CS5805 : Machine Learning I
Lecture #7

Reza Jafari, Ph.D

Collegiate Associate Professor
rjafari@vt.edu

What Is statistical Learning?

- Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product (TV, radio and newspaper).

What Is statistical Learning?

- Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product (TV, radio and newspaper).
- It is not possible for our client to directly increase sales of the product.

What Is statistical Learning?

- Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product (TV, radio and newspaper).
- It is not possible for our client to directly increase sales of the product.
- On the other hand, they can control the advertising expenditure in each of the three media.

What Is statistical Learning?

- Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product (TV, radio and newspaper).
- It is not possible for our client to directly increase sales of the product.
- On the other hand, they can control the advertising expenditure in each of the three media.
- Our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

What Is statistical Learning?

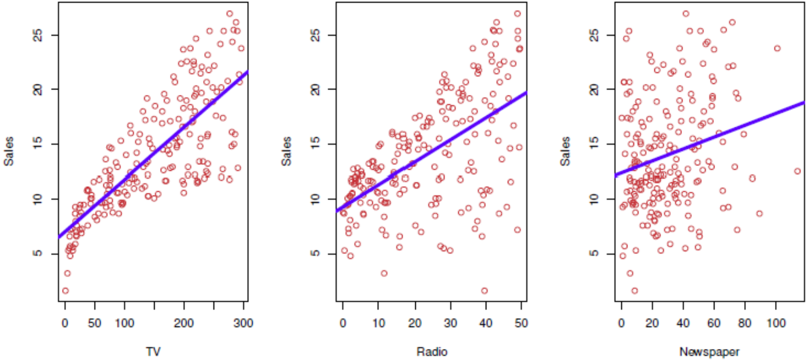
- Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product (TV, radio and newspaper).
- It is not possible for our client to directly increase sales of the product.
- On the other hand, they can control the advertising expenditure in each of the three media.
- Our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.
- The advertising budgets→ *input* variables while sales→ *output* variable.

What Is statistical Learning?

- Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product (TV, radio and newspaper).
- It is not possible for our client to directly increase sales of the product.
- On the other hand, they can control the advertising expenditure in each of the three media.
- Our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.
- The advertising budgets→ *input* variables while sales→ *output* variable.
- Input variables (independent variables, predictors, features, variables) denoted by symbol X and Output variable(response, dependant variable) denoted by symbol Y.

Example

- The dotted displays sales in thousands of units as a function of TV, radio and newspaper in thousands of $.



Statistical Learning

- Suppose that a quantitative response Y is observed and p different predictors $X_1, ..., X_p$.

Statistical Learning

- Suppose that a quantitative response Y is observed and p different predictors $X_1, ..., X_p$.
- We assume that there is a relationship between Y and $X = (X_1, ..., X_p)$, which can be written as :

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

Statistical Learning

- Suppose that a quantitative response Y is observed and p different predictors $X_1, ..., X_p$.
- We assume that there is a relationship between Y and $X = (X_1, ..., X_p)$, which can be written as :

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

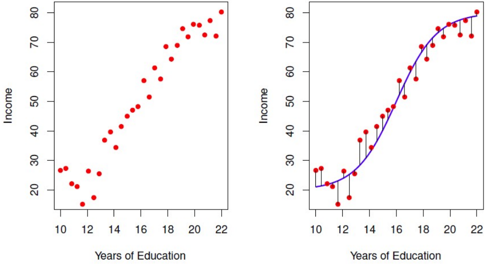
- f is some unknown function of $X_1, ..., X_p$ and ϵ is a random *error term* which is independent of X and has zero mean.

Statistical Learning

Statistical Learning: $\boxed{\text{refers a set of approaches for estimating } f}$.

