Predictive Analytics (ISE529)

Predictive Modeling

Dr. Tao Ma
ma.tao@usc.edu

Tue/Thu, May 22 - July 1, 2025, Summer

USC Viterbi
School of Engineering
*Daniel J. Epstein
Department of Industrial
and Systems Engineering*

USC University of
Southern California

INTRODUCTION TO PREDICTIVE MODELS

Statistical vs Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much **overlap** — both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on **large scale** applications and **prediction accuracy**.
  - Statistical learning emphasizes models and their **interpretability**, **precision** and **inference**.
- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.
- Machine learning has the upper hand in Marketing!

A Brief History

- At the beginning of the 19th century, the method of **least squares** was developed, implementing the earliest form of **linear regression** for predicting **quantitative** values.
- **Linear discriminant analysis** was proposed in 1936 to predict **qualitative** values (categorical). In the 1940s, **logistic regression**
- In the early 1970s, the **generalized linear model** was developed to describe an entire class of statistical learning methods that include both **linear** and **logistic** regression
- In the mid 1980s, **classification** and **regression trees** were developed, followed shortly by **generalized additive models**.
- **Neural networks** gained popularity in the 1980s, and **support vector machines** arose in the 1990s.

Since that time, statistical learning has emerged as a new subfield in statistics, focused on supervised and unsupervised modeling and prediction.

Examples of Prediction Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack based on demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Traffic volume prediction over highway network and urban road network
- Weather forecast, earthquake forecast, etc., … … …

Supervised vs Unsupervised Learning

Statistical/machine learning refers to a vast set of tools for understanding data. These tools can be classified as **supervised** or **unsupervised**.

- Supervised learning involves building a statistical model for predicting an output based on inputs. For each observation of the predictor measurement(s) x_i, $i = 1, . . . , n$ there is an **associated response** measurement y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (**prediction**) or better understanding the relationship between the response and the predictors (**inference**), e.g., linear regression and logistic regression

- Unsupervised learning describes the situation where for every observation $i = 1, . . . , n$, we observe a vector of measurements x_i without associated response y_i. There are inputs **but no output**; nevertheless, we can learn relationships and structure from such data, e.g. cluster analysis.

The Supervised Learning

Specifically,

- Outcome measurement *Y* (also called dependent variable, response, target).
- Vector of *p* predictor measurements *X* (also called inputs, regressors, covariates, features, independent variables).
- In the **regression** problem, *Y* is quantitative (e.g., price, blood pressure).
- In the **classification** problem, *Y* takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample). *Y* is also called categorical variable.
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

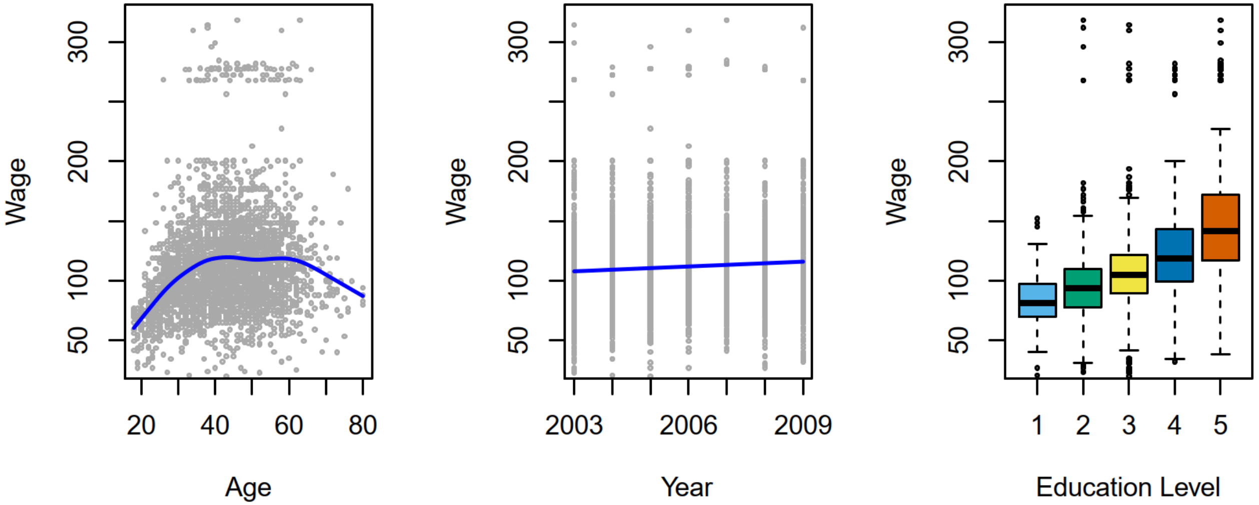
Objectives and example

Based on the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Example:

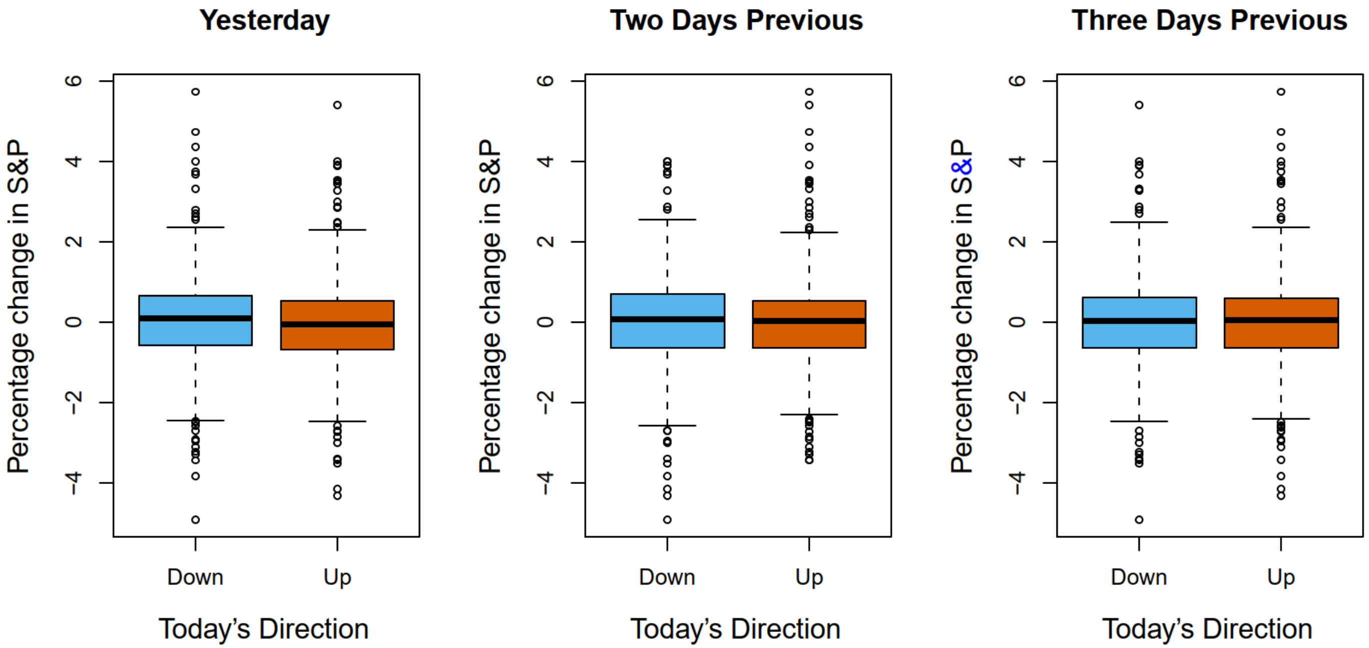
we examine a number of **factors** that relate to wages for a group of men from the Atlantic region of the United States. We wish to understand the association between an employee’s **age** and **education**, as well as the **calendar** year, on his wage.



Example

In certain cases, we may wish to predict a non-numerical value—that is, a **categorical or qualitative** output.

- A stock market data set that contains the daily movements in the Standard & Poor's 500 (S&P) stock index over a 5-year period between 2001 and 2005.
- The goal is to predict whether the index will ***increase or decrease*** on a given day, using the past 5 days' percentage changes in the index. This is known as a **classification problem**.



one for the 648 days for which the market increased on the subsequent day, and one for the 602 days for which the market decreased. The two plots look almost identical, suggesting that there is no simple strategy for using yesterday's movement in the S&P to predict today's returns.

