# Notes: Introduction to Statistical Learning

A. John Woodill

Revision 1 – October 2018

## Contents

# 1 Introduction

The following notes are from Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. For updated document and Python code see https://github.com/johnwoo Intro-to-Stat-Learning. For online book content see www.statlearning.com

# 2 Statistical Learning

## 2.1 Overview

Suppose we observe a quantitative response $Y$ with different predictors, $X_1, X_2, ..., X_p$. We assume there is some relationship between $Y = X_p$. A general form is,

$$Y = f(x) + \epsilon$$

where $f$ is some systematic information that $X_p$ provides about $Y$. $\epsilon$ is the random error term, which is **independent** of $X_p$ and mean zero.

Statistical learning refers to approaches that estimate $f$ where we can estimate to provide predictions or inference.

### 2.1.1 Predictions

Predictions assume a set of inputs $X_p$ are available but outputs $Y$ may not be. Predictions follow the form,

$$\hat{Y} = f(\hat{X})$$

The accuracy of $\hat{Y}$ by predicting $f(\hat{X})$ is not perfect, which introductions an error of two quantities:

**Reducible Error**: Error can be improved with appropriate modeling strategies.

**Irreducible Error**: variability of $\epsilon$ affects the accuracy of the prediction; thus, cannot reduce the error introduced by $\epsilon$. The error may contain (1) unmeasured variables, (2) unmeasured variation.

Derive reducible and irreducible errors by simply differencing $Y - \hat{Y}$, then find mean squared.

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(\hat{X})]^2$$

$$E(Y - \hat{Y})^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}}$$

Thus, the irreducible error depends on the variation in the error. Statistical Learning focuses on improving (minimizing) the reducible error. Note that the irreducible error will always provide an upper bound on the accurary of the prediction, which is almost always unknown in practice

### 2.1.2 Inference

Inference relates to understanding the relationship between $X$ and $Y$, or how $Y$ changes in response to $X$.

- Which predictors are associated with the response?

- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

### 2.1.3 Estimating $f()$

Our goal is to apply statistical learning method to train data to estimate and unknown function $f$. Methods include parametric and nonparametric methods.

**Parametric**: Methods that use distributional assumptions are called parametric methods, because we estimate the parameters of the distribution assumed for the data. In OLS, assumptions about function form are linear with fixed parameters. No matter how much data you have, there will always be fixed parameters.

Examples:

- Logistic Regression
- Linear Discriminant Analysis
- Perceptron
- Naive Bayes
- Simple Neural Networks

Advantage: simplifies estimating $f()$ because it is easier to estimate a set of parameters, $\beta_0, \beta_1$.

Disadvantage: model does not usually match $f()$.

**Non-Parametric**: Methods do not make explicit assupmtions about the functional form of $f()$. Goal is to get as close to the data points as possible without being to rough or widdly.

Examples:

- k-Nearest Neighbors
- Decision Trees like CART and C4.5
- Support Vector Machines

Advantange: potential to accurately fit a wider range of possible shapes for $f()$.

Disadvantage: do not reduce the problem of estimating $f$ to a small number of parameters, thus a large number of observations is required to accurately estimate $f$.

### 2.1.4 Prediction Accuracy versus Model Interpretability

*Why would we ever choose to use a more restrictive method instead of a very flexible approach?*

Restrictive models, such as linear models, are more interpretable. In contrast, flexible approaches, such as splines, can provide complicated estimates of $f()$ that may improve prediction accuracy.