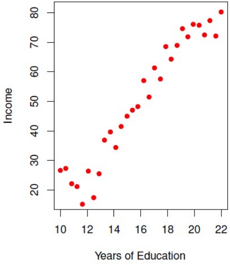
Statistical Learning

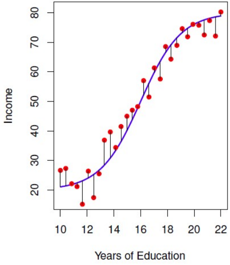
- The red dots are the observed values of income (in tens of thousands) and years of education for 30 individuals.



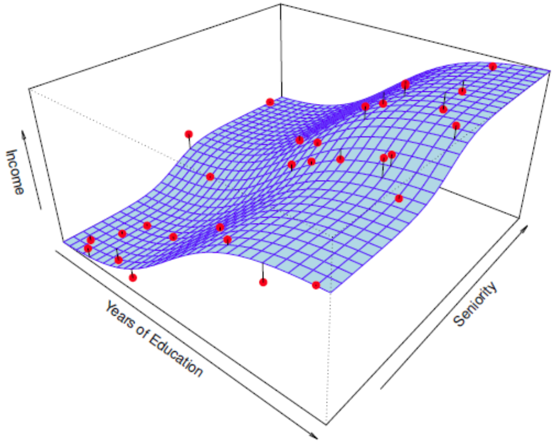
Statistical Learning

- The red dots are the observed values of income (in tens of thousands) and years of education for 30 individuals.
- The blue curve the true underlying relationship between income and years of education, which is generally unknown. And the $\boxed{\text{black}}$ lines are the positive or negative errors.





Statistical Learning-3 dimensional



Why Estimate f?

- There are two reasons that we may wish to estimate f: prediction & inference.

Prediction

- In many situations, set of inputs X are readily available, but the output Y cannot be easily obtained. Since the error term averages to zero, we can predict Y using

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

where $\hat{f}$ represents estimate for f and $\hat{Y}$ represents prediction for Y.
- Accuracy of $\hat{Y}$ depends on reducible error and the irreducible error.

$$\begin{aligned} E(Y - \hat{Y})^2 &= E\left[f(X) + \epsilon - \hat{f}(X)\right]^2 \\ &= \underbrace{\left[f(X) - \hat{f}(X)\right]^2}_{Reducible} + \underbrace{Var(\epsilon)}_{Irreducible} \end{aligned}$$

Inference

Inference

- We are often interested in understanding the association between Y and $X_1, \ldots, X_p$.

Inference

Inference

- We are often interested in understanding the association between Y and $X_1, ..., X_p$.
- In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y.

Inference

Inference

- We are often interested in understanding the association between Y and $X_1, ..., X_p$.
- In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y.
- In the inference, we are interested in answering the following questions:
  - Which predictors are associated with the response? It

Inference

Inference

- We are often interested in understanding the association between Y and $X_1, ..., X_p$.
- In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y.
- In the inference, we are interested in answering the following questions:
  - Which predictors are associated with the response? It
  - What is the relationship between the response and each predictor?

Inference

Inference

- We are often interested in understanding the association between Y and $X_1, ..., X_p$.
- In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y.
- In the inference, we are interested in answering the following questions:
  - Which predictors are associated with the response? It
  - 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?

Example

Real Estate

- We seek to relate values of homes to input such as crime rate, zoning, distance from river, air quality, schools, income level of community, size of houses.

Example

Real Estate

- We seek to relate values of homes to input such as crime rate, zoning, distance from river, air quality, schools, income level of community, size of houses.
- *How much extra will a house be worth if it has a view of the river?* This is an inference problem.

Example

Real Estate

- We seek to relate values of homes to input such as crime rate, zoning, distance from river, air quality, schools, income level of community, size of houses.
- *How much extra will a house be worth if it has a view of the river?* This is an inference problem.
- One may simply be interested in predicting the value of a home given its characteristics: is this house under- or over-valued? This is a prediction problem.

Example

Real Estate

- We seek to relate values of homes to input such as crime rate, zoning, distance from river, air quality, schools, income level of community, size of houses.
- *How much extra will a house be worth if it has a view of the river?* This is an inference problem.
- One may simply be interested in predicting the value of a home given its characteristics: is this house under- or over-valued? This is a prediction problem.

