Applied Data Analysis — Introduction

Dr. Lan Du & Dr Ming Liu

Faculty of Information Technology, Monash University, Australia

Week 1

Outline

Data Mining

- **Data mining** is the process of *automatically extracting information from large data sets*
- These data sets are usually *so large that manually examining them is impractical*
- The data sets can be *structured* (e.g., a database) or *unstructured* (e.g., free-form text in documents)
  - **Text data mining** uses *natural language processing* to extract information from *large text collections*
  - Quantitative data mining extracts information from numerical data
  - It's also possible to *integrate quantitative and qualitative information sources*

Data Mining

MONASH University

- Data mining permits businesses to *exploit the information present in the large data sets* they collect in the course of their business
- Typical business applications:
  - in *medical patient management*, data mining identifies patients likely to *benefit from a new drug or therapy*
  - in *customer relationship management*, data mining identifies customers likely to be *receptive to a new advertising campaign*
  - in *financial management*, data mining can help *predict the credit-worthiness of new customers*
  - in *load capacity management*, data mining predicts the fraction of customers with airline reservations that will actually *turn up for the flight*
  - in *market basket* and *affinity analysis*, data mining identifies *pairs of products likely (or unlikely) to be bought together*, which can help design advertising campaigns

Data Mining

- Diverse range of data mining tasks:
  - software packages exist for standard tasks, e.g., affinity analysis
  - but *specialised data mining applications require highly-skilled experts* to design and construct them
- Data mining is often *computationally intensive* and involve advanced algorithms and data structures
- Data mining may involve *huge data sets* too large to store on a single computer
  - often requires *large clusters* or *cloud computing* services

Machine Learning

MONASH University

- **Machine learning** is a branch of *Artificial Intelligence* that studies *methods for automatically learning from data*
- It focuses on *generalisation* and *prediction*
  - typical goal is to *predict the properties of yet unseen cases*

  $\Rightarrow$ split training set/test set methodology, which lets us estimate accuracy on *novel test data*
- Data mining can use machine learning, but it doesn't have to:
  - E.g., "who is the phone system's biggest user?" doesn't necessarily involve machine learning
  - E.g., "which customers are likely to increase their phone usage next year?" does involve machine learning

Statistical Modelling

- **Probability theory** is the branch of mathematics concerned with *random phenomena* and *systems whose structure and/or state is only partially known*
  - $\Rightarrow$ probability theory is a *mathematical foundation of machine learning*
- **Statistics** is the science of the *collection, organisation and interpretation of data*
  - A **statistic** is a function of data sets (usually numerically-valued) intended to *summarise the data* (e.g., the *average* or *mean* of a set of numbers)
- A **statistical model** is a mathematical statement of the *relationship between variables that have a random component*
  - many machine learning algorithms are based on statistical models
  - statistical models also play a central role in natural language processing

Statistics vs Machine Learning

- Statistics and machine learning often use *the same statistical models*
  $\Rightarrow$ very strong cross-fertilisation between fields
- Machine learning often involves data sets that are *orders of magnitude larger* than those in standard statistics problems
  - Machine learning is concerned with algorithmic and data structure issues that statistics doesn't deal with
- Statistics tends to focus on *hypothesis testing*, while machine learning focuses on *prediction*
  - **Hypothesis testing:** *Does coffee cause cancer?*
  - **Prediction:** *Which patients are likely to die of cancer?*

Supervised vs Unsupervised Learning

MONASH University

- Machine learning algorithms usually learn from training data.
- **Supervised training data** contains the labels y that we want to predict.
  - E.g., in *Part-of-Speech (PoS) tagging*, the training data may be a *corpus* containing words *labelled with their parts-of-speech*

  Unsupervised training data does not contain the labels y that we want to predict.
  - E.g., in *topic modelling* we are given a large collection of documents without any topic labels. Our goal is to group them by topic (i.e., $\mathcal{Y}$ is a set of topics, and our goal is to learn a function f that maps each document x to its topic $y = f(x)$).

There are intermediate possibilities between supervised and unsupervised training data. **Semi-supervised training data** partially identifies the labels y, or identifies the labels on some but not all of the training examples.
  - E.g., in *PoS* tagging, we may be given a small corpus of PoS-tagged words, and a much larger corpus of words without PoS tags

Different Types of Learning Algorithms

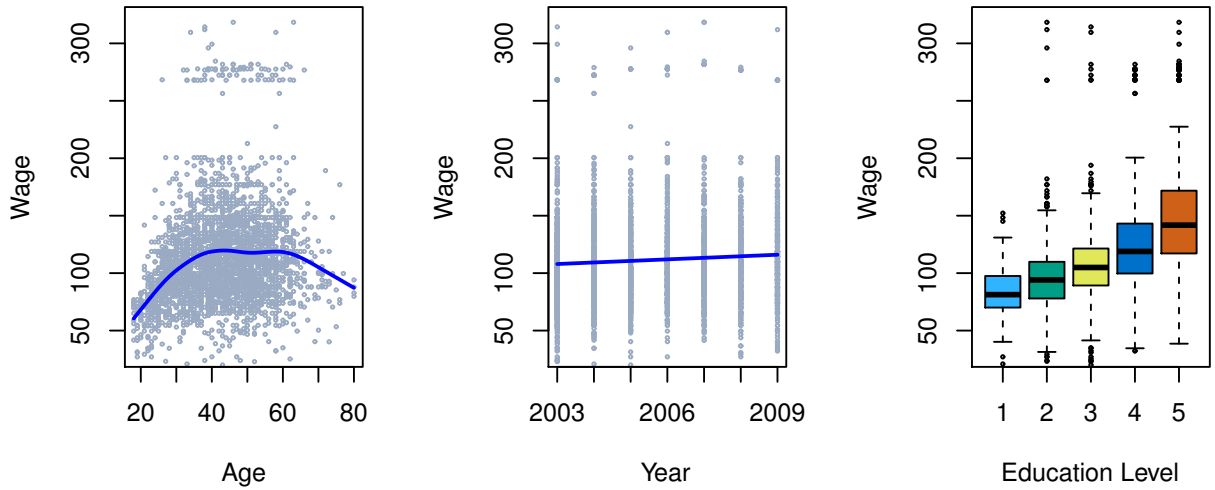
MONASH University

- The kinds of machine learning algorithms used *depend on whether the labels are discrete or continuous, and whether the data is supervised or unsupervised*

	Discrete labels	**Continuous labels**
Supervised data	*classification*	*regression*
Unsupervised data	*clustering*	*dimensionality reduction*

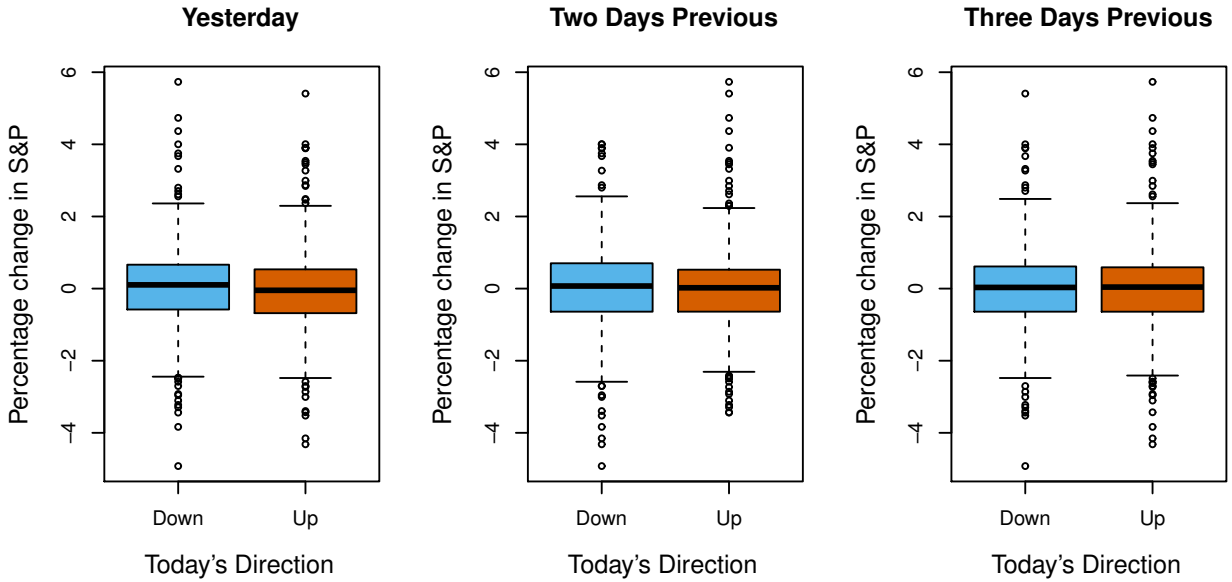
Applications: Wage Data [Chapters 3 and 7]