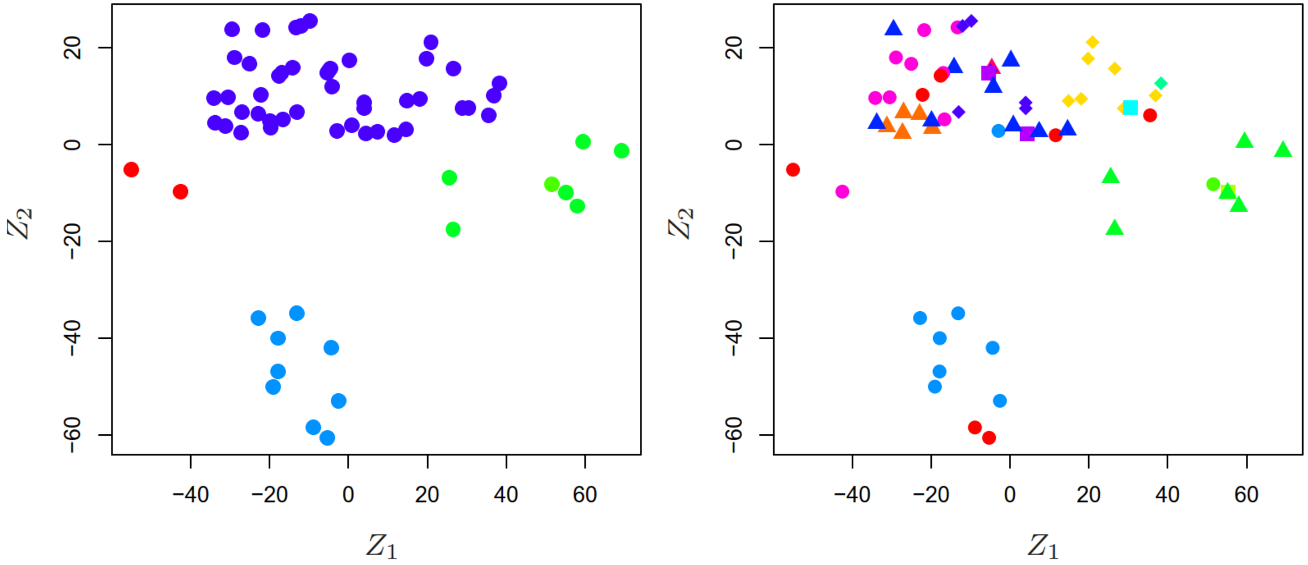
Unsupervised Learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is fuzzy:
  - find groups of samples that behave similarly
  - find features that behave similarly
  - find linear combinations of features with the most variation
- Difficult to know how well you are doing.
- Different from supervised learning but can be useful as a pre-processing step for supervised learning.

Example

In the situations where we only observe input variables, with no corresponding output. We may wish to understand which types of individuals are similar to each other by grouping them according to their observed characteristics. This is known as a **clustering problem.**

- A gene expression data set consists of 6,830 gene expression measurements for each of 64 cancer cell lines.
- We want to determine whether there are groups, or clusters, among the cell lines based on their gene expression measurements.



Representation of the gene expression data set in a two-dimensional space, Z1 and Z2. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One must understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately **assess the performance of a method**, to know how well or how badly it is working
- **Simpler methods often perform as well as fancier ones**!
- This is an exciting research area, having important applications in science, industry and finance.
- Statistical/Machine learning is a fundamental ingredient in the training of a modern data scientist.

STATISTICAL LEARNING

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. Our client cannot directly increase sales of the product, but adjust advertising budgets, thereby indirectly increasing sales. The data set looks like:

#	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
…	…	…	…	…
199	283.6	42	66.2	25.5
200	232.1	8.6	8.7	13.4

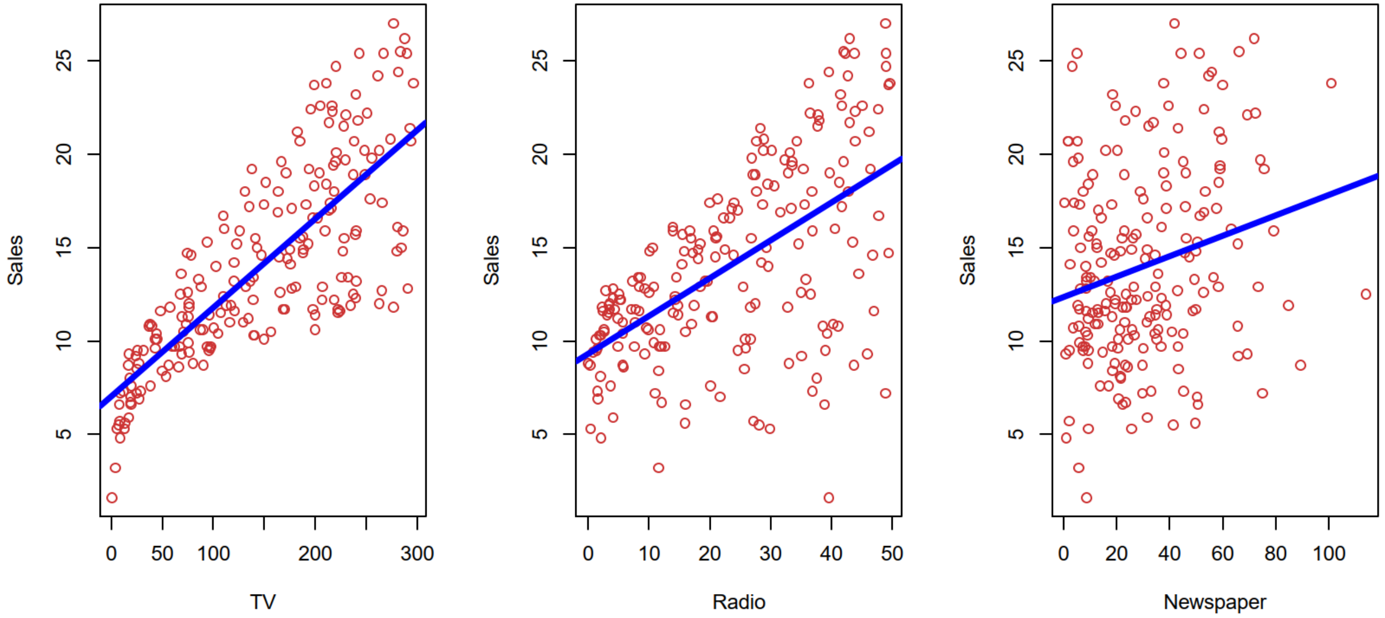
Our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

- Input variables = Advertising budgets
- output variable = Sales

$$\text{Sales} \approx f\,(\text{TV, Radio, Newspaper})$$

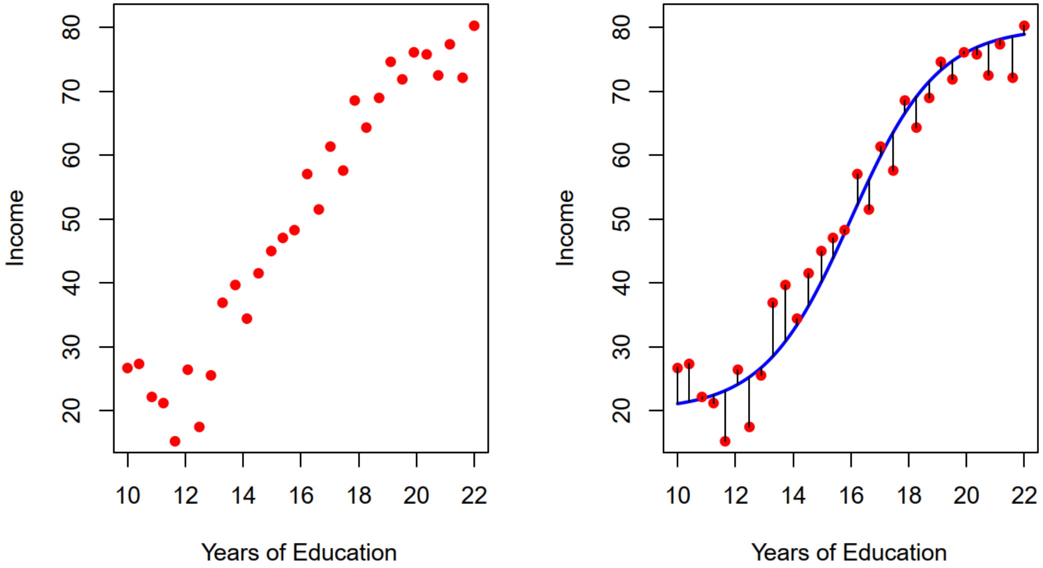
What is Statistical Learning?

- We use ***X*** with subscript to denote the input variables. x_1 = TV budget, x_2 = the radio budget, x_3 = the newspaper budget. The inputs also refer to as predictors, independent variables, features, predictor, or sometimes just variables.
- Sales is a response or target that we wish to predict, generically refer to as the response or dependent variable denoted by Y. (Y = sales)



What is Statistical Learning?

Suppose to predict income using years of education. A plot of shows income versus years of education for 30 individuals in the Income data set.



$$\text{income} \approx f\,(\text{years of education}\ )$$

The function *f* that connects the input variable to the output variable is in general unknown. We must estimate *f* based on the observed points.

Mathematical Notation

More generally, suppose that we observe a quantitative response *Y* and *p* different predictors, $x_1, x_2, \ldots, x_p$. We assume that there is some relationship between *Y* and $\boldsymbol{X}^T = (x_1, x_2, \ldots, x_p)$,

Now we write our model as

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

Here *f* is a fixed but **unknown** function of $x_1, x_2, \ldots, x_p$, and ϵ is a random error term, which is independent of $\boldsymbol{X}$ and has mean zero and variance σ^2.