- Whether the goal is prediction, inference, or a combination , different methods for estimating f may be appropriate.

Example

Real Estate

- We seek to relate values of homes to input such as crime rate, zoning, distance from river, air quality, schools, income level of community, size of houses.
- *How much extra will a house be worth if it has a view of the river?* This is an inference problem.
- One may simply be interested in predicting the value of a home given its characteristics: is this house under- or over-valued? This is a prediction problem.

- Whether the goal is prediction, inference, or a combination , different methods for estimating f may be appropriate.
- E.g., linear models allow a simple and interpretable inference, but may not yield as accurate predictions as other approaches.

How do We Estimate f?

- Observations divides into training data that is used to teach our method how to estimate f.

How do We Estimate f?

- Observations divides into training data that is used to teach our method how to estimate f.
- Let x_{ij} represent the value of j^{th} predictor for observation $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$.

How do We Estimate f?

- Observations divides into training data that is used to teach our method how to estimate f.
- Let x_{ij} represent the value of j^{th} predictor for observation $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$.
- Let y_i represent the response variable for the i^{th} observation.

How do We Estimate f?

- Observations divides into training data that is used to teach our method how to estimate f.
- Let x_{ij} represent the value of j^{th} predictor for observation $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$.
- Let y_i represent the response variable for the i^{th} observation.
- Then training dataset consist of $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$.

How do We Estimate f?

- Observations divides into training data that is used to teach our method how to estimate f.
- Let x_{ij} represent the value of j^{th} predictor for observation $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$.
- Let y_i represent the response variable for the i^{th} observation.
- Then training dataset consist of $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$.
- The goal is to apply a Statistical Learning method to find unknown function f.

How do We Estimate f?

- Observations divides into training data that is used to teach our method how to estimate f.
- Let x_{ij} represent the value of j^{th} predictor for observation $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$.
- Let y_i represent the response variable for the i^{th} observation.
- Then training dataset consist of $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$.
- The goal is to apply a Statistical Learning method to find unknown function f.
- We want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y). There are two methods : $\boxed{\text{parametric}}$ and $\boxed{\text{non-parametric}}$.

Parametric Methods

- Parametric methods involve a two-step model-based approach

1.

We make assumption about the functional form. E.g, a linear model that f is linear in X

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

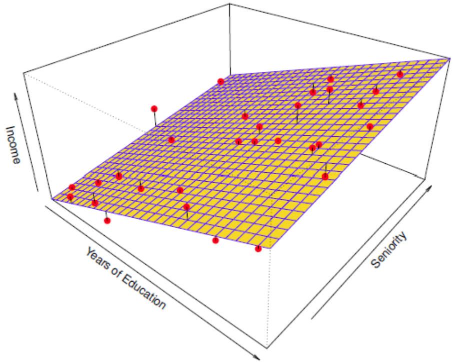
2.

A procedure that uses training data to fit or train the model. The most common approach $\rightarrow$ $\boxed{\text{Least Equare Estimate (LSE)}}$

Linear Model Example

- A linear model fit by **least squares** to the Income data.

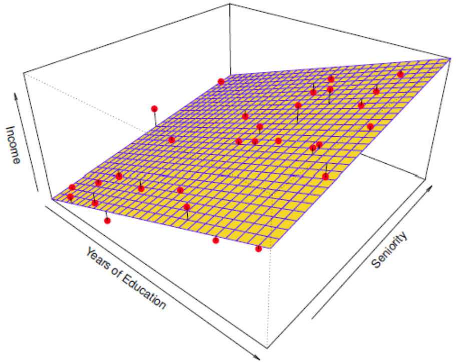
$$\textit{income} \approx \beta_0 + \beta_1 \times \textit{education} + \beta_2 \times \textit{seniority}$$



Linear Model Example

- A linear model fit by **least squares** to the Income data.
- Observations are in red and the yellow plane indicates the least squares fit to the data.

$$income \approx \beta_0 + \beta_1 \times education + \beta_2 \times seniority$$



Parametric Modeling Disadvantages

- A potential disadvantage of parametric approach: the model we choose will usually not match the true unknown form of f.