- We wish to understand the association between an employee's wage and
  - his age, education and calendar year
- Left: wage, as a function of age, increases by age but not after 60!
- The blue curve shows an estimate of average wage for a given age
  - Regression: using this curve to predict someone's wage, a continuous value,
- Centre: wage as a function of year
- Right: boxplots showing wage as a function of education



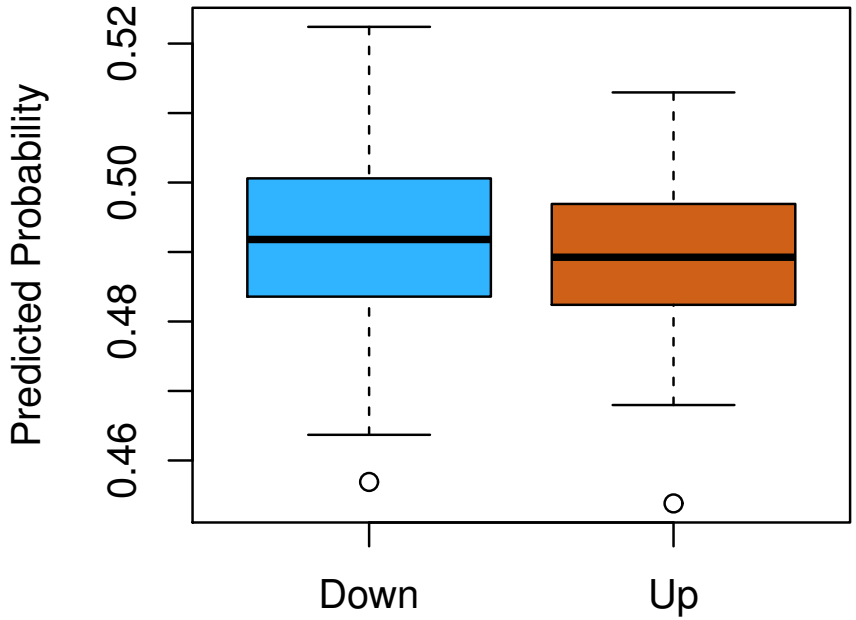
Application: Stock Market [Chapter 4]

- Stock market data on daily movements over a 5-year period
- Aim: predict if the market is going to be Up or Down (categorical label)
- This aim needs classification.
- Left: 648 days the market increased on the next day, and 602 days for the market decreased.
- Centre: considering 2 days percentage change
- Right: considering 3 days percentage change



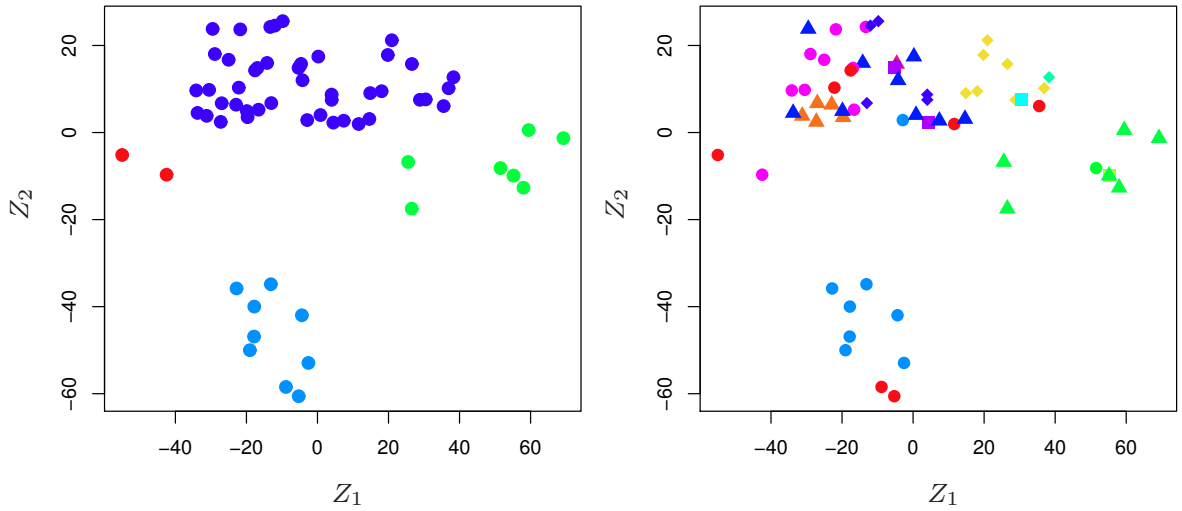
Application: Stock Market

- A quadratic discriminant analysis model is fitted to the subset of the market data (2001-2004) Training Set
- Predicted the probability of a stock market decrease using the 2005 data Test Set
- Correctly predict the direction of movement in the market 60% of the time



Clustering Gene Expression Data

- NCI60 dataset, 6,830 gene expression measurements for 64 cancer cell lines
- Instead of predicting a particular output variable
- we are interested in determining whether there are groups, or clusters, among the cell lines
- based on their gene expression measurements
- thousands of gene expression measurements per cell line, hard to visualise
- Left: gene expression data in 2-dimensional space
- Right: same data with different colour for 14 types of cancer



Outline

Unit Outcomes

- **Synopsis**: This unit aims to provide students with the necessary analytical and data modeling skills for the roles of a data scientist or business analyst. Students will be introduced to established and contemporary Machine Learning techniques for data analysis and presentation using widely available analysis software. They will look at a number of characteristic problems/data sets and analyse them with appropriate machine learning and statistical algorithms. Those algorithms include regression, classification, clustering and so on. The unit focuses on understanding the analytical problems, machine learning models, and the basic modeling theory. Students will need to interpret the results and the suitability of the algorithms.
- Note that
  - Overlap with FIT5201 and FIT5197 on the standard regression/classification method.

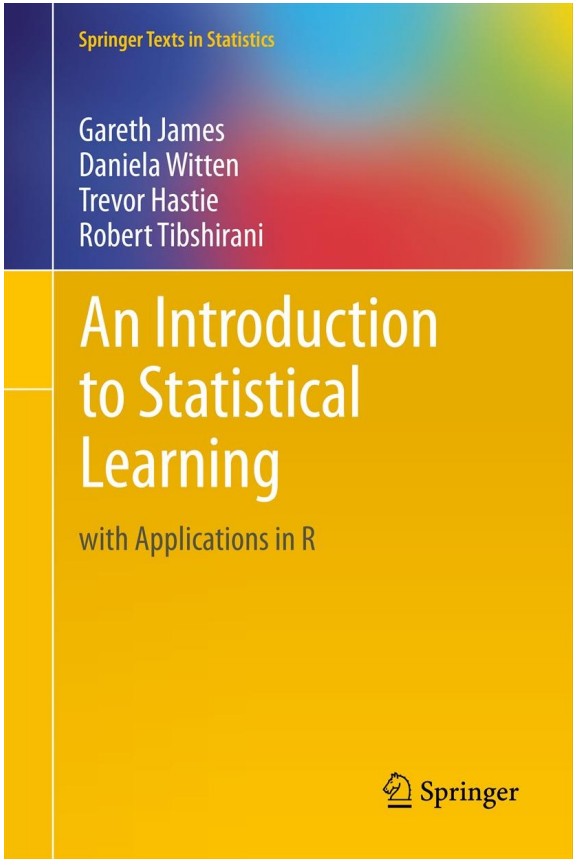
Unit Outcomes

Students are expected to

- Analyse data sets with a range of statistical, graphical and machine-learning tools;
- Evaluate the limitations, appropriateness and benefits of data analytics methods for given tasks;
- Design solutions to real world problems with data analytics techniques;
- Assess the results of an analysis;
- Communicate the results of an analysis for both specific and broad audiences.

For example, apply a statistical learning model to a toy dataset.