Take expectation on both sides, we get:

$$E\left(Y \mid X = x\right) = f\left(X\right)$$

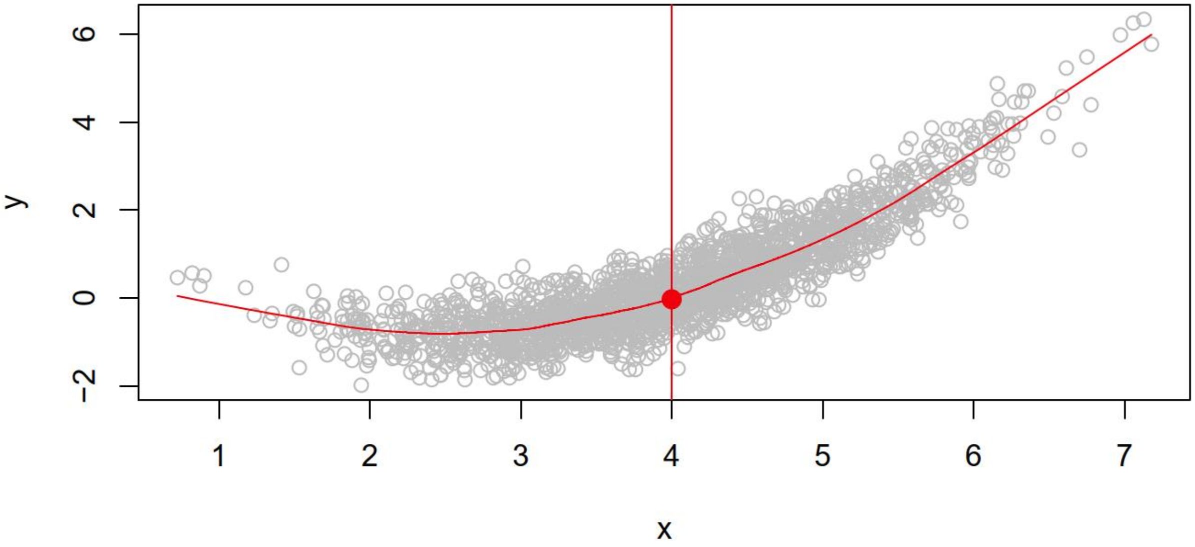
Mathematical Notation

What is a good value for $f(X)$ at any selected value of X, say $X = 4$? There can be many Y values at $X = 4$. According to the model, we get:

$$f(4) = E(Y \mid X = 4)$$

$E(Y|X = 4)$ is the expected value (average) of Y given $X = 4$.

This ideal $f(X) = E(Y \mid X = x)$ is called the **regression function**.



Why Estimate *f* ?

Two main reasons that we may wish to estimate *f*:

Prediction

A good f yields accurate predictions of Y at new points $\boldsymbol{X}^T = (x_1, x_2, \ldots, x_p)$.

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

where $\hat{f}$ represents our estimate for f *and* $\hat{Y}$ represents the resulting prediction for Y .

Inference

We can understand the association between Y and $x_1, x_2, \ldots, x_p$, which components are important in explaining Y , and which are irrelevant.

Depending on the complexity of *f*, we may be able to understand how each component x_j of $\boldsymbol{X}$ affects Y and to what extent.

Why Estimate *f* ?

Prediction

For instance, a company is interested in conducting a direct-marketing campaign. The goal is to identify individuals who are likely to respond positively to a mailing, based on observations of demographic variables measured on each individual.

- The demographic variables serve as predictors, and
- The response to the marketing campaign (either positive or negative) serves as the outcome.

The company is not interested in obtaining a deep understanding of the relationships between each individual predictor and the response; instead, the company simply wants to accurately predict the response using the predictors.

Why Estimate *f* ?

Inference

1. Consider the advertising data set, answer the following questions:

- Which media are associated with sales?
- Which media generate the biggest boost in sales?
- How large of an increase in sales is associated with a given increase in TV advertising?

2. Modeling the brand of a product that a customer might purchase based on variables such as price, store location, discount levels, competition price, and so forth. In this situation one might really is interested in the association between each variable and the probability of purchase. Answer the following questions:

- To what extent is the product's price associated with sales?

Why Estimate *f* ?

Combination for prediction and inference

In a real estate setting, one may seek to relate **values of homes** to inputs such as crime rate, zoning, distance from a river, air quality, schools, income level of community, size of houses, and so forth.

Answer the following questions:

- How much extra will a house be worth if it has a view of the river? (inference)
- Is this house under- or over-valued given its characteristics? (prediction)

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating *f* may be appropriate to serve different purposes, e.g., linear models for interpretable inference; non-linear approaches for prediction.

How to estimate *f* ?

- In essence, statistical learning refers to a set of approaches for estimating *f*. We focus on the key theoretical concepts that arise in **estimating** *f*, as well as tools for **evaluating** the estimates obtained. *f* represents the relationship between ***X*** and *Y*. *f* must be determined by observed data.
- A set of *n* different data points is called the **training data set.**
- Let x_{ij} data represent the value of the j^{th} predictor, or input, for observation *i*, where $i = 1, 2, . . . , n$ and $j = 1, 2, . . . , p$.
- Correspondingly, let y_i represent the response variable for the i^{th} observation.
- Then our training data consist of $\{(x_1, y_1), (x_2, y_2), . . . , (x_n, y_n)\}$ where $x_i^T= (x_{i1}, x_{i2}, . . . , x_{ip})$.
- Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function *f*, i.e., to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).

How to estimate *f* ?

Broadly speaking, most statistical learning methods for this task can be characterized as either **parametric** or **non-parametric**.

Parametric Methods: a two-step approach.

1. Make an assumption about the functional form, or shape of *f*.

   e.g., *f* is linear in *X*:

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

Once we have assumed that *f* is linear, the problem of estimating *f* is boiled down to estimating the (*p*+1) coefficients β_0, β_1, . . . , β_p.

How to estimate *f* ?

2. After a model form has been selected, we need a procedure that uses the training data to fit or train the model, *i.e.*, find values of these parameters such that

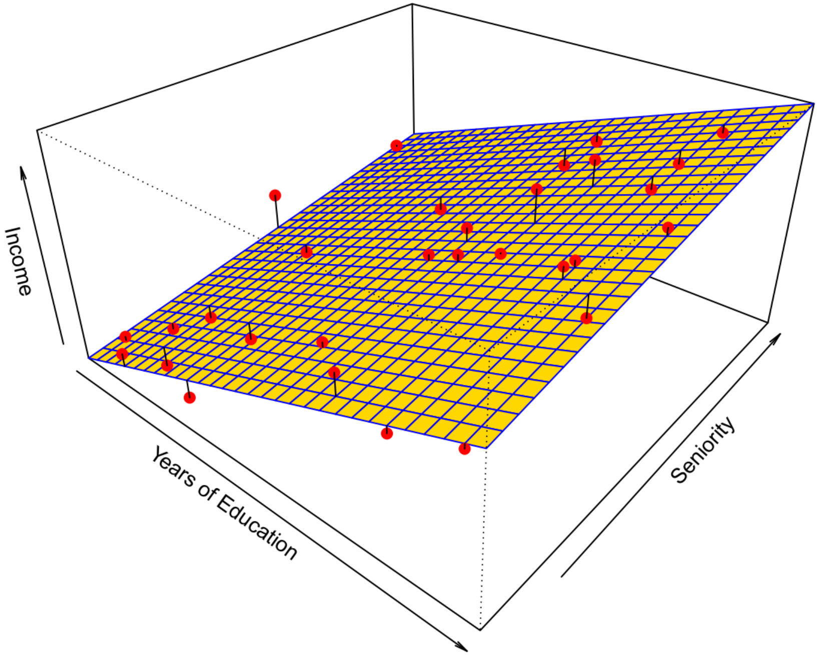
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

The most common approach to fitting the model form above is referred to as (ordinary) **least squares**.

The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of *f*.

How to estimate *f* ?

A linear model fit by least squares to the Income data . The observations are shown in red, and the yellow plane indicates the **least squares** fit to the data.



We have fit a linear model of the form

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

How to estimate *f* ?

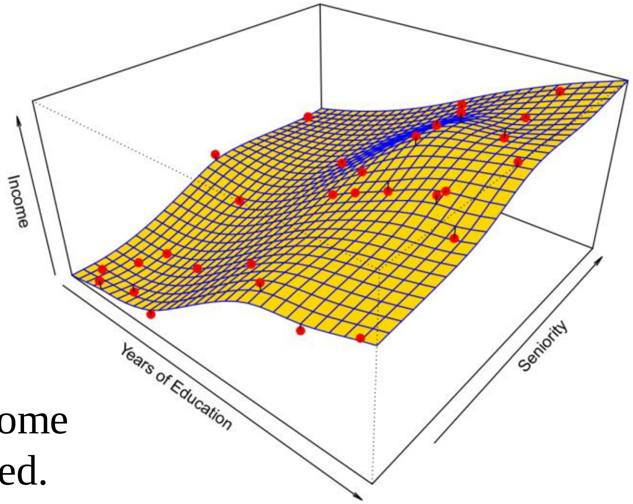
Non-parametric methods do not make explicit assumptions about the functional form of *f*. It instead attempts to produce an estimate for *f* that is as close as possible to the observed data.

By avoiding the assumption of a particular functional form for *f*, they have the potential to accurately fit a wider range of possible shapes for *f*.

Disadvantage:

Since they do not reduce the problem of estimating *f* to a small number of parameters, it **requires a very large data set** in order to obtain an accurate estimate for *f*.