Parametric Modeling Disadvantages

- A potential disadvantage of parametric approach: the model we choose will usually not match the true unknown form of f.
- To address the performance issue, we choose flexible models $\rightarrow$ which means estimating greater number of parameters.

Parametric Modeling Disadvantages

- A potential disadvantage of parametric approach: the model we choose will usually not match the true unknown form of f.
- To address the performance issue, we choose flexible models $\rightarrow$ which means estimating greater number of parameters.
- The more complex models can lead to a phenomenon known as overfitting the data, which means they follow the errors, or noise, too closely.

Non-Parametric Methods

- Non-Parametric methods do not make explicit assumption about the functional form of f.

Non-Parametric Methods

- Non-Parametric methods do not make explicit assumption about the functional form of f.
- Instead seek an estimate of f that gets close to the data points as possible.

Non-Parametric Methods

- Non-Parametric methods do not make explicit assumption about the functional form of f.
- Instead seek an estimate of f that gets close to the data points as possible.
- Non-Parametric Advantage: No need to assumption of particular functional form for f.

Non-Parametric Methods

- Non-Parametric methods do not make explicit assumption about the functional form of f.
- Instead seek an estimate of f that gets close to the data points as possible.
- Non-Parametric Advantage: No need to assumption of particular functional form for f.
- Non-Parametric Disadvantage: Needs far more number of observations compared to parametric approach.

Non-Parametric Methods

- Non-Parametric methods do not make explicit assumption about the functional form of f.
- Instead seek an estimate of f that gets close to the data points as possible.
- Non-Parametric Advantage: No need to assumption of particular functional form for f.
- Non-Parametric Disadvantage: Needs far more number of observations compared to parametric approach.
- A popular example : Decision Tree is a **non-parametric** supervised machine learning algorithm used for both classification and regression problems.

Non-Parametric Methods

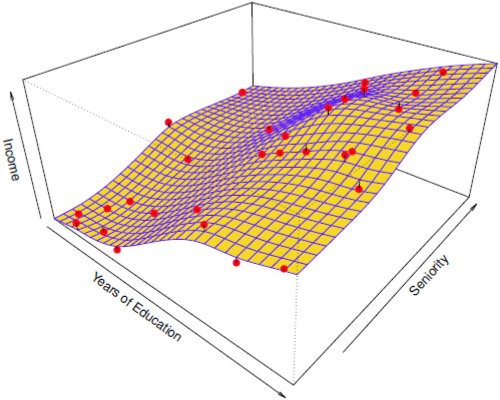
- Non-Parametric methods do not make explicit assumption about the functional form of f.
- Instead seek an estimate of f that gets close to the data points as possible.
- Non-Parametric Advantage: No need to assumption of particular functional form for f.
- Non-Parametric Disadvantage: Needs far more number of observations compared to parametric approach.
- A popular example : Decision Tree is a **non-parametric** supervised machine learning algorithm used for both classification and regression problems.
- In the Parametric approach it is possible that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well.

Non-Parametric Methods

- Non-Parametric methods do not make explicit assumption about the functional form of f.
- Instead seek an estimate of f that gets close to the data points as possible.
- Non-Parametric Advantage: No need to assumption of particular functional form for f.
- Non-Parametric Disadvantage: Needs far more number of observations compared to parametric approach.
- A popular example : Decision Tree is a **non-parametric** supervised machine learning algorithm used for both classification and regression problems.
- In the Parametric approach it is possible that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well.
- Non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made.