Books



R and Data Mining
Examples and Case Studies

Yanchang Zhao

RDataMining.com

ELSEVIER
AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE
SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Unit Contents

Week	Activities	Assessment
0		No formal assessment or activities are undertaken in week 0
1	Introduction to statistical learning	
2	Exploratory data analysis	
3	Linear regression	
4	Classification	
5	Re-sampling methods	
6	Linear model selection, and regularization	
7	Nonlinear methods	Assessment 1: Mining Knowledge from Data due Friday 1 May 2020
8	Tree-based methods	
9	Support Vector Machines	
10	Semi-supervised learning	
11	Unsupervised learning	
12	Unsupervised learning + summary	Assessment 2: Data Analysis Challenge due Sunday 7 June 2020
	SWOT VAC	No formal assessment is undertaken in SWOT VAC
	Examination period	LINK to Assessment Policy:http://policy.monash.edu.au/policy-bank/academic/education/assessment/assessment-in-coursework-policy.html

Assessments

Table: Assessment Summary

Assessment task	Value	Due Date
Mining Knowledge from Data	15%	Friday 1 May 2020
Data Analysis Challenge	35%	Sunday 7 Jun 2020
Final Exam	50%	To be advised

- **It is a very bad idea to leave the assessments until the last minute.**

Programming Requirements

MONASH University

- R
- Jupyter notebook/R Markdown
- Lynda.com through Monash library
- Google it!

Lecture/Tutorial Schedule

Contacts

- Lecturers
  - **Lan Du**: lan.du@monash.edu
    - Office: Room 329, Room 329, Building 6 - 29 Ancora Imparo Way, Clayton, VIC 3800
    - Consultation: Appointment needed for meetings outside the consultation times.
  - **Ming Liu**: grayming.liu@monash.edu
    - Consultation: Appointment needed for meetings outside the consultation times.
- Tutors
  - **Minh Le** : minh.le@monash.edu
  - **Tam Vo**: tam.vo@monash.edu
  - **Yuan Jin**: yuan.jin@monash.edu
  - **Dan Nguyen**: dan.nguyen2@monash.edu
- consultations
  - Consultation will start in **Week 3**.
  - **Consultation times will be announced in Moodle.**
  - **Make use of all the consultations**

Outline

Motivation

MONASH University

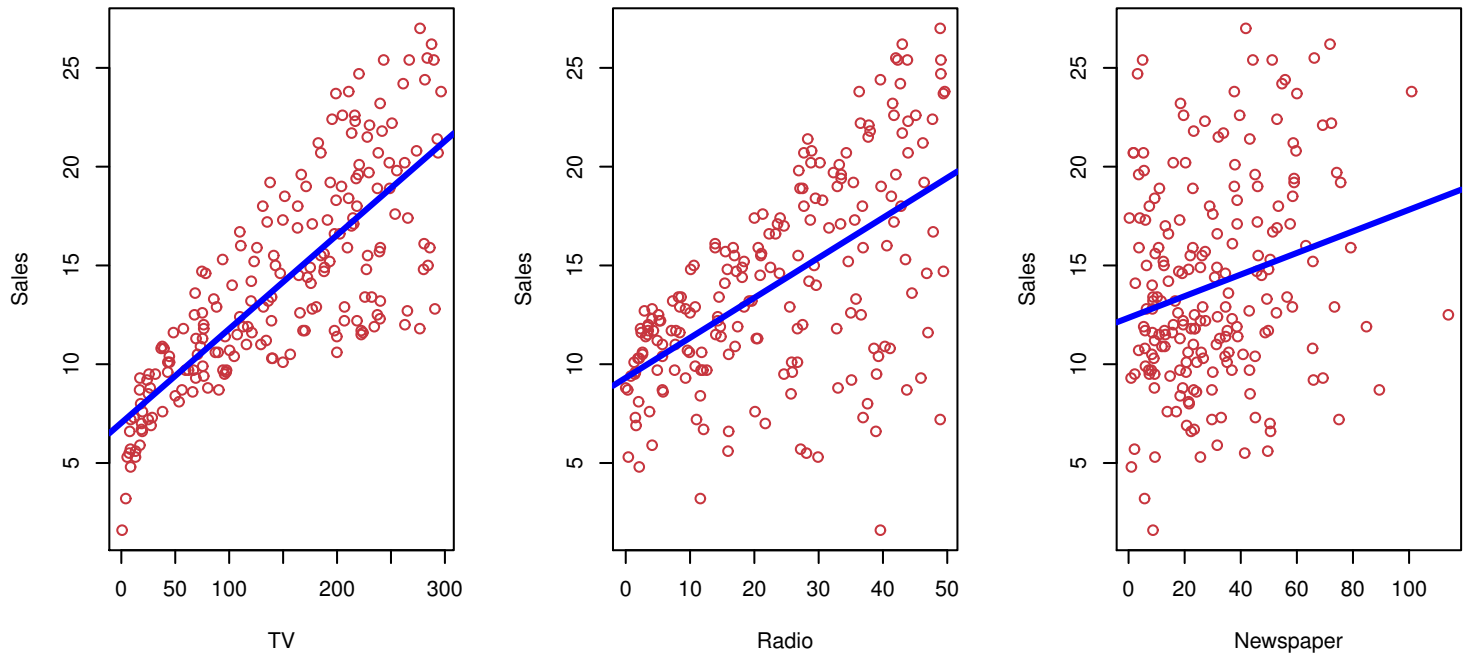
- Suppose you are a consultant
- **Task\Question** How to improve sales of a particular product
- The Advertising data, sales in thousands of dollars, as a function of TV, Radio and Newspaper budgets for 200 different markets
- They cannot directly increase sales of the product
- They can control the advertising expenditure in each of the three media

	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
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6
12	214.7	24	4	17.4
13	23.8	35.1	65.9	9.2

- If there is an association between advertising and sales, then client can adjust advertising budgets, to indirectly increasing sales

Motivation

- Our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets
- Simple least square fit of sales to the variables
- Each blue line represents a simple model that can be used to predict sales using TV, Radio, and Newspaper



Definitions

- Input variables: advertising budget
  - denoted by X with a subscript. $\mathbf{X} = (X_1, X_2, X_3)$
  - X_1 the TV budget, X_2 the Radio budget, and X_3 the Newspaper budget
  - Predictors, independent variables, features, variables
- Output variable: sales
  - denoted by Y
  - response, and dependent variable

	TV X1	Radio X2	Newspaper X3	Sales Y
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
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6
12	214.7	24	4	17.4
13	23.8	35.1	65.9	9.2

Definitions

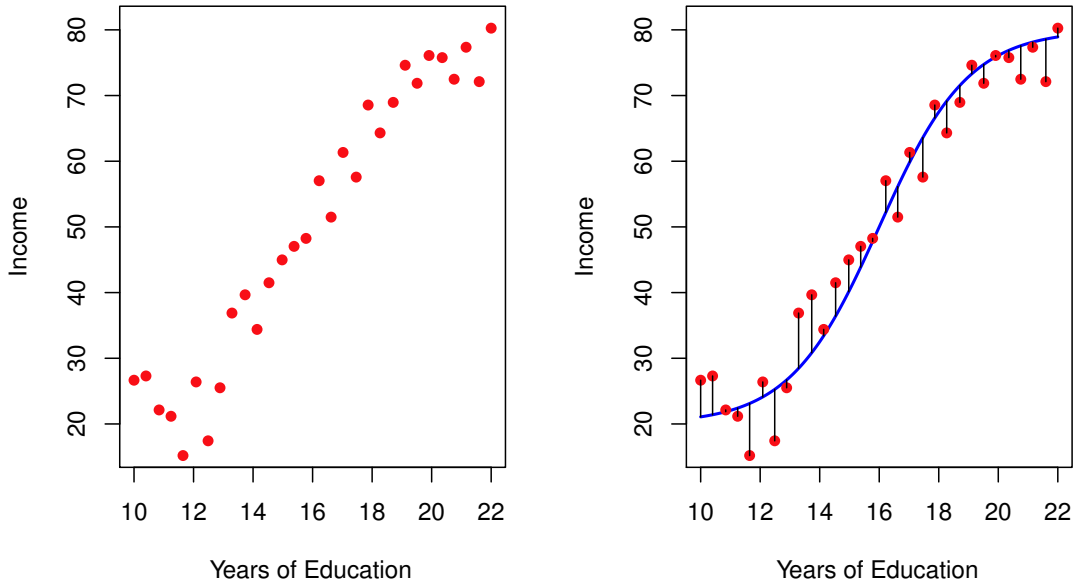
- Suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$
- Assume some relationship between Y, and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$

$$Y = f(\mathbf{X}) + \epsilon$$

  - Here $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ is an observation (input)
  - With features $X_1, X_2, \ldots, X_p$
  - Y is the output or response variable
- f is a fixed but unknown function of $X_1, X_2, \ldots, X_p$
- ϵ is a random error term, independent of $\mathbf{X}$ and $E[\epsilon] = 0$
- f represents the systematic information that $\mathbf{X}$ provides about Y

