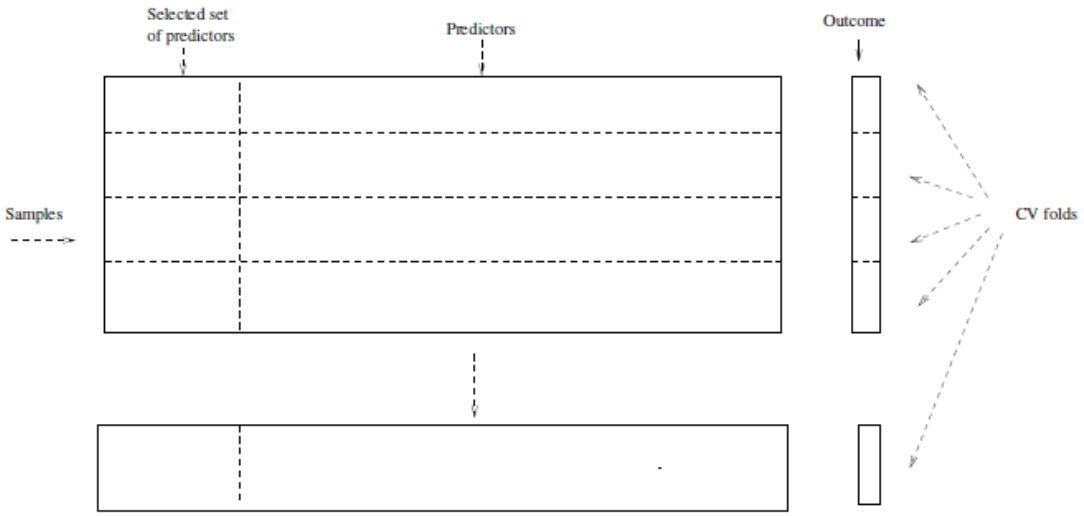
Definitions/Notes
Notes and discussion topics:

- Why is this chapter called **resampling** method?
    - Explain the difference between *population* and *sample* and provide a real life examples. Sample is a subset of population, which may be more cost effective and more feasible to acquire. Keywords in determining the difference between population and sample: feasibility, cost efficiency, etc.
    - Samples are often taken from a population with a sampling technique.
        - One of the sampling method is stratified random sampling where you "subgroup" your population and randomly sampling n number of sample from each stratum (group):
    - 

    - What is resampling?
        - To sample from the sample
    - 

        - Motivation: To estimate the test error

    - Why can't we use training error to estimate test error? We want to know how well it's doing in a general case, i.e., using the model to predict "unseen" data
- How do we estimate test error?
    - Best solution: *use a large designated test set*, but what is considered large dataset? A big test set is available only, e.g., if we have an unlimited dataset
    - Other ways:
        - Use an *adjustment* of training error.
            - Cp statistic, Aic, Bic -
        - *Validation (or Cross Validation) - holding out* a subset of training error.
            - How do you describe *cross validation* (in your own words)?
            - What metrics are used to estimate validation errors? MSE
            - Discuss the wrong vs. right way.. Where does bias come in when we cherry-pick predictors using CV (applying <u>step2</u> only, below)?
                - <u>Step 1</u>:Starting w/ 5000 predictors and 50 samples having the largest correlations with the labels
                - <u>Step 2</u>:We then apply a classifier only these 100 predictors



- Two resampling methods:
    - Cross-validation
        - Used for getting a good idea of the test set error of a model
            - **K-fold CV**

$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k$$

                - Best choices for k is 5-10, why?

            - **Leave One Out Cross Validation** (**LOOCV**).

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

- A special case! **With least-squares linear or polynomial regression**, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit!
- Advantages? You don't have to retrain your model on each fold. **The main advantage is we could simply use the formula** (above) to get the error estimate, but only on only for linear regression. Low bias (best estimate)
- Disadvantages? Computationally expensive, high variance

- **Bootstrap** - best to get the variability or standard deviation of estimate, confidence intervals and bias
  - Define bootstrap *in your own words*. Sampling with replacements and calculating statistics on the resampled sets.
  - How do we use bootstrap to estimate prediction error?

- Bootstrap
  - To sample datasets from the original dataset, with replacement, so that the resulting dataset is the same size as the original and may contain duplicates from the original
  - Is most useful to get an idea of the variability or standard deviation of an estimate and its bias
  - Can provide an estimate of the standard error of or confidence interval for a coefficient
- What's a typical procedure for including a CV estimate in model-building workflow:

- Test error
  - How well we do on new data that the model has never seen
- Training error
  - How well we do on data from which the model has been trained on
- Bias
  - How far off on the average the model is from the truth
- Variance
  - How much the estimate varies around its average
- The bias-variance tradeoff
- Training set, validation or hold-out set, test set
- Generalization error
- Overfitting
- K-fold CV
  - Partition data into K folds. Use one fold as the validation set and the remaining data to train on. Repeat using the remaining K-1 folds as the validation set. The average out-of-fold error is an estimate of the test error.
  - There is a bias-variance tradeoff in choosing K. In the extreme case when K=n (leave-one-out CV) we have high variance but low bias
- Leave-one-out CV
- Confidence interval

- Bootstrap
    - Can be used to calculate confidence intervals and estimate test error
    - The bootstrap estimate of error will be biased downward
- Out-of-bag sample
    - The approximately ⅓ of the original sample that is not included in the bootstrap sample
- Bagging
    - Training different models on bootstrap samples of your data

Questions

- The validation set error may tend to <u>overestimate</u> the test error for the model fit on the entire data set. Why? And is this still true when using k-fold cross-validation?
    - As an aside, should we take the average error across all k-fold validation set errors as our final validation error? → Yes
- Why is it that if you filter out your predictors (say, from 100,000 down to 1,000) that you need to apply cross-validation to both the unfiltered and filtered datasets?
    - Here, do they actually mean that you're filtering out the datasets (so that the number of samples is smaller), or are we just not taking the other features into account (leaving the sample size in tact)?
        - The number of features is reduced, not the number of samples.
- If we use 10-fold cross-validation as a means of model selection, the cross-validation estimate of test error may be biased upward or downward. Why? Let's discuss the course explanation:
    - There are competing biases: on one hand, the cross-validated estimate is based on models trained on smaller training sets than the full model, which means we will tend to overestimate test error for the full model.
    - On the other hand, cross-validation gives a noisy estimate of test error for each candidate model, and we select the model with the best estimate. This means we are more likely to choose a model whose estimate is smaller than its true test error rate, hence, we may underestimate test error. In any given case, either source of bias may dominate the other.
- Since estimating the prediction error using the bootstrap method is complicated, when should we actually use the bootstrap method in the real world and in Kaggle competitions (where cross-validation is the norm)?
- If we have a small dataset should we choose K in K-fold cross-validation to be large or small?
- Can we use the bootstrap to estimate other statistics such as the median or mode of the error?
- About two-thirds of the original data points appear in each bootstrap sample. Can you prove this?
- Are the bootstrap estimates unbiased?