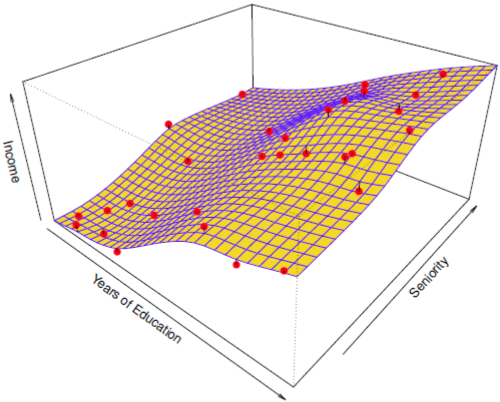
Non-Parametric- Example

- A thin-plate spline is used to estimate f.



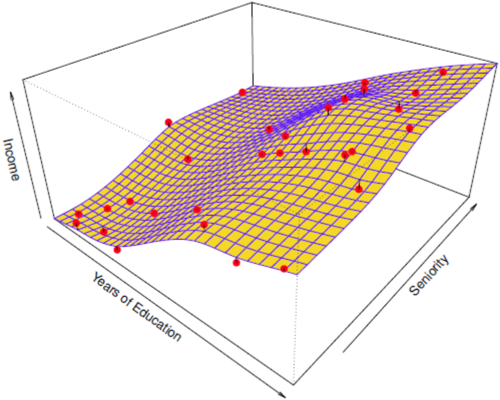
Non-Parametric- Example

- A thin-plate spline is used to estimate f.
- No assumption of f. Attempts to minimize MSE.



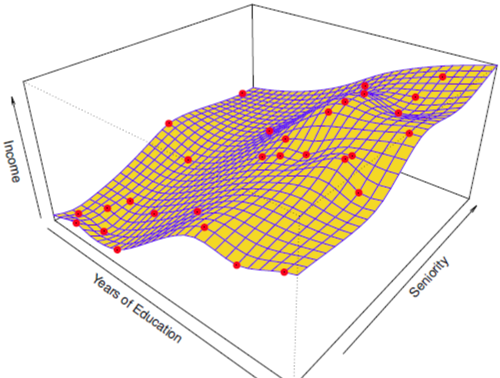
Non-Parametric- Example

- A thin-plate spline is used to estimate f.
- No assumption of f. Attempts to minimize MSE.
- Produce remarkably accurate estimate of f.



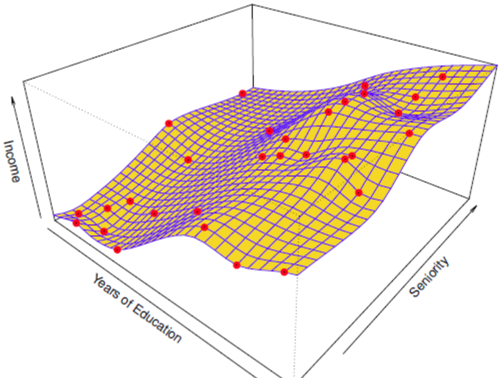
Non-Parametric- Example Rough fit

- A lower level of smoothness allowing rougher fit.



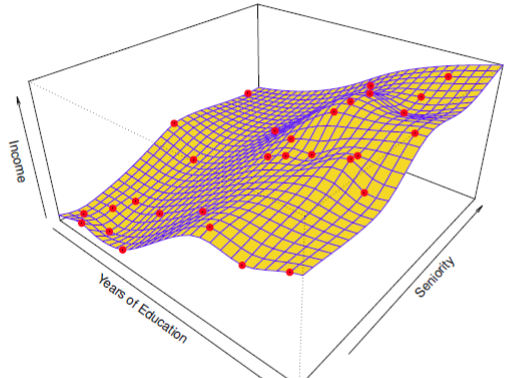
Non-Parametric- Example Rough fit

- A lower level of smoothness allowing rougher fit.
- Far more variable (flexible & less restrictive) but subject to overfitting.