Another example: f with one variable

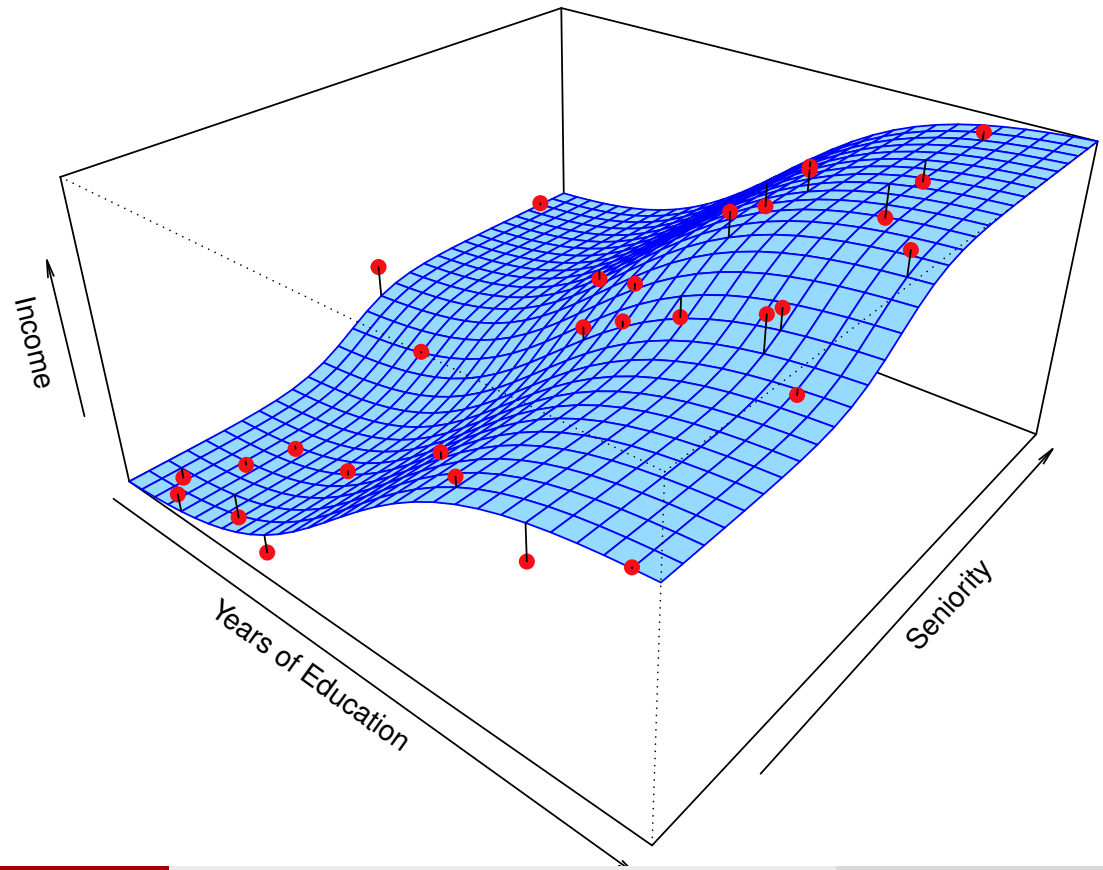
- Income versus years of education for 30 individuals
- One might be able to predict income using years of education
- However, the function f is unknown
- One must estimate f based on the observed points
- Income is a simulated data set, f is known and is shown by the blue curve
- The vertical lines are the error ϵ [why are some points above or under?]



Another example: f with more than one variable

- Income as a function of years of education and seniority
- f is a two-dimensional surface (true underlying relationship)
- Must be estimated based on the observed data

$$\text{Income} = f(\text{Years}, \text{Seniority})$$



Statistical Learning

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$
$$\hat{Y} = \hat{f}(X_1, X_2, \ldots, X_p)$$
$$Y \approx \hat{Y}$$

- In essence, statistical learning refers to a set of approaches for estimating f
- Statistical learning: A set of tools for understanding data
- In this lecture
  - We outline some of the key concepts that arise in estimating f
  - And tools for evaluating the estimates obtained

Two Main Reasons: Prediction and Inference

MONASH University

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \text{ or } Y = f(\mathbf{X}) + \epsilon$$

- Prediction: to find an estimator $\hat{Y} = \hat{f}(\mathbf{X})$
  - $\hat{f}$ is treated as a black box, we are not concerned what is it
  - We are interested in accurate predictions for Y
  - The accuracy of $\hat{Y}$ as a prediction of Y depends on
    - Reducible error: $|f - \hat{f}|$
    - We can potentially improve the accuracy by using better learning technique to estimate
    - Irreducible error: Y is a function of ϵ and the error cannot be predicted using $\mathbf{X}$
    - No matter how well we estimate f, we cannot reduce the error ϵ

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

  - $E[(Y - \hat{Y})^2]$ the average value of squared difference between the prediction and true value
  - $Var(\epsilon)$ the variance associated with the error term ϵ
  - $E[(Y - \hat{Y})^2] \geq Var(\epsilon)$
  - The focus of this unit is to minimize the reducible error

Two Main tasks: Prediction and Inference

MONASH University

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \text{ or } Y = f(\mathbf{X}) + \epsilon$$

- Inference: to understand the way Y is affected by $X_1, X_2, \ldots, X_p$
- We wish to estimate f, but our goal is not necessarily to make predictions
- We want to understand the relationship between $\mathbf{X}$ and Y
- How Y changes as a function of $X_1, X_2, \ldots, X_p$
- $\hat{f}$ cannot be a black box; we need to know its exact form
- We want to answer the following questions:
  - 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?

Estimate f

MONASH University

- Training data: n different data points (or observations) $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
- For $i = 1, \ldots, n$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$
- x_{ij} is the j^{th} predictor of observation i for $i = 1, \ldots, n$ and $j = 1, \ldots, p$
- y_i response variable for the i^{th} observation
- training set is $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$
- Goal: to apply a statistical learning method to the training data to estimate the unknown function f
- We want to find $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any (X, Y)
- Most statistical learning methods are categorized as
  - Parametric Methods
  - Non-parametric Methods

Parametric Methods

Two-step model-based approach: reduces the problem of estimating f down to one of estimating a set of parameters:

1. We make an assumption about the functional form, or shape, of f

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

   - This is a linear model
   - Instead of estimating an entirely arbitrary p-dimensional function $f(\mathbf{X})$,
   - One only needs to estimate the $p + 1$ coefficients $\beta_0, \ldots, \beta_p$

2. After a model has been selected, we need a procedure that uses the training data to fit or train the model
   - we need to estimate the parameters $\beta_0, \ldots, \beta_p$

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p$$

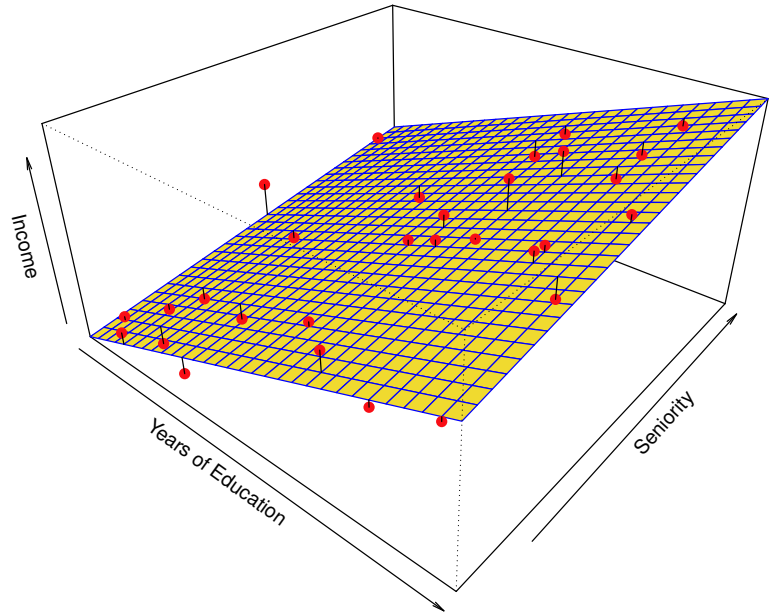
   - The most common approach to fitting the model is referred to as (ordinary) least squares