It suffers from overfitting problem. It may fit the training data well but **will not yield accurate predictions** of the response on new observations that were not part of the original training data set.



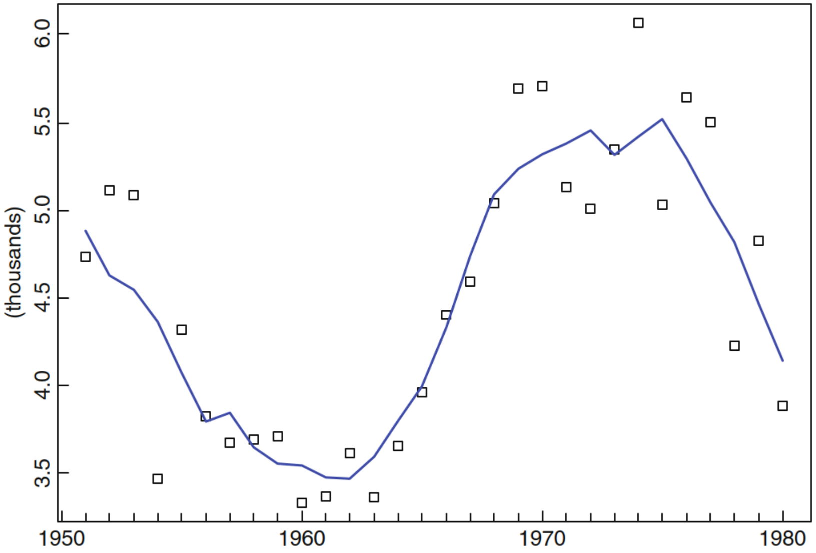
A smooth thin-plate ***spline*** fit to the Income data, the observations are displayed in red.

How to estimate f ?

Non-parametric methods: moving average filter

$$\hat{f}(x_i) = Ave\left(Y \mid X \in N(x_i)\right)$$

where $N(x_i)$ is some neighborhood of x_i . The parameter is bin width q.

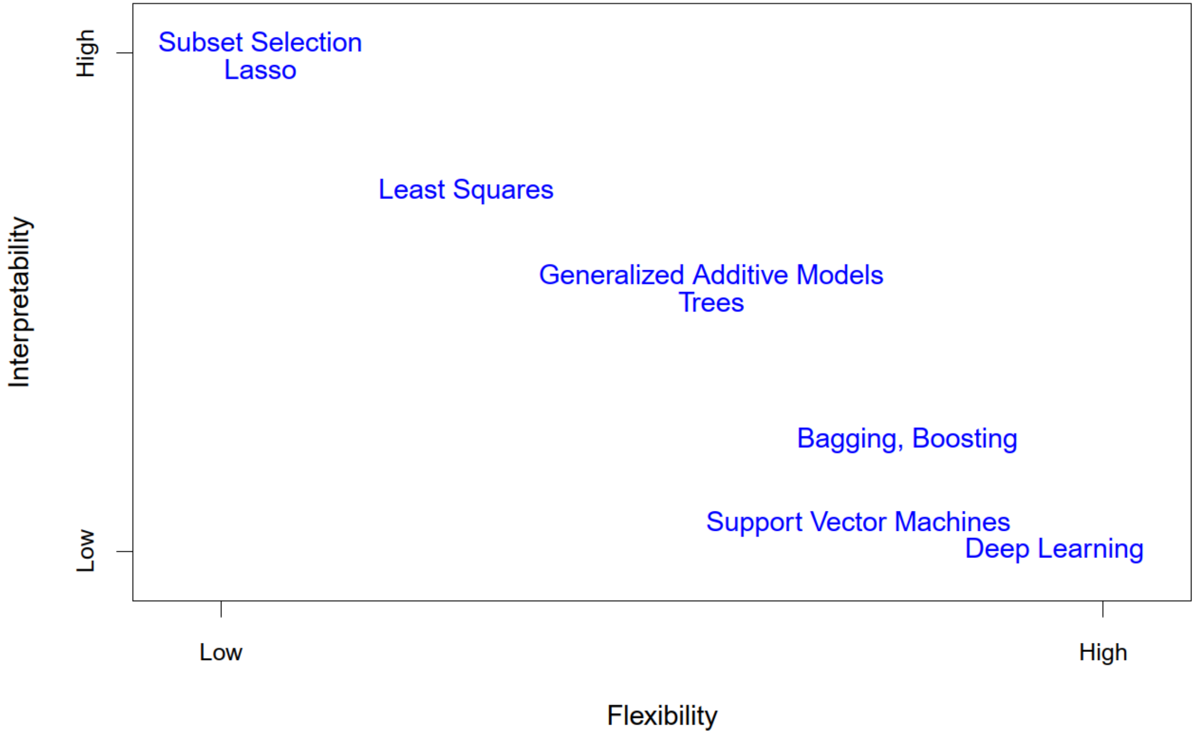


Example: we use moving average filter to estimate trend m_t over years.

$$\hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}, \quad q+1 \le t \le n-q.$$

Trade-Off

How do we choose a model form of f between inflexible and flexible approaches, such as linear regression versus smoothing spline ?



The figure above represents the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

The rule of thumb

- When **inference** is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods.

- In some settings, however, we are only interested in **prediction**, and the interpretability of the predictive model is simply not of interest. It will be best to use the most flexible model available.

- **Surprisingly, this is not always the case!** We will often obtain more accurate predictions using a less flexible method. This may seem counter-intuitive. This phenomenon is due to the potential for **overfitting** in highly flexible methods.

- **Parsimony** versus black-box. We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Regression vs Classification

- Variables can be characterized as either **quantitative** or **qualitative** (also known as **categorical**).
- Quantitative variables take on numerical values, e.g., income, the value of a house, stock price, etc.
- Qualitative variables take on values in one of *K* different classes, or categories, e.g., marital status (married or not), travel mode choice, etc.
- Regression quantitative response - regression problems

  e.g., least squares linear regression
- Regression qualitative response - classification problems

  e.g., logistic regression
- We select statistical learning methods on the basis of whether the response is quantitative or qualitative. However, whether the **predictors** are qualitative or quantitative is generally considered less important. Most of the statistical learning methods can be applied **regardless of the predictor variable type** as long as they are properly coded.

ASSESSING MODEL ACCURACY

How to choose the best *f* ?

“**All models are wrong, but some are useful.**”
a quote by British statistician George E. P. Box, published in 1976.

There is no free lunch in statistics: no one method dominates all others over all possible data sets. On a particular data set, one specific method may work best, but some other methods may work better on a similar but different data set. **Hence it is an important task to decide for any given set of data which method produces the best results**.

Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice. We need to find a way to measure the Goodness of Fit of a model.

Measuring the Goodness of Fit

To evaluate the performance of a statistical learning method on a given data set, we need to measure how well its predictions match the observed data.

In the regression setting, the most commonly-used measure is the mean squared error (**MSE**), given by

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

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the i^{th} observation, y_i is response value of the i^{th} observation. If the MSE is computed using the training data that was used to fit the model, it is called the **training MSE**.

Measuring the Goodness of Fit

Suppose that we fit our statistical learning method on our training observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and we obtain the estimate $\hat{f}$.

If $\hat{f}(x_i) \approx y_i,\ i = 1, 2, \ldots, n$, then the **training MSE** is small. However, we are not interested in whether the **training MSE** is small, instead, we want to know whether $\hat{f}(x_0) \approx y_0$, where (x_0, y_0) is a **test observation previously unseen** in training data set.

We want to choose the method that gives the lowest test MSE, as opposed to the lowest training MSE. if we had a large number of **test observations**, we could compute

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

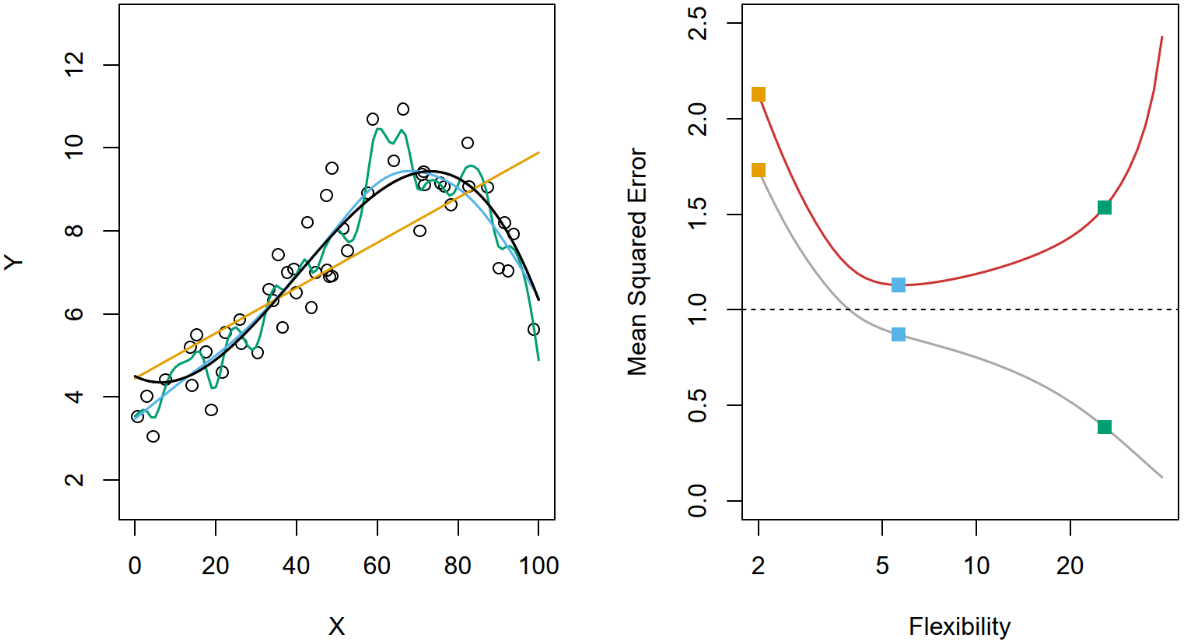
the **average squared prediction error** for these test observations (x_0, y_0). We'd like to select the learning method for which the test MSE is smallest.