Non-Parametric- Example Rough fit

- A lower level of smoothness allowing rougher fit.
- Far more variable (flexible & less restrictive) but subject to overfitting.
- The model fits the training data perfectly but poor performance on the test data.



Prediction Accuracy & Model Interpretability

- Some models are less restrictive or more restrictive.
- E.g., Linear regression is a relatively inflexible approach.

- If we are interested in prediction only and interpretability is not of interest, then we may expect that more flexible model is the best. Surprisingly, this is not always the case!.

Prediction Accuracy & Model Interpretability

- Some models are less restrictive or more restrictive.
- E.g., Linear regression is a relatively inflexible approach.

- If we are interested in prediction only and interpretability is not of interest, then we may expect that more flexible model is the best. $\boxed{\text{Surprisingly, this is } \underline{\text{not}} \text{ always the case!}}$.

Prediction Accuracy & Model Interpretability

- Some models are less restrictive or more restrictive.
- E.g., Linear regression is a relatively inflexible approach.
- E.g., Thin plate spline are considerably more flexible.

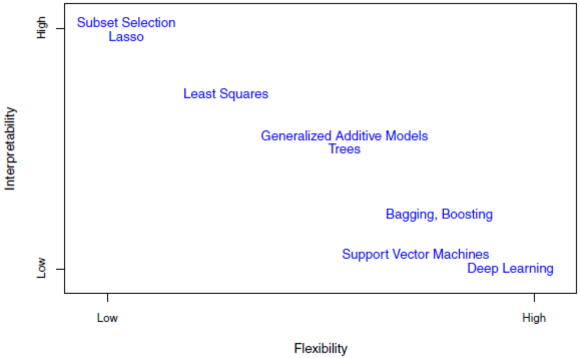
Question:why would we ever choose to use a more restrictive method instead of a very flexible approach?

- If we are mainly interested in inference, then restrictive models are much more interpretable.
- It is easier to understand the relationship between Y and $X_1, \ldots, X_p$ in a linear regression compared to a deep learning model.
- E.g., If we seek to predict the stock price, then interpretability is not a concern and we need an accurate prediction.

- If we are interested in prediction only and interpretability is not of interest, then we may expect that more flexible model is the best. Surprisingly, this is not always the case!.

Prediction Accuracy & Model Interpretability

- We will often obtain more accurate predictions using a less flexible model. (Highly flexible model suffers from potential overfitting)



Supervised Learning

Supervised Learning

- For each observation of the predictor measurement(s) $x_i, i = 1, ..., N$ there is an associated response measurement y_i.
- Goal: Fit a model that relates the response to the predictors with the aim of accurately predicting the response for the future observations ($\boxed{\text{prediction}}$) or better understanding the relationship between the response and the predictors ($\boxed{\text{inference)}}$).
- Classical statistical learning : E.g., linear regression & logistic regression.
- Modern statistical learning : E.g., support vector machines with non-linear kernels or Neural networks (deep learning).

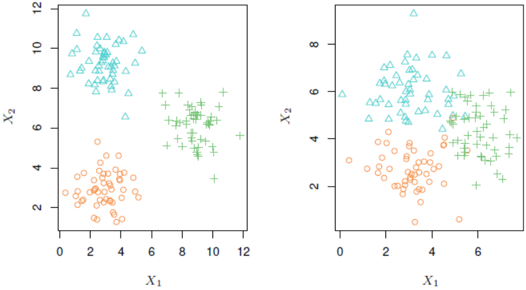
Unsupervised Learning

Unsupervised Learning

- Unsupervised learning describes more challenging situation in which every observed observation $i = 1, ..., n$ is a vector of measurements x_i but no associated response y_i.
- It is not possible to fit a linear regression model since there is no response variable to predict. In this case, we are in some sense working blind.
- The statistical learning tool in this case, is cluster analysis, or clustering.
- The goal of cluster analysis cluster is to ascertain, on the basis of $x_1, ..., x_n$, whether the observations fall into analysis relatively distinct groups.
- E.g., zip code, family income & shopping habit. Big spenders versus low spenders.