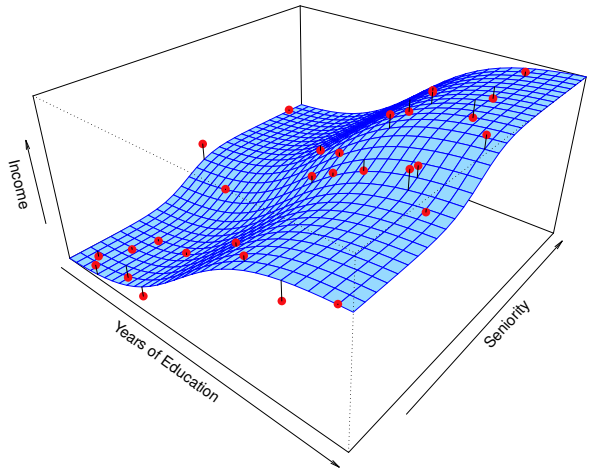
Parametric Methods - Disadvantage

MONASH University

- Assuming a parametric form for f simplifies the problem of estimating f
- It is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f
- Potential disadvantage is that the model we choose will usually not match the true unknown form of f
- If the chosen model is too far from the true f, then our estimate will be poor
- We can try to address this problem by choosing flexible models
- Fitting a more flexible model requires estimating a greater number of parameters
- More complex models leads to a phenomenon known as overfitting the data
- Which essentially means they follow the errors, or noise, too closely

Example





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

- Entire fitting problem reduces to estimating β_0, β_1 and β_2
- Capturing the positive relationship between years of education and income

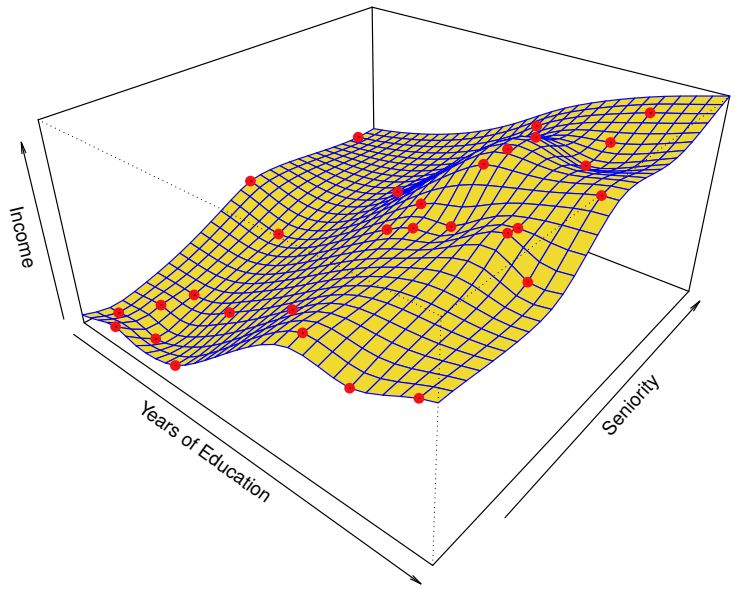
- Income as a function of years of education and seniority in the Income data set

Non-parametric Methods

MONASH University

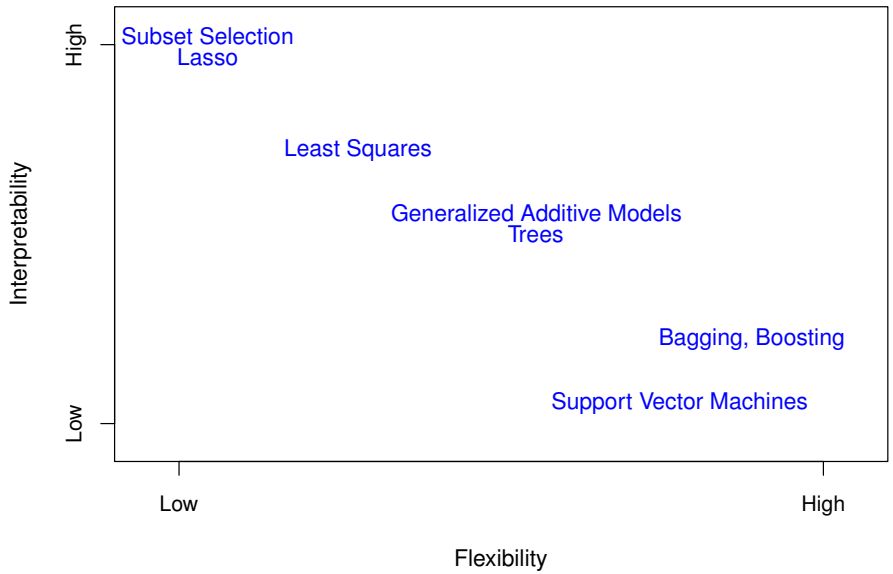
- Non-parametric methods do not make explicit assumptions about the functional form of f
- Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly
- By avoiding the assumption of a particular functional form for f,
- They have the potential to accurately fit a wider range of possible shapes for f

- Disadvantage: a large number of observations is required in order to obtain an accurate estimate for f
- A thin-plate spline is used to estimate f
- No prespecified model on f
- A remarkably accurate estimate of the true f



Trade-Off Between Accuracy and Model Interpretability

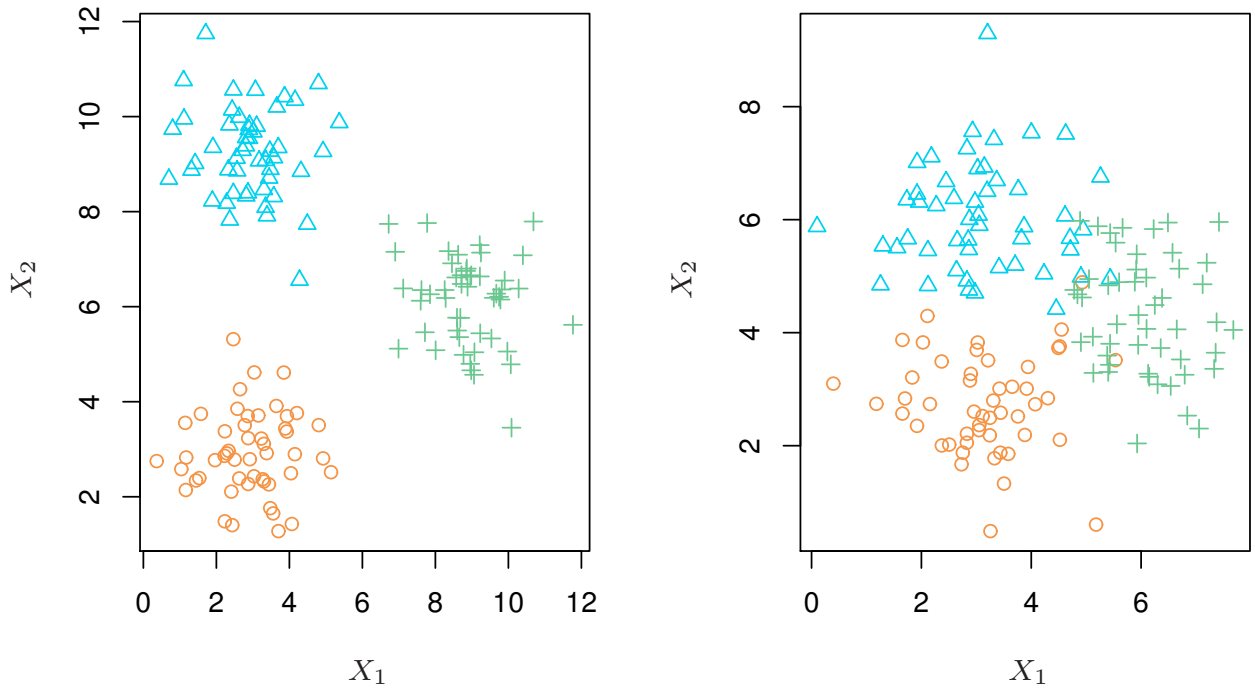
- Some methods are less flexible
- Linear regression is a relatively inflexible approach
- Other methods are considerably more flexible
- They can generate a much wider range of possible shapes to estimate f
- Why would choose a more restrictive method instead of a very flexible?
- Mainly interested in inference



Supervised Learning v.s. Unsupervised Learning

MONASH University

- A clustering data set involving three groups
- Left: a clustering approach should successfully identify the three groups
- The three groups are well-separated
- Right: There is some overlap among the groups.
- Now the clustering task is more challenging