Measuring the Goodness of Fit

Why not use the training MSE to select model?
There is no guarantee that the model with the lowest **training MSE** will also have the lowest **test MSE**. Roughly speaking, the problem is that many statistical methods specifically estimate coefficients so as to minimize the training set MSE. For these methods, the training set MSE can be quite small, but the test MSE is often much larger.

True *f* shown in black, three estimates of *f* are shown: the linear regression line (orange), and two smoothing spline fits (blue and green).

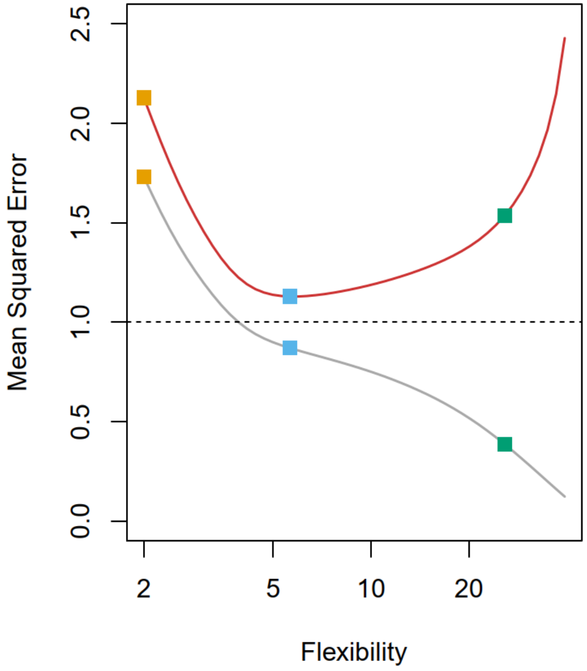
Training MSE (grey curve)
Test MSE (red curve)



The grey curve displays the average training MSE as a function of flexibility for a number of smoothing splines. The red curve denotes the test MSE. The horizontal dashed line indicates Var(ϵ), the lowest achievable test MSE.

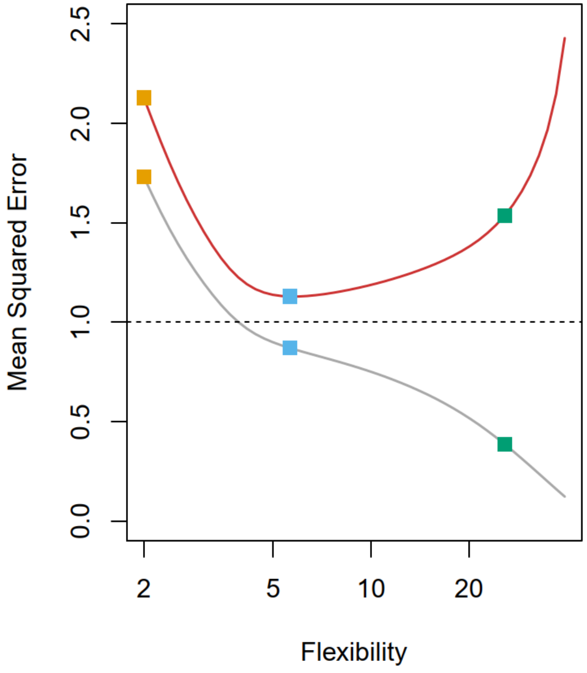
Overfitting Phenomenon

- As the flexibility of the statistical learning method increases, there is a monotone decrease in the training MSE and a **U-shape** in the test MSE. This is a fundamental property of statistical learning that holds regardless of the data set at hand and regardless of the statistical method being used.
- As model flexibility increases, the training MSE will decrease, **but the test MSE may not**. When a given method yields a small training MSE but a large test MSE, it is said to be **overfitting** the data.
- This is because our statistical learning procedure is working too hard to find patterns in the training data, and may be **picking up some** patterns that are just caused by random chance (**noise**) rather than by true properties of the unknown function *f*.

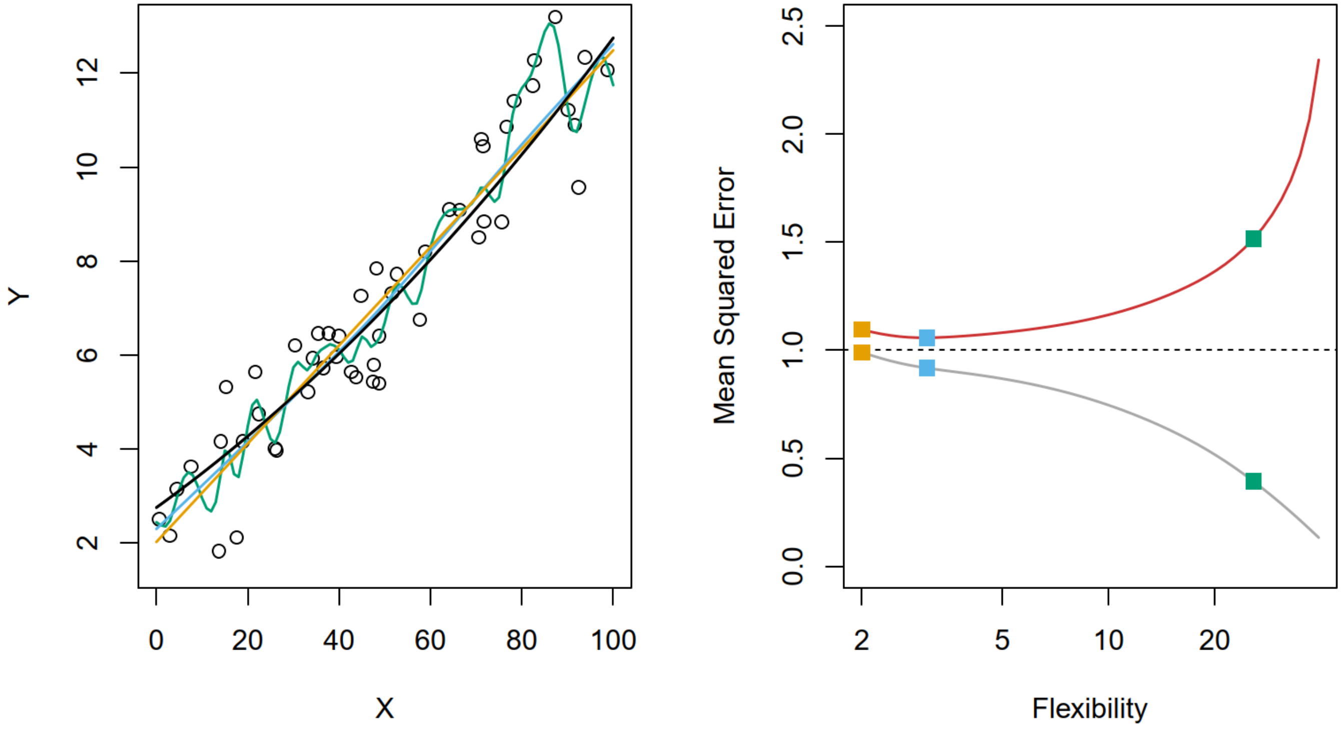


Overfitting Phenomenon

- As the flexibility of the statistical learning method increases, there is a monotone decrease in the training MSE and a **U-shape** in the test MSE. This is a fundamental property of statistical learning that holds regardless of the data set at hand and regardless of the statistical method being used.
- As model flexibility increases, the training MSE will decrease, **but the test MSE may not**. When a given method yields a small training MSE but a large test MSE, it is said to be **overfitting** the data.
- This is because our statistical learning procedure is working too hard to find patterns in the training data, and may be **picking up some** patterns that are just caused by random chance (**noise**) rather than by true properties of the unknown function *f*.