Unsupervised Learning

- 150 observations with two variables X_1 & X_2
- Each observation corresponds to one of the three distinct groups.
- The left-hand side is well-separated. The right-hand-side is a more challenging due to some overlap between the groups.
- If p variables, then $\frac{p(p-1)}{2}$ distinct scatter plots$\longrightarrow$ Visual inspection is not viable.



Measuring the Quality of Fit

- In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data.
- In the regression, the most common-used measure is mean squared error(MSE).

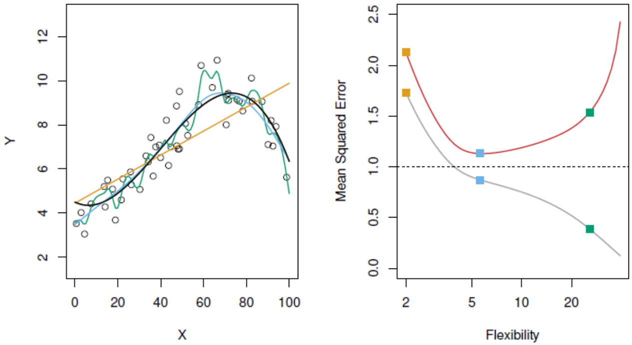
$$MSE(train) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- Suppose we fit out machine learning model on our training observations $(x_1, y_1), ..., (x_n, y_n)$ and we obtain $\hat{f}$. If $\hat{f}(x_i \approx y_i$ for $i = 1, ..., n$ then MSE is small.
- But we are interested to know if $\hat{f}(x_0) \overset{?}{\approx} y_0$ where (x_0, y_0) is a previously unseen test observation.

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

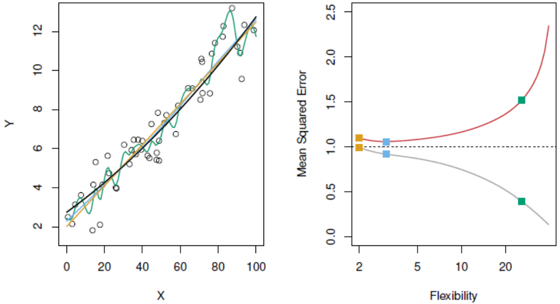
Mean squared Error and overfitting

- Left : Linear regression line, two smoothing spline fits.
- Right : Training MSE and Test MSE
- The green model (flexible) matches the data very well, but it fit the true f poorly. $\boxed{\text{Small Train MSE but high Test MSE}}$.
- Horizontal dashed line indicates $Var(\epsilon)$ irreducible error.



Mean squared Error and overfitting

- When a model yields a small <u>training</u> MSE but a large <u>test</u> MSE we are said to be overfitting the data.
- In this example linear regression provides a very good fit to the data.
- Estimation of the test MSE is difficult since usually no test data are available.
- One method to estimate test MSE is cross validation.



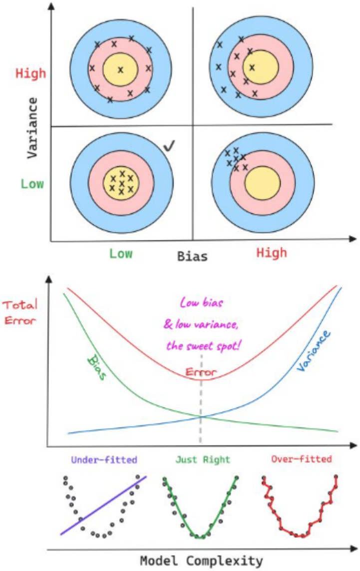
The Bias-Variance Trade-Off

- The U-shape observed in the test MSE is due to two competing properties of statistical learning methods.
- Test MSE for a given x_0 can always to be decomposed into:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias(\hat{f}(x_0))\right]^2 + Var(\epsilon)$$