Regression Versus Classification Problems

MONASH University

- Variables can be characterized as either quantitative (numerical) or qualitative (categorical)
- Regression problems: problems with a quantitative response
- Classification problems: those involving a qualitative response are often referred
- the distinction is not always that crisp
  - Least squares linear regression is used with a quantitative response
  - Logistic regression is typically used with a qualitative response
  - Logistic regression estimates class probabilities
  - K-nearest neighbours and boosting can be used in the case of either quantitative or qualitative responses
- Whether the predictors are qualitative or quantitative is generally considered less important

Outline

Selecting a statistical learning procedure

MONASH University

- No free lunch in statistics: No one method dominates all others over all possible data sets
- 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

Measuring the Quality of Fit

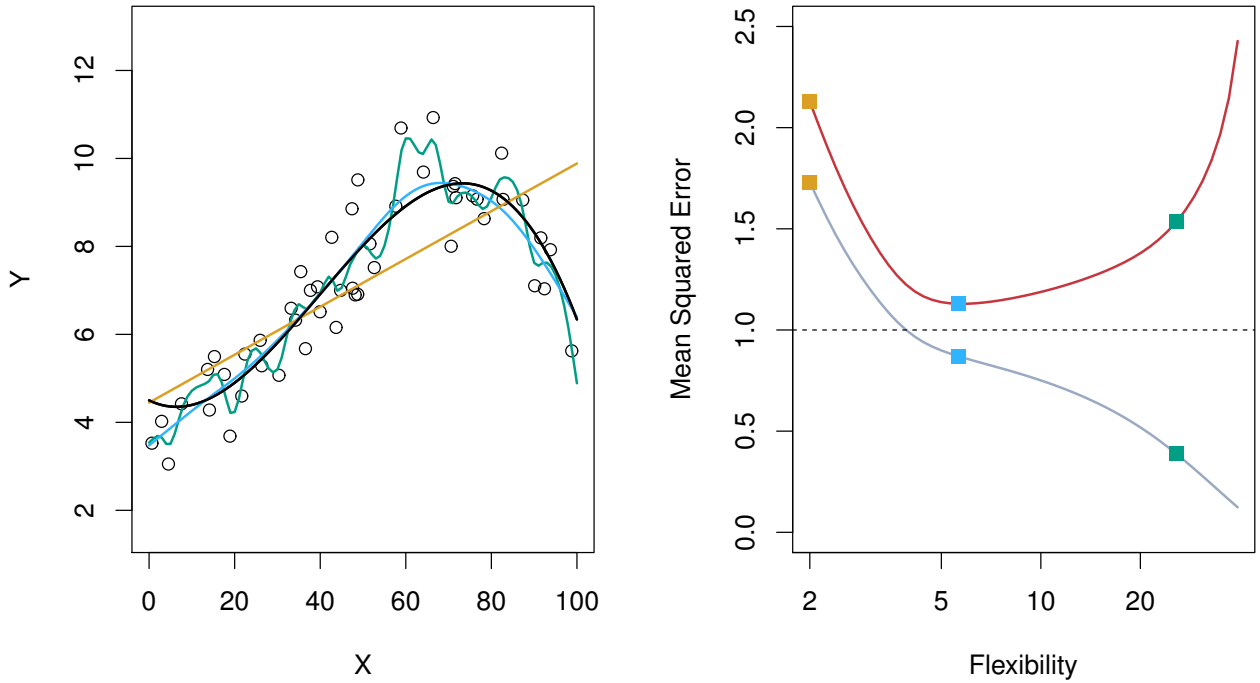
MONASH University

- We need some way to measure how well a method's predictions actually match the observed data
- In the regression setting, the most commonly-used measure is the mean squared error (MSE)

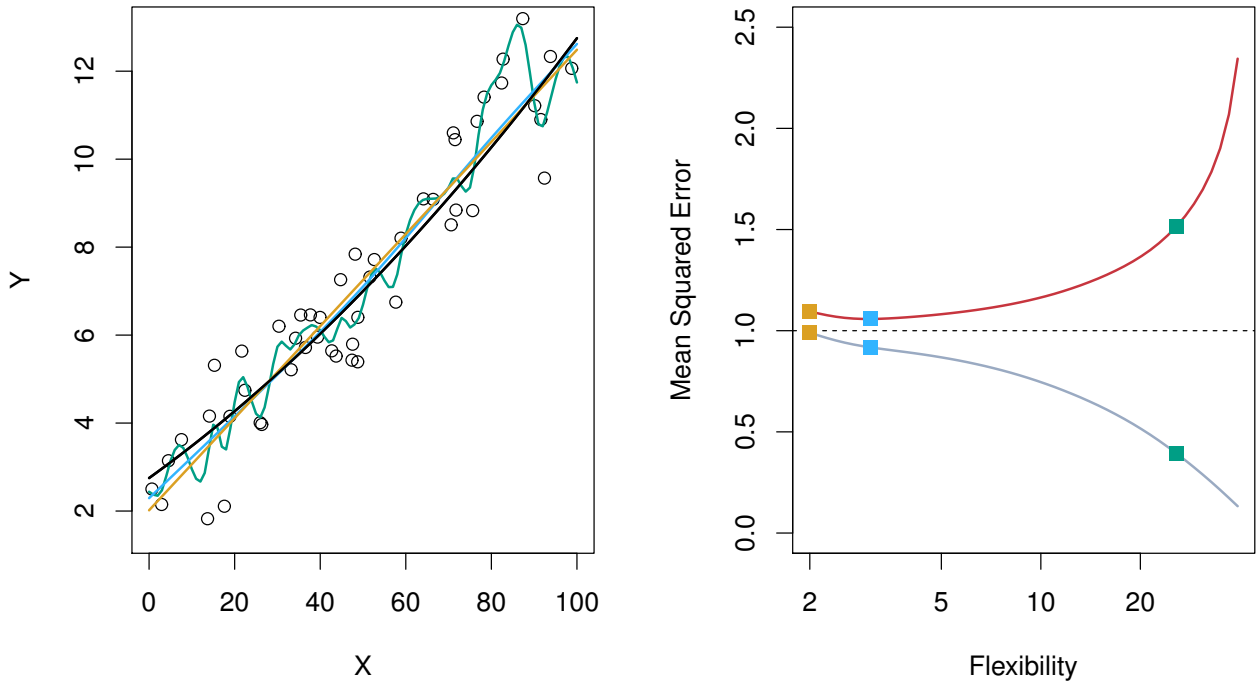
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2$$

- Training MSE
  - We are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data
- Test Set MSE:
  - previously unseen observation not used to train the statistical learning method
  - Average squared prediction error for these test observations

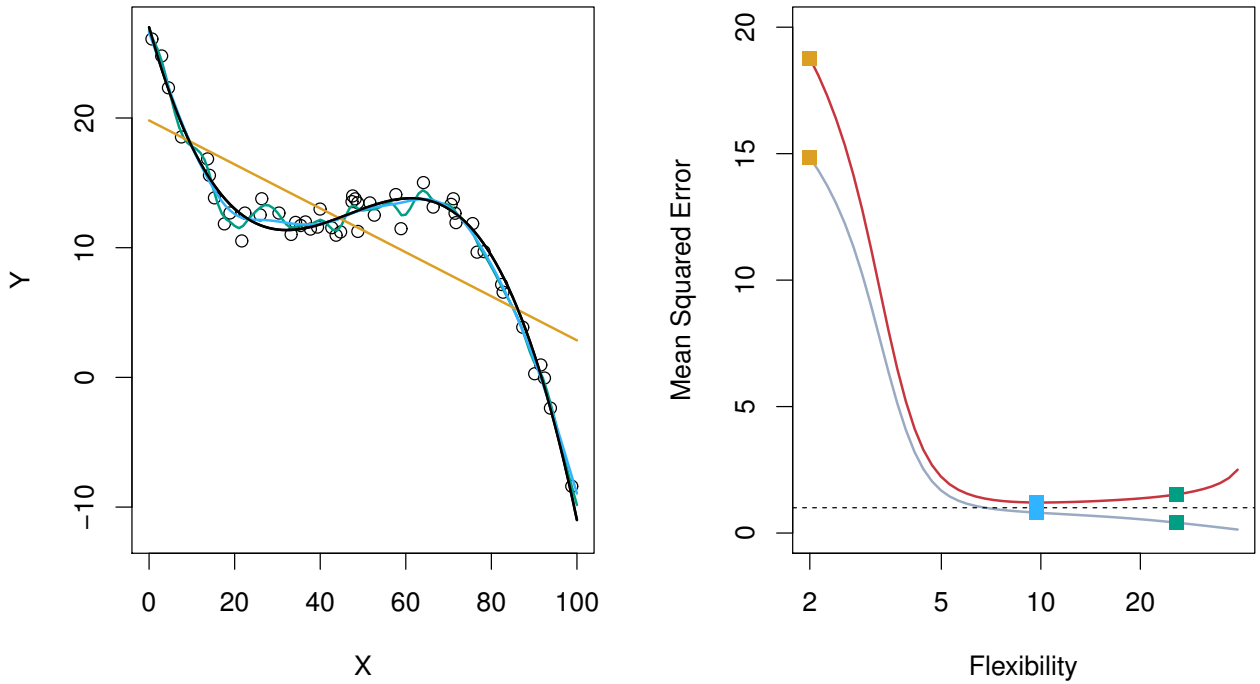
- How can we go about trying to select a method that minimizes the test MSE?
- We may have a test data set available
- But what if no test observations are available?
- Selecting a method based on training MSE!
- There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE
- In practice, one can usually compute the training MSE with relative ease,
- But estimating test MSE is considerably more difficult because usually no test data are available.
- One important method is cross-validation, which is a method for estimating test MSE using the training data.