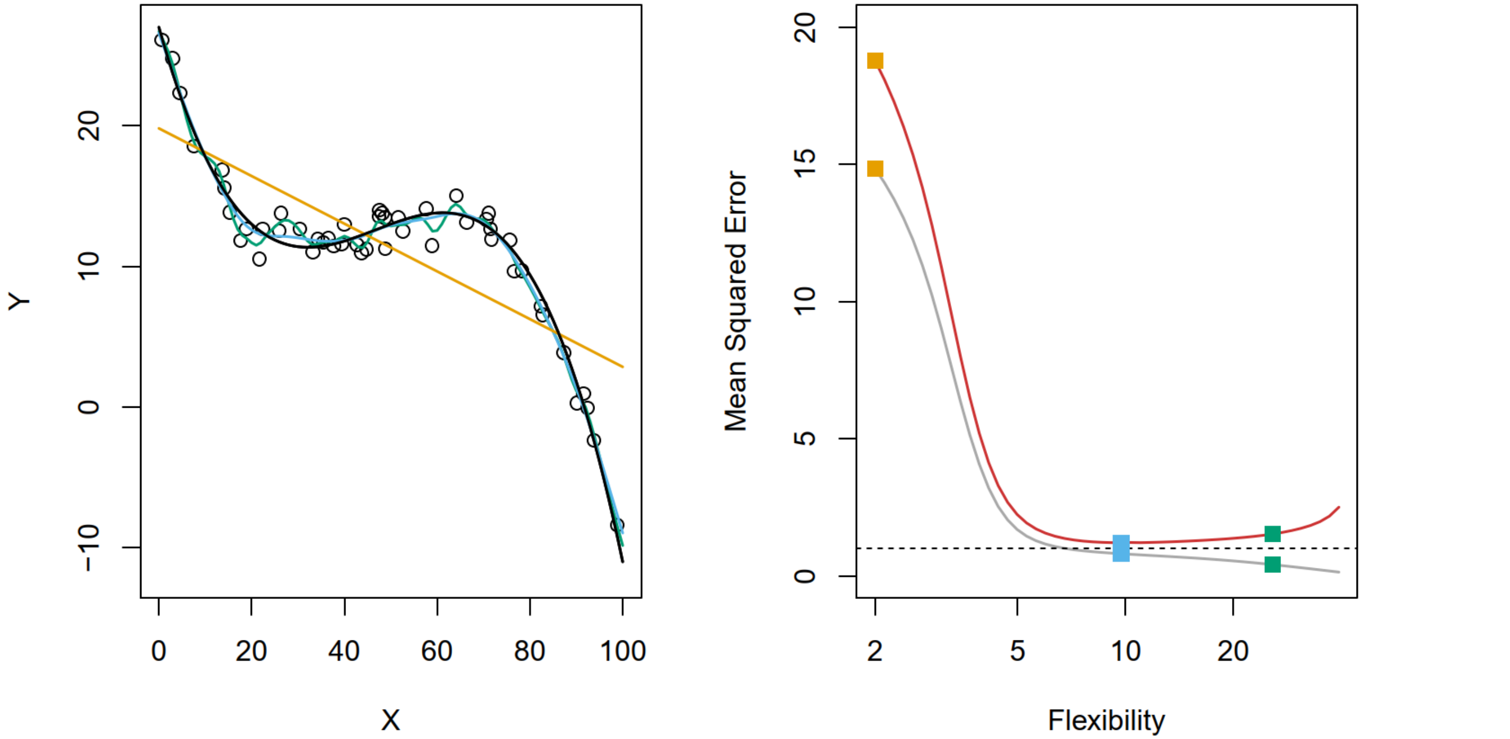


Overfitting Phenomenon



Another example: using a different true f that is much closer to linear. observe that the training MSE decreases monotonically as the model flexibility increases, and that there is a U-shape in the test MSE. However, because the truth is close to linear, the test MSE only decreases slightly before increasing again, so that the orange least squares fit is substantially better than the highly flexible green curve.

Overfitting Phenomenon



Another example: using a different true *f* that is highly non-linear. The training and test MSE curves still exhibit the same general patterns, but now there is a rapid decrease in both curves before the **test MSE** starts to increase slowly.

Unbiased estimators

An estimator should be "close" in some sense to the true value of the unknown parameter.

Definition

The point estimator $\hat{\Theta}$ is an unbiased estimator for the parameter θ if

$$E\left(\hat{\Theta}\right) = \theta$$

If the estimator is not unbiased, then the difference

$$E\left(\hat{\Theta}\right) - \theta$$

is called the **bias** of the estimator $\hat{\Theta}$.

This is equivalent to saying that the mean of the sampling distribution of $\hat{\Theta}$ is equal to θ. When an estimator is unbiased, the bias is zero; that is, $E\left(\hat{\Theta}\right) - \theta = 0$.

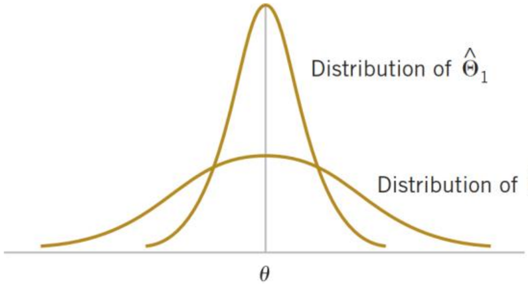
Unbiased estimators

Sample Mean and Variance are Unbiased.

- Suppose that X is a random variable with mean μ and variance σ^2. Let X_1 , X_2 ,…, X_n be a random sample of size n from the population represented by X.
- Show that the sample mean $\overline{X}$ and sample variance S^2 are unbiased estimators of μ and σ^2, respectively.

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

If we consider all unbiased estimators of θ, the one with the smallest variance is called the **minimum variance unbiased estimator.**



For example, we wish to estimate the mean of a population, two possible unbiased estimators for μ: the sample mean $\overline{X}$ and a single observation from the sample, say, X_i, which one should we use?

The Bias-Variance Trade-Off

The **U-shape** test MSE curves turns out to be the result of **two competing properties** of statistical learning methods. The expected test MSE, for a given value x0, can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error term ϵ. That is,

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

where

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

$$Bias\left(\hat{f}(x_0)\right) = f(x_0) - E\left[\hat{f}(x_0)\right]$$

$$Var(\varepsilon) = E\left(y_0 - f(x_0)\right)^2$$

Here the notation $E\left(y_0 - \hat{f}(x_0)\right)^2$ defines the **expected test MSE** at x_0, refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets and tested each at x_0.

Variance and Bias

- Variance of a statistical learning method

Using different training data sets will result in a different $\hat{f}$. Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. If a method has high variance, then small changes in the training data can result in large changes in $\hat{f}$. In general, more flexible statistical methods have higher variance.

- Bias of a statistical learning method

Bias refers to the error between the model and the underline true *f* that is introduced by approximating method chosen. Generally, more flexible methods result in less bias.

- Irreducible error of a statistical learning method

Variability associated with ϵ also affects the accuracy of our predictions. This is known as the irreducible error, because no matter how well we estimate f, we cannot reduce the error introduced by ϵ. The quantity ϵ may contain unmeasured variables (information) that are useful in predicting *Y*.

Reducible and irreducible error

Consider a given estimate $\hat{f}$ and a set of predictors $\boldsymbol{X_0}$, which yields the prediction $\hat{y}_0 = \hat{f}\left(X_0\right)$. Then, it is easy to show that

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

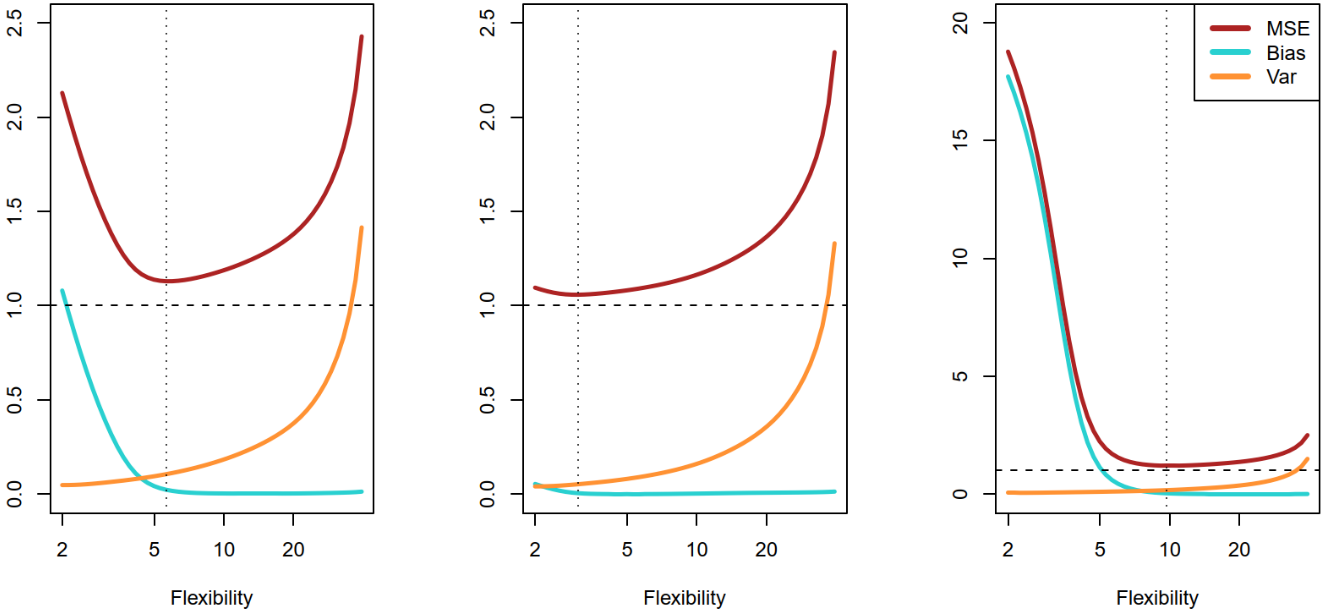
$$= \underbrace{E\left[f\left(x_0\right) - \hat{f}\left(x_0\right)\right]^2}_{reducible} + \underbrace{Var\left(\varepsilon\right)}_{irreducible}$$

Can you show it is equal to the following?

$$= \underbrace{Var\left(\hat{f}\left(x_0\right)\right) + \left[Bias\left(\hat{f}\left(x_0\right)\right)\right]^2}_{reducible} + \underbrace{Var\left(\varepsilon\right)}_{irreducible}$$

Variance and Bias

As a general rule, as we use more flexible methods, the variance will increase, and the bias will decrease and vice versa. Good test set performance of a statistical learning method requires low variance as well as low squared bias.



Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve is the sum of these three quantities.) for the three data sets in previous examples. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

The Classification Setting

In the classification setting, model accuracy measure needs some modifications because y_i is qualitative. Given that f is estimated on the basis of training observations $\{(x_1, y_1), (x_2, y_2), . . . , (x_n, y_n)\}$, where now $y_1, . . . , y_n$ are qualitative. We use the **training error rate** to quantify the accuracy of our estimate $\hat{f}$.

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

where $\hat{y}_i$ is the predicted class label for the i^{th} observation using $\hat{f}$. And $I\left(y_i \neq \hat{y}_i\right)$ is an indicator variable that equals 1 if $y_i \neq \hat{y}_i$ and zero if $y_i = \hat{y}_i$.

$$I = \begin{cases} 1 \quad \text{if } \; y_i \neq \hat{y}_i \quad \text{i.e., } y_i \text{ was misclassified} \\ 0 \quad \text{if } \; y_i = \hat{y}_i \; \text{ i.e., } y_i \text{ was correctly classified} \end{cases}$$

The **training error rate** computes the fraction of incorrect classifications.

The Classification Setting

Similar to the regression setting, the **test error rate** associated with a set of test observations of the form (x_0, y_0) is given by

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

where $\hat{y}_0$ is the predicted class label that results from applying the classifier $\hat{f}$ to the test observation with predictor x_0. And $I\left(y_0 \neq \hat{y}_0\right)$ is an indicator variable.

$$I = \begin{cases} 1 \quad \text{if } y_0 \neq \hat{y}_0 \text{ i.e., } y_0 \text{ was misclassified} \\ 0 \quad \text{if } y_0 = \hat{y}_0 \text{ i.e., } y_0 \text{ was correctly classified} \end{cases}$$

The **test error rate** also computes the fraction of incorrect classifications. A good classifier is one for which the test error rate is smallest.

Example: Bayes Classifier

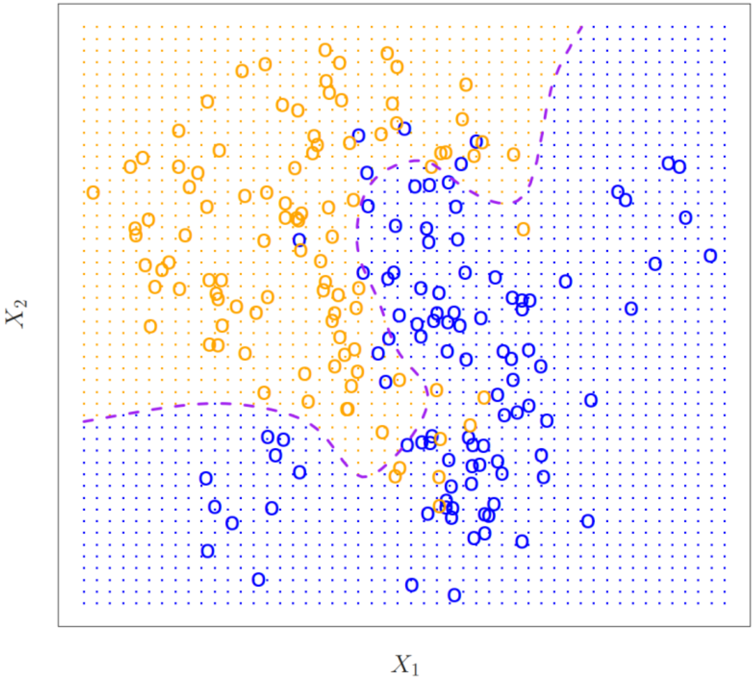
The Bayes Classifier simply assigns a test observation y_0 with predictor vector x_0 to the class j for which

$$\max_j \Pr\left(y_0 = j \mid X = x_0\right)$$

where $\Pr(y_0 \mid x_0)$ is a conditional probability, i.e., it is the probability that y_0 is assigned to class j, given the observed predictor vector x_0.

- For each value of X_1 and X_2, the Bayes classifier outputs a different probability of the response being orange or blue.
- The purple dashed line represents the points where the probability is exactly 50 %, called the **Bayes decision boundary**.
- In general, the overall Bayes error rate is given by

$$1 - E\left(\max_j \Pr\left(y_0 = j \mid X = x_0\right)\right)$$



Example: KNN classifier

In theory we would always like to predict qualitative responses using the Bayes classifier, because the test error rate is always minimized. However, if we do not know the conditional distribution of *Y* given *X*, computing the Bayes classifier is impossible.

The ***K*-nearest neighbors** (*KNN*) classifier attempts to **estimate** the conditional distribution of *Y* given *X*, and then classify a given observation to the class with highest **estimated** probability.

Given a positive integer *K* and a test observation x_0, the *KNN* classifier

- first identifies the *K* points in the training data that are closest to x_0, (by Euclidean distance), N_0 denotes *K* nearest points.
- then estimates the conditional probability for class *j* as the fraction of points in N_0 whose response values equal *j* :

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

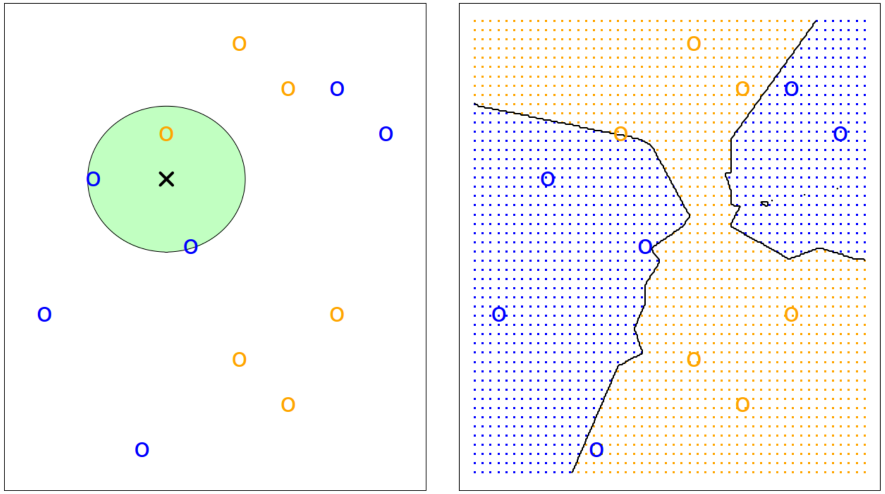
- Finally, *KNN* assigns the test observation y_0 to the class with the largest probability.

Example: KNN classifier

Euclidean distance, a.k.a, L^2 norm:

$$d(p,q) = \|p-q\|_2 = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

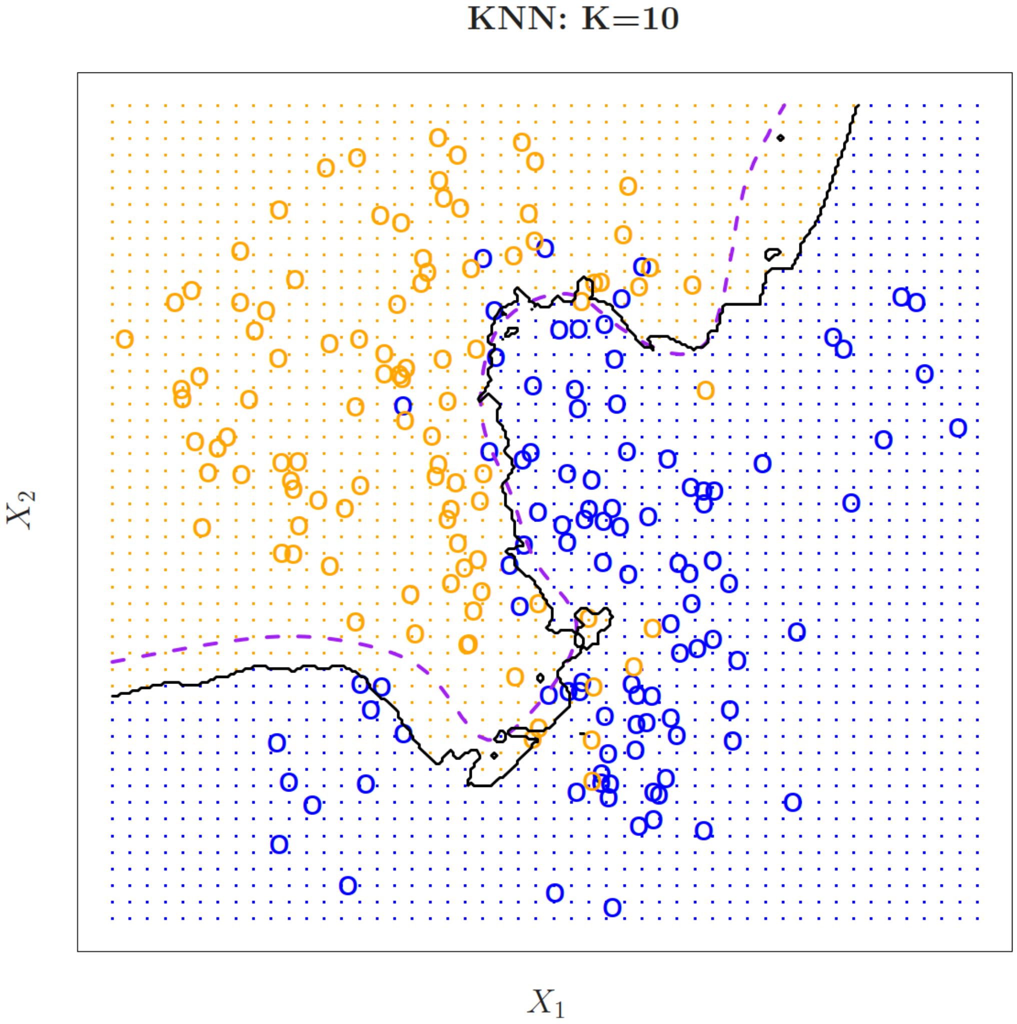
p, *q* = Euclidean vectors represent two points in Euclidean *n* dimensional space



The training data set consists of six blue and six orange observations. Choose *K*=3, *KNN* identifies the three nearest observations around black cross point, two blue points and one orange, resulting in estimated probabilities of 2/3 for the blue class and 1/3 for the orange class. Hence *KNN* will predict that the black cross belongs to the blue class. The right graph shows ***KNN* decision boundary** after applied the *KNN* at all the possible values for X_1 and X_2.

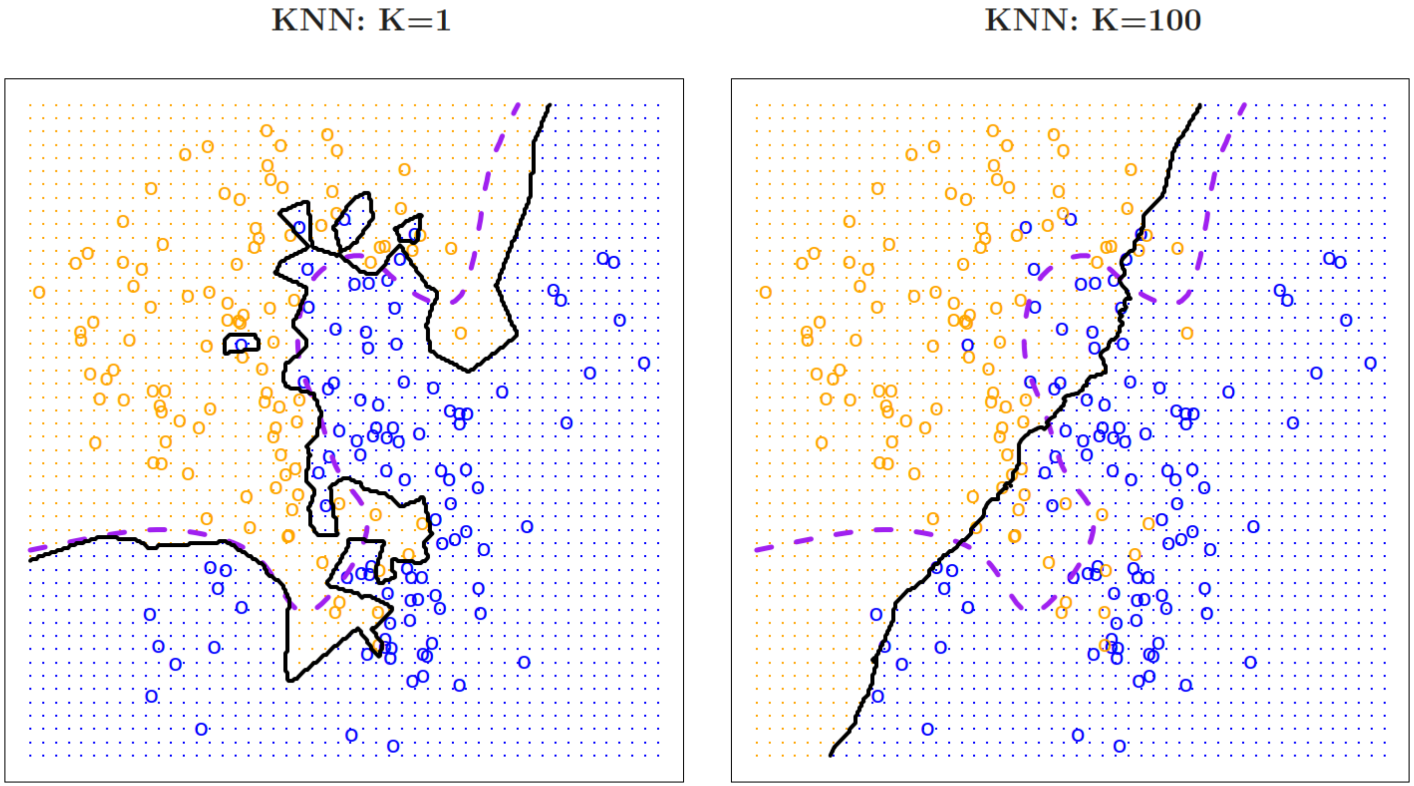
KNN vs. Bayes classifier

Even though the true distribution is not known by the *KNN* classifier, the *KNN* decision boundary using K = 10 is very close to that of the Bayes classifier.



KNN vs. Bayes classifier

The choice of *K* has a drastic effect on the *KNN* classifier.

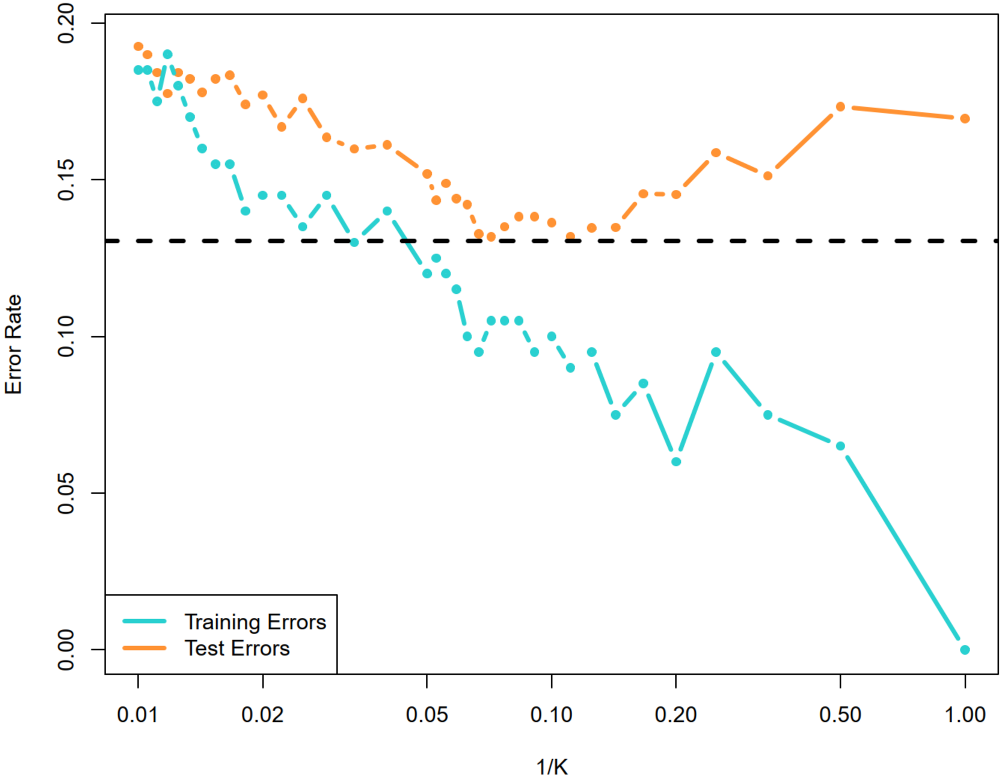


When K = 1, the decision boundary is overly flexible. This corresponds to a classifier that has low bias but very high variance.

As *K* =100, *KNN* becomes less flexible and produces a decision boundary that is close to linear. This corresponds to a low variance but high bias classifier.

Variance and Bias in Classifiers

In general, as we use more flexible classification methods, the training error rate will decline but the test error rate may not. The graph plotted the *KNN* test and training errors as a function of *1/K*. As *1/K* increases, the method becomes more flexible.



As in the regression setting, the training error rate consistently declines as the flexibility increases. However, the test error exhibits a *U*-shape, (with a minimum at approximately $K = 10$) before overfits.