- We need to select a statistical learning method that simultaneously achieves low variance and low bias.
- Variance is non-negative quantity. Squared of bias is non-negative. Hence the test MSE can never lie below $Var(\epsilon)$

The Bias-Variance Trade-Off



The Bias-Variance Trade-Off

Variance

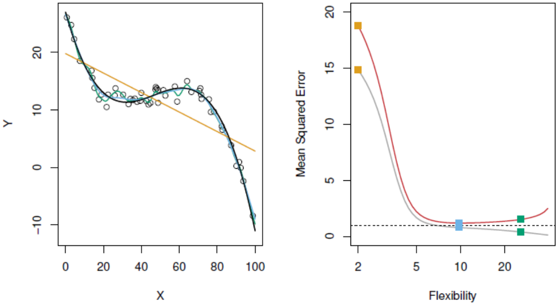
- Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- If a model has a high variance, small changes in training data $\rightarrow$ large changes in $\hat{f}$. Generally $\uparrow$ flexible model,$\uparrow$ variance.

Bias

- Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- It is unlikely that any real-life problem truly has such a simple linear relationship between Y and $X_1, ..., X_p$ so performing linear regression will result large bias in the estimate of f.
- Generally $\uparrow$ flexible model,$\downarrow$ bias.

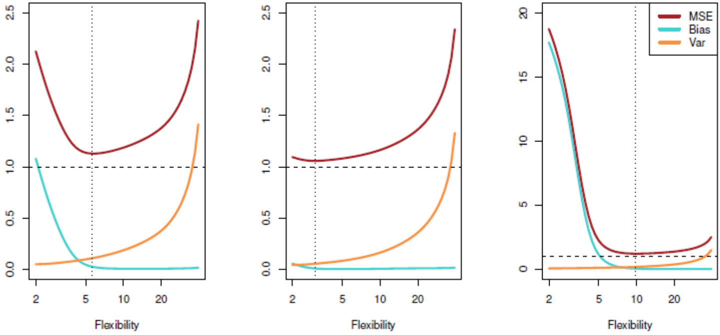
The Bias-Variance Trade-Off

- As the flexibility of a class of methods increases, the bias tends to initially decrease faster than the variance increases. Consequently, the expected test MSE declines.
- However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance. When this happens the test MSE increases.



The Bias-Variance Trade-Off

- Blue line → squared bias. Orange line → variance. Horizontal dashed line → $Var(\epsilon)$. Vertical dashed line → optimum flexibility level corresponding to the smallest test MSE.



The Bias-Variance Trade-Off

Bias-variance Trade off

- Relationship between bias, variance and test MSE is called bias-variance trade off.
- It is easy to obtain a method with extremely low bias but high variance (by drawing a curve that passes through every single training observation)
- It is easy to obtain a method with very low variance but high bias (by fitting a horizontal line to the data).
- The challenge lies in finding a method for which both the variance and the squared bias are low.

The Classification Setting

- Many of the concepts that we have encountered, such as the bias-variance trade-off, transfer over to the classification setting with only some modifications since y_i is no longer quantitative.
- Suppose we seek to estimate f on the basis of training observations $(x_1, y_1), \ldots, (x_n, y_n)$ where $y_1, \ldots, y_n$ are qualitative.

Training Error Rate

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

- $\hat{y}_i$ is the predicted class label for the i^{th} observation using $\hat{f}$.
- $I(y_i \neq \hat{y}_i) \Rightarrow 1$ if $y_i \neq \hat{y}_i$
- $I(y_i \neq \hat{y}_i) \Rightarrow 0$ if $y_i = \hat{y}_i$

The Classification Setting

Test Error Rate

- The test error rate associated with a set of test observations of the form (x_0, y_0) is given by

$$Ave(I(y_0 \neq \hat{y}_0))$$

- $\hat{y}_0$ is the predicted class label that results from applying the classifier to the test observations with predictor x_0.
- A good classifieris one for which the test error is small.