- 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)
- Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line)



- True f that is much closer to linear, shown in black
- Linear regression provides a very good fit to the data
- Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line)



- f is far from linear
- llinear regression provides a very poor fit to the data
- Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line)

The Bias-Variance Trade-Off

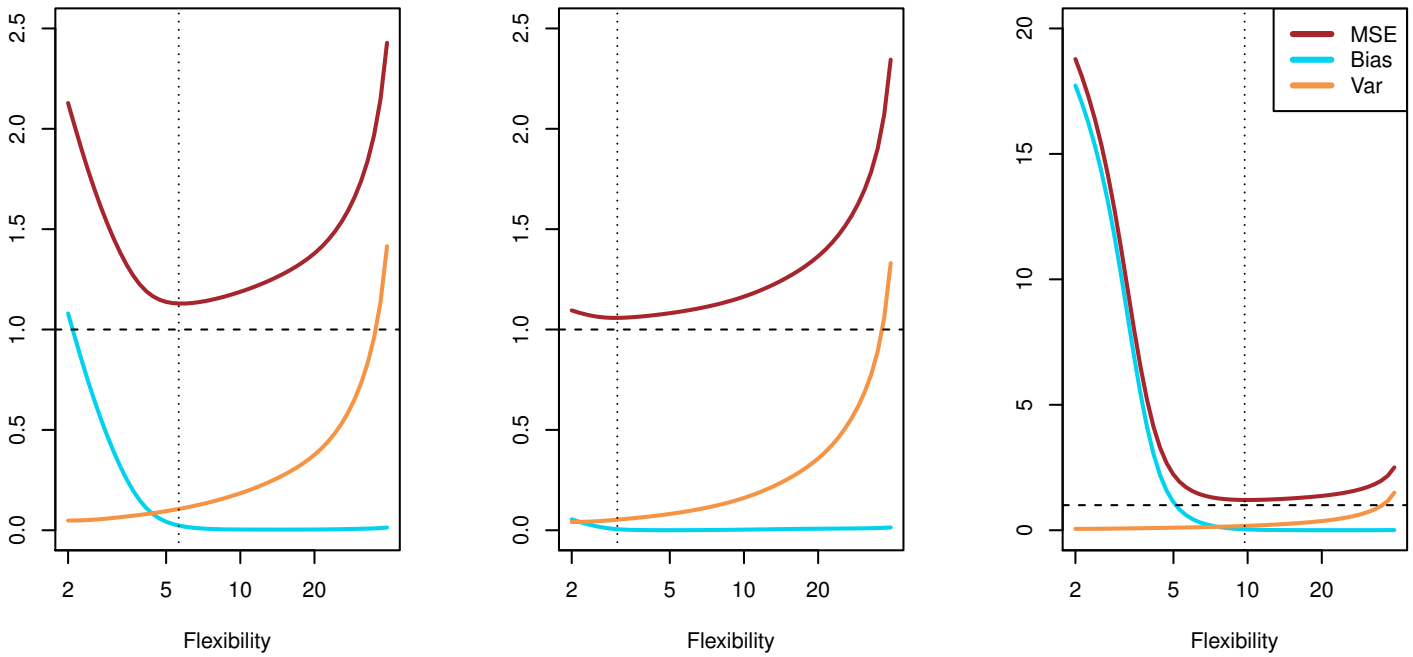
MONASH University

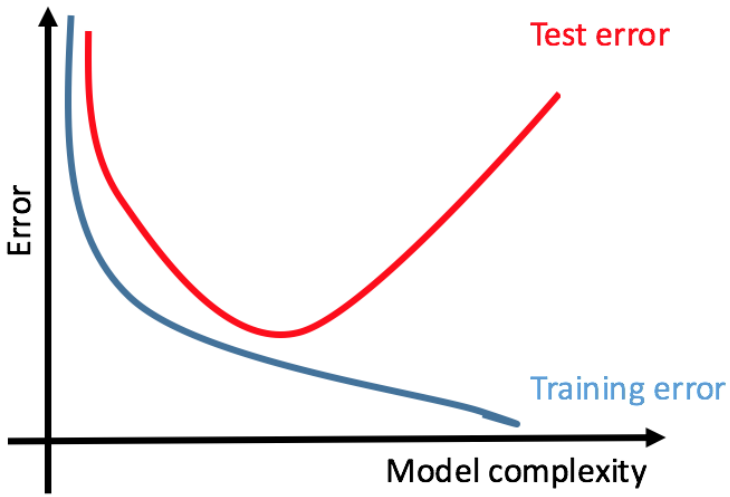
- U-shape test MSE is due to two competing properties of learning methods
- Test MSE at a point $\mathbf{x}_0$ has three components: Three sources of errors:
  - The variance of $\hat{f}$
  - The squared bias of $\hat{f}$
  - The variance of error term ϵ

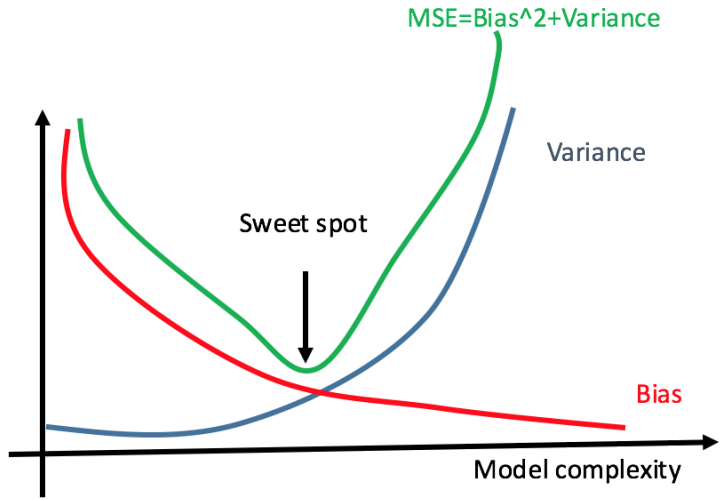
$$\underbrace{E\left[(y_0 - \hat{f}(\mathbf{x}_0))^2\right]}_{\text{expected test MSE}} = \underbrace{Var(\hat{f}(x_0)) + [Bias(\hat{f}(\mathbf{x}_0))]^2}_{\text{Reducible error}} + \underbrace{Var(\epsilon)}_{\text{Irreducible error}}$$

- If we have many training sets, for each of them one can find its β's such that
- $\bar{f} = E[\hat{f}(\mathbf{x}_0)]$
- $Var(\hat{f}(\mathbf{x}_0)) = E[(\hat{f}(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0))^2]$
- $Bias(\hat{f}(\mathbf{x}_0)) = f(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)$
  - Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- Low variance and low bias
- Expected test MSE can never lie below $Var(\epsilon)$

- Squared bias (blue curve)
- Variance (orange curve)
- $Var(\epsilon)$ dashed line
- And test MSE (red curve)
- The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE







- The challenge lies in finding a method for which both the variance and the squared bias are low.
- This trade-off is one of the most important recurring themes in this unit

- In classification, y_i are qualitative
- Training observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- Quantifying the accuracy of our estimate $\hat{f}$ by means of the training error rate
- The proportion of mistakes that are made if we apply our estimate $\hat{f}$ to the training observations, missclassifications

