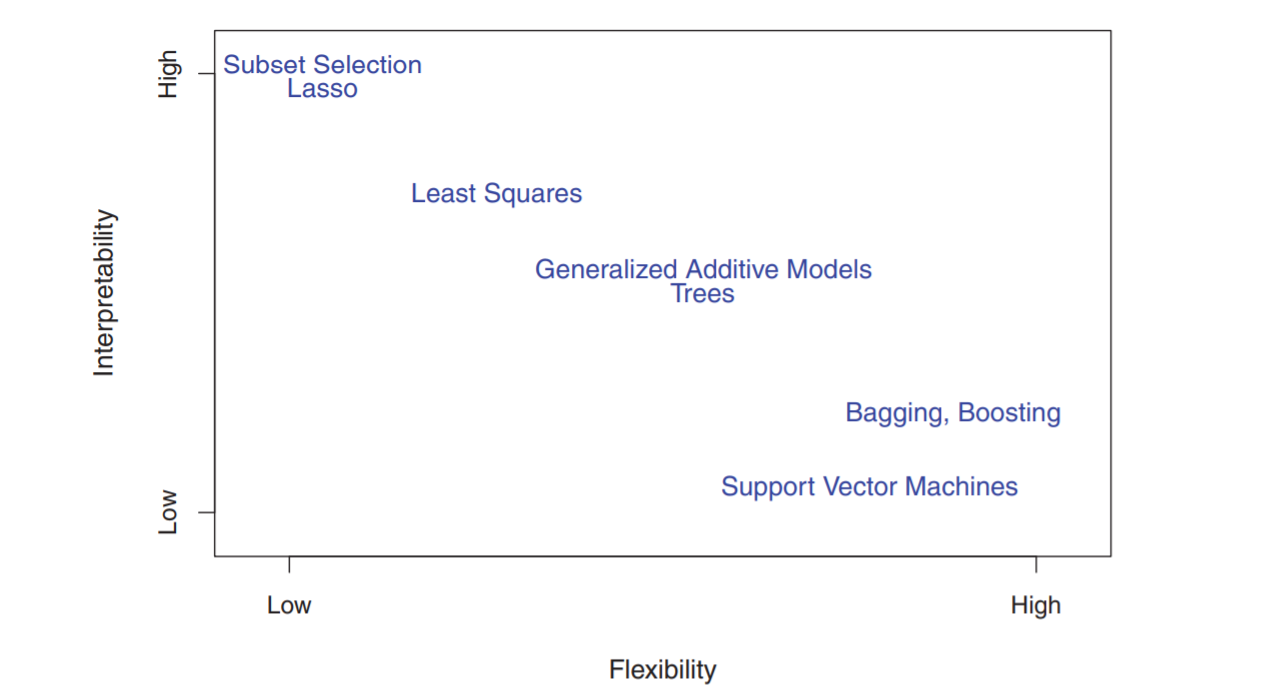
**Notes for Introduction to Statistical Learning**

**Chapter 2 – Statistical Learning**

* We estimate *f* for
  + *prediction (for a given X what is Y)*
  + *inference (how does Y change with respect to X)*
* Accuracy of our model depends on *reducible error*  and *irreducible error(e)*
* **Parametric** Models – reduces the problem of estimating f down to one of estimating a set of parameters (coefficients)
  + First, make an assumption about the functional form ie linear Y = B0 + B1X1 + …+ BnXn
  + After model has been selected fit to training data ie estimate the coefficients B0 … Bn using ordinary least squares
* **Non-parametric** Models – do not make explicit assumptions about the functional form of *f*
  + Main advantage – by avoiding making assumptions have the potential to accurately fit a wider range of possible shapes for *f*
  + Main disadvantage – since we do not reduce problem to estimating parameters, need a large number of observations
  + Example – thin-plate spline
* Tradeoff Prediction Accuracy vs Model Interpetability
  + If we are mainly interested in inference, then restrictive models are muchFo more interpretable
* In unsupervised learning there is labeled training data we only see X vars but no outcome variable – ie cluster analysis like market segmentation
* Quantative Y var = regression, qualitative Y var = classification
* Measuring Model accuratcy
  + For regressions, mean squared error (MSE) = Ave (y – yhat)^2
  + As model flexibility increases training MSE will decrease and test MSE will increase
* Cross validation = method for estimating test MSE using the training data
* Bias-Variance Tradeoff =
  + Expected MSE = variance of *f + (*bais of *f)^2 +* variance of error term *e*
* To minimize expected error need to select a statistical method with low variance and low bias
* **Variance =** the amount by which fhat will change if we estimated using a different training set
  + High variance of a method means small changes in training set will lead to large changes in *f*
  + More flexible methods have high variance
* **Bias =** the error that is introduced by approximating real life problems
  + Ie – linear regressions assume linear relationship b/w dependant & outcome vars – unlikely in real life so linear regression has linear bias
* More flexible models mean high variance and bias will decrease
  + As we increase flexibility = variance increases slower than bais decreases but at certain point bais stops decreasing and variance continues increasing
* Good test set performance of a statistical learning method requires low variance as well as low squared bias

**Classification**

* **Bayes Classifier** – able to minimize the error rate for a classifier by a simple Bayes classifier that assigns each observation to the most likely class given its predictor values
  + Pr(Y = j | X = x0)
* Bayes classifier prediction is determined by the Bayes decision boundary
  + Solve for lowest possible *Bayes error rate*  1 – E(maxPr(Y = y|X) (expectation avarges the probability of over all possible boundaries of X
* K-Nearest Neighbors – KNN identifies the K closest objects and estimates conditional probability based on proportion of x that are a certain class
  + KNN applies Bayes rule and classiﬁes the test observation x0 to the class with the largest probability
  + K = 100 leads to a classifier with low bias but high variance (very flexible)
  + Higher K leads to decreasing training error but U shaped test error

**Summary of Chapter 2:** In both the regression and classiﬁcation settings, choosing the correct level of ﬂexibility is critical to the success of any statistical learning method. The bias-variance tradeoﬀ, and the resulting U-shape in the test error, can make this a diﬃcult task

**Chapter 3 – Linear Regression**

* Linear regression is a useful tool for predicting a quantitative response
* To measure closeness we minimize the least squares error
* **Residual Sum of Squares (RSS)** = e1^2 + e2^2 +….+en^2 where e = y1 – yhat
* **Least squares** approach chooses coefficients which minimize RSS
* We assume the coefficients B0 & B1 are unbiased with a large enough sample size because the estimates of coeffcients are close to the population(real) estimates
* The **standard error** is the average amount by which an estimate differs from the actual value
  + Standard error reduces with larger N
  + We general don’t know o^2 but estimate of o^2 is known as the residual standard error (RSE) = square root of RSS / (n-2)
* **Variance** is standard error squared
* **Confidence intervals** calculated from RSE
  + A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter
  + Ie There is a 95% chance that the interval will contain the true value of the coefficient
* **Standard errors** also used to perform hypothesis test
  + Null hypothesis is that there is no relationship b/w Y and X ie B1 = 0
  + Alternative hypothesis is that there is a relationship b/w Y and X ie B1 != 0
* Compute **t-statistic** to test hypothesis = B1 – 0 / SE(B1)
  + If there is no relationship b/w X and Y we expect t will have a t-distribution with n-2 degrees of freedom
* From t-statistic we compute the **p-value** which is the probability of observing any number equal to |t| or larger in absolute value, assuming B1=0.
  + A small p-value is better – means low probability of seeing that value if we assume B1 = 0 s
  + Small p = reject the null hypothesis
* To assess how well model fits data we use – residual standard error (RSE) & R^2 statistic
  + **RSE =** an estimate of the standard deviation of error term – average amount the response will deviate from true regression line
  + RSE is a measure of lack of fite
* **R^2 =** Total Sum of Squares (TSS) – Residual Sum of Squares (RSS) / TSS
  + **R^2** measures the proportion of variability in Y that can be explained using X
    - Larger R2 is better
  + **TSS =** measures the total variance in the response Y
* Corr(X,Y) (look up formula) is a measure of the linear relationship between X and Y
* **In multiple linear regressions** ie more than 1 coefficient, it is important to check the correlation between X variables before estimating coefficients

We perform multiple linear regression to answer the following questions

* **Is at least one of the predictors X1,X2..Xp useful in predicting the response?**
  + Null hypothesis is B1=B2=B3..Bp=0
  + Use the F-statistic computed from Total Sum of Squares and Residual Sum of squares
    - If no relationship F statistic close to 1, reject null if F greater than 1
    - With small n need a larger F statistic to reject
    - The
  + The square of each t-statistic is the corresponding F-statistic if that varwas omited
* **Do all predictors help to explain Y or is only a subset useful?**
  + Can use following techniques to select subset of most useful variables – Mallows Cp, Akaike information criterion, Bayesian information criterion and adjusted R
  + 3 main approaches
    - **Forward selection –** begin with a null model ie just intercept but no predictors – we then add the variable that results in lowest RSS, we then add to that model variable that results in the lowest RSS (very similar to decision trees use node that reduces entropy by most)
    - **Backward selection –** start with all variables in the model and remove the variable with largest p-value – least significant var, continue until reach stopping rule- **Cannot be used if p>n**
    - **Mixed selection –** Start with no vars in model, continue adding based on reducing RSS, but when one p-value falls below threshold, remove that var until all p-values are below desired threshold
* **How well does the model fit the data?**
* **Given a set of predictor values , what response value should we predict and how accurate is our prediction?**