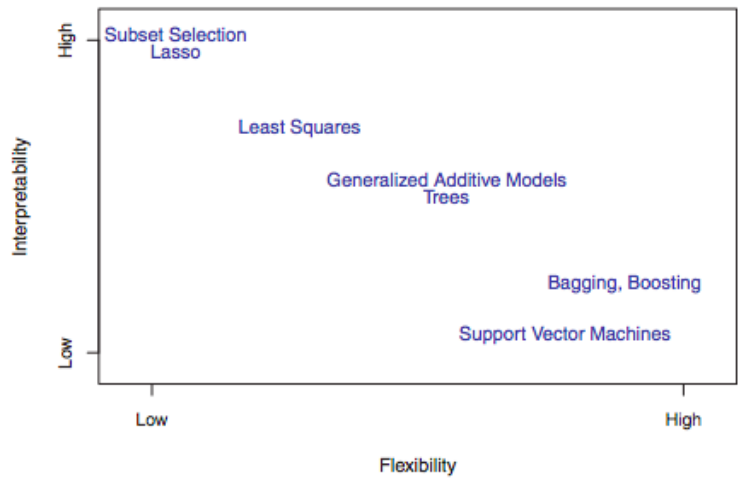


**FIGURE 2.7.** *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

The choice of modeling strategy comes down to the end goal: prediction or inference. Less flexible models are easy to interpret, thus prefered when the goal is inference. If prediction is prefered, then more flexible models may be prefered – although, more flexible models are not always prefered for prediction acurracy due to overfitting.

### 2.1.5 Supervised Versus Unsupervised Learning

**Supervised**: Each observation of the predictor measurements $x_i$ there is an associated response measurement $y_i$. Goal is to fit a model that relates to the response predictors with an aim to accurately predict the response variable in the future.

Examples:

- Linear Regression
- Logistic Regression
- GAM
- Boosting
- Support Vector Machines

**Unsupervised**: Observations of the predictor $x_i$ does not contain a response variable, $y$.

Examples:

- Cluster Analysis
- PCA

### 2.1.6 Regression versus Classification Problems

Variables are characterized by quantitative or qualitative (categorical). Quantitative values are numerical whereas qualitative variables take values in classes or categories.

- Regression Analysis: Uses quantitative variables
- Classification Analysis: Uses qualitatie variables

We select statistical learning methods based on the response variable being quantitative or qualitative.
**Note**: distribution of predictors being qualitative or quantitative is less important.

## 2.2 Assessing Model Accurary

### 2.2.1 Measuring Quality of Fit

To assess performance of statistical learning methods, we need to quantify the extent to which the predicted response value is close to the true value. The most commonly-used measure is mean-squared-error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Calculating training MSE (MSE on training data) doesn't tell us much about out-of-sample performace, which is prefered. Suppose $(x_0, y_0)$ are previously unseen test observations. The test MSE is

$$Ave(y_0 - \hat{f}(x_0))^2)$$

It is important to note that minimizing the training data provides no guarantee that the method will also minimize the test data.

There is a trade-off between inflexibility versus flexible models. Degrees of freedom define the flexibility of a curve. A more restricted (smoother) curve has fewer degrees of freedom than a wiggly curve. As flexibility increases training MSE declines monotonically.

In the figure below, as the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE. As model flexibility increases, training MSE will decrease, but the test MSE may not.

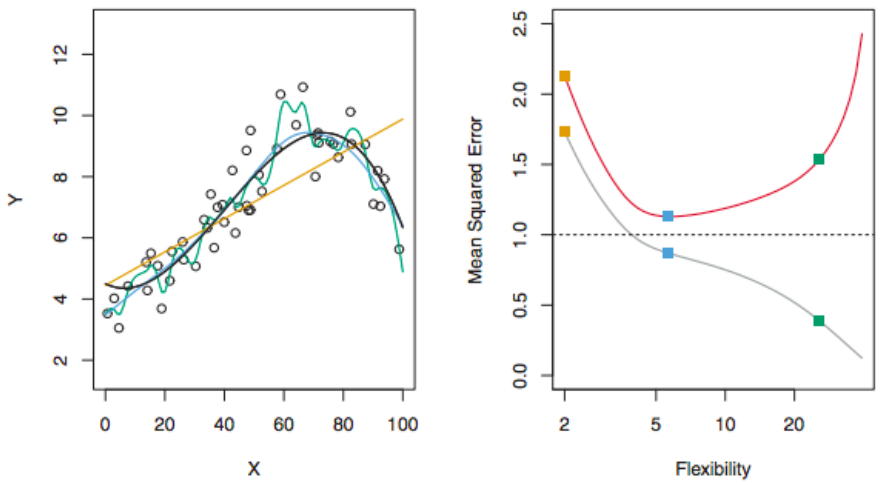


**FIGURE 2.9.** Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

**Note**: Overfitting occurs when a small training MSE but a large test MSE exists. This happens because the training model is find patterns in the data and not establishing the signal.

### 2.2.2 The Bias-Variance Trade-Off

The U-shaped observed in the test MSE curve is a result of two competing properties in statistical learning:

The test MSE, for a given value $x_0$, can be decomposed into the sum of three fundamental quantities: variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$, and the variance of the error terms $\epsilon$,

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{Expected Test MSE}} = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

To minimize the expected test error, we need to select a stat. method that achieves a low variance and a low bias.

- Variance: amount by which $\hat{f}$ changes if estimated using different raining data.
- Bias:error that is introduced by approximating a real-life problem

Variance between training data sets shouldn't change $\hat{f}$ too much; however, methods that are more flexible have higher variance that will shift the MSE larger whereas restricted methods have low variance and will only cause small shifts.

In terms of bias, the inverse is true. Restricted methods do not identify the true response variable, which results in large bias; however, flexible methods are usually better at predicting the true response variable which provides less bias.

**Bias-Variance Trade-Off**

- Flexible Methods: Variance increases and bias will decrease MSE

- Restricted Methods: Variance decreases and bias will increases MSE

**Note**: The challenge lies in finding a method for which both the variance and the squared bias are low.

### 2.2.3 Classification Strategy

Model accuracy transfers over to classification problems. The most common approach is to quantify the accuracy of $\hat{f}$ using a training error rate, or the proportion of mistakes that are made to the training observations,

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y_i})$$

where $I(y_i \neq \hat{y_i})$ is an indicator variable that equals 1 if $y_i \neq \hat{y}_i$, and zero if $y_i = \hat{y}_i$. If $I(y_i \neq \hat{y_i}) = 0$, then the observation was classified correctly. The test error is calculated as,

$$Ave(I(Y - 0 \neq \hat{y}_0))$$

**The Bayes Classifer**

The error rate acn be classified by assigning each observation to themost likely class, given its predictor values.The Bayes Classifer is,

$$Pr(Y = j|X = x_0)$$

or the propability that $Y = j$ given the observed predictor vector $x_0$. The Bayes Classifer establishes a Bayes decision boundary that falls on one side or the other of the classification.

Bayes error rate maximizes the probability of selecting,

$$1 - E(\max_j PR(Y = j|X))$$

and is analogous to the irreducible error.

**K-Nearest Neighbors**

In theory, Bayes Classifier is the gold standard, but we don't always know the conditional distribution of $Y$ given $X$;thus, we need to stimate the probability – K-Nearest Neighbor (KNN) is one method.

Given a positive integer $K$, and a test observation $x_0$, the KNN classifer identigies the $K$ points in the training data that are closest to $x_0$, represented by $N_0$. Condition probabilities are estimated for class $j$ as a fraction of $N_0$ whose response values equal $j$:

$$Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = J)$$

The choice of $K$ has a drastic effect on the classifer obtained; small $K = 1$ provides a boundary that is overfly flexible and has a low bias but high variance. As $K$ increases, method becomes less flexible and is closer to linear (high bias low variance). No strong relationship between test and train error rates. Flexible $K = 1$ have a low training rate (0), but test error will be high.

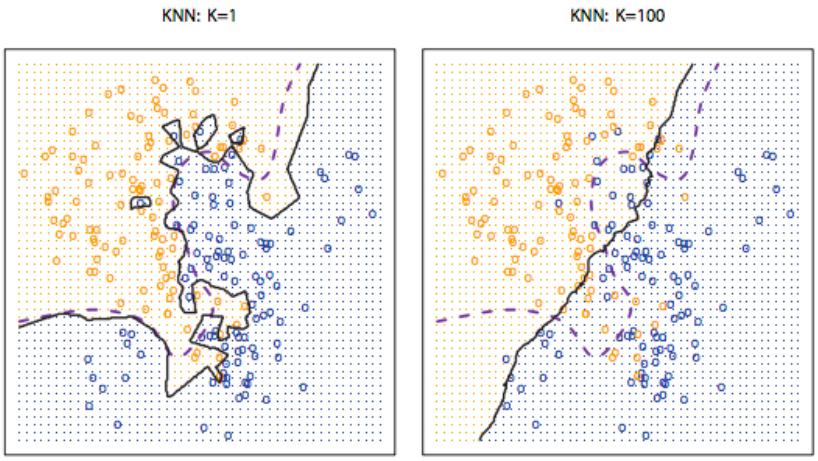


**FIGURE 2.16.** *A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.*

**Note:** In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method.