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

- $\hat{y}_i$ predicted class label for the i^{th} observation using $\hat{f}$
- $I(y_i \neq \hat{y}_i)$ 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 i^{th} observation was classified correctly
- Otherwise it was misclassified
- The formula gives the fraction of incorrect classifications

- Training error rate

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

- Test error rate associated with a set of test observations

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

$$= \frac{1}{m}\sum_{j=1}^{m} I(y_j \neq \hat{y}_j)$$

- A good classifier is one for which the test error is smallest

The Bayes Classifier

MONASH University

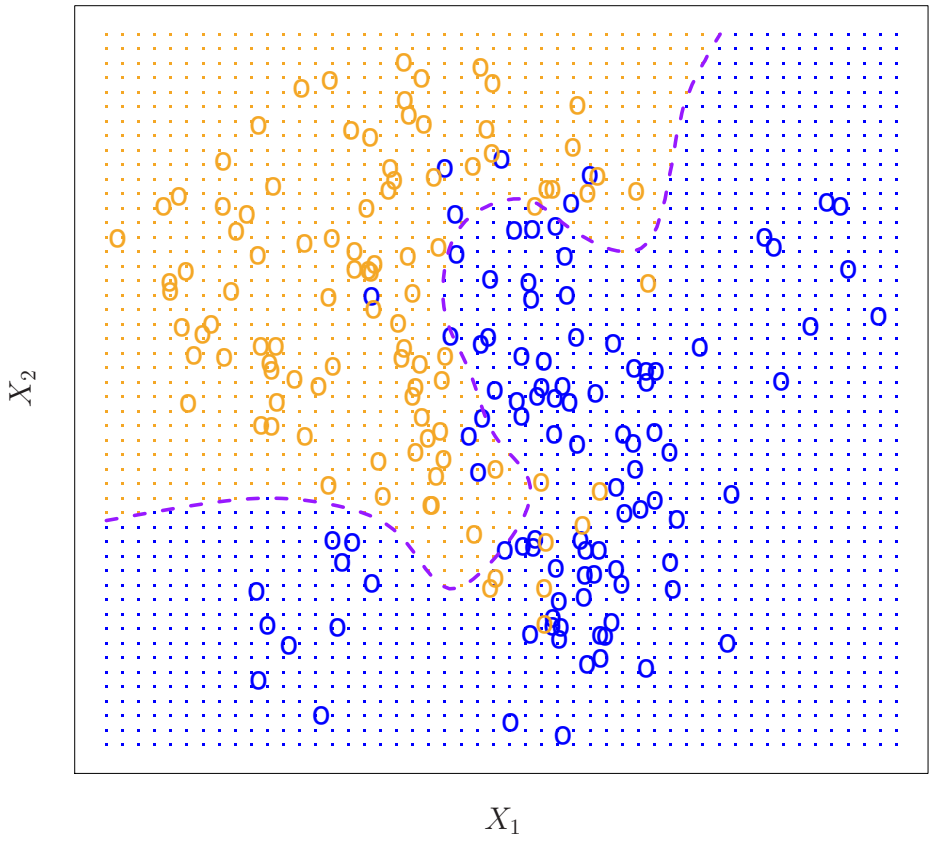
- The test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values
- We should simply assign a test observation with predictor vector x_0 to the class j for which the probability is largest

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

- This very simple classifier is called the Bayes classifier
- If we have 2 classes: class one and class two
- It predicts the class is one if $\Pr(Y = 1|X = x_0) > 0.5$
- The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate
- Bayes error rate is given by

$$1 - E\left[\max_j \Pr(Y = j|X)\right]$$

- A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange.
- The purple dashed line represents the Bayes decision boundary
- The orange background grid indicates the region in which a test observation will be assigned to the orange class,
- The blue background grid indicates the region in which a test observation will be assigned to the blue class



K-Nearest Neighbors

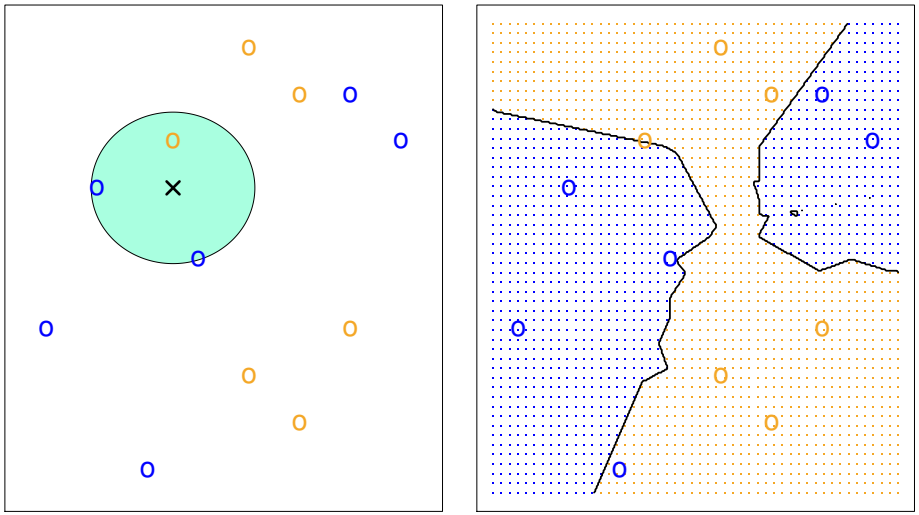
MONASH University

- Bayes classifier is good in theory
- For real data, we do not know the conditional distribution of Y given X
- Many approaches attempt to estimate the conditional distribution of Y given X,
- And then classify a given observation to the class with highest estimated probability
- One such method is the K-nearest neighbors (KNN) classifier
- Given a positive integer K and a test observation x_0
  1. The KNN classifier first identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$
  2. Then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j

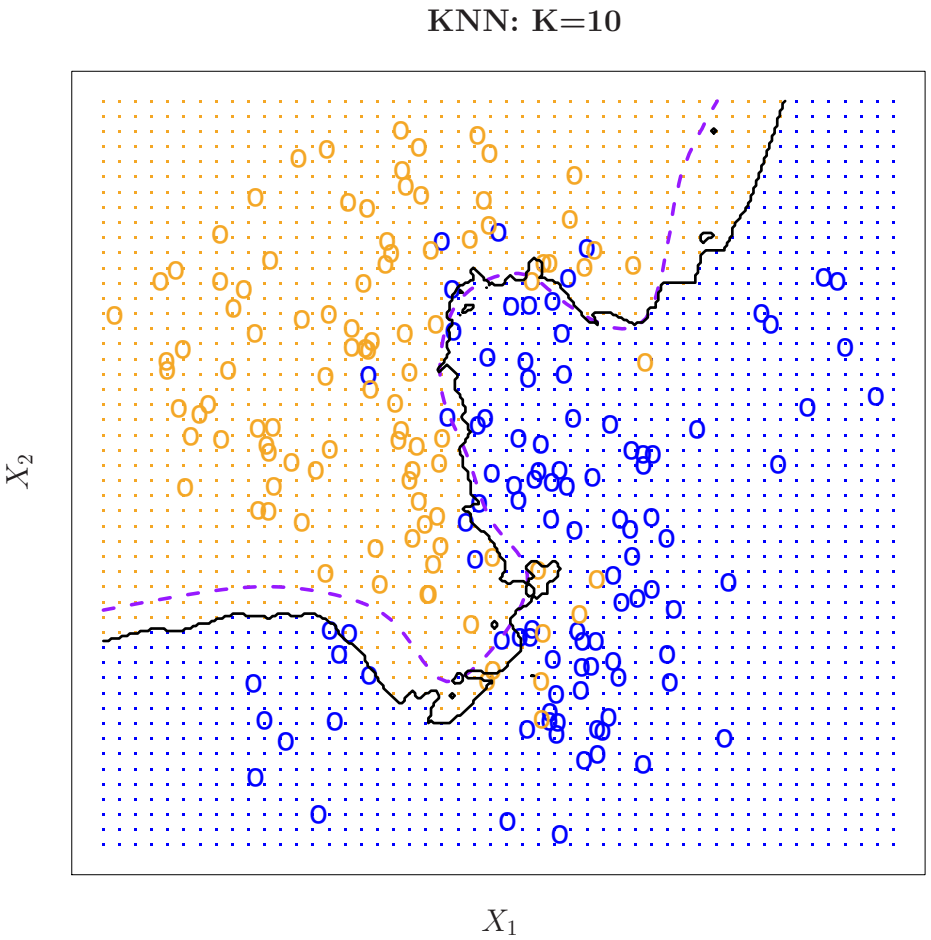
$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

  3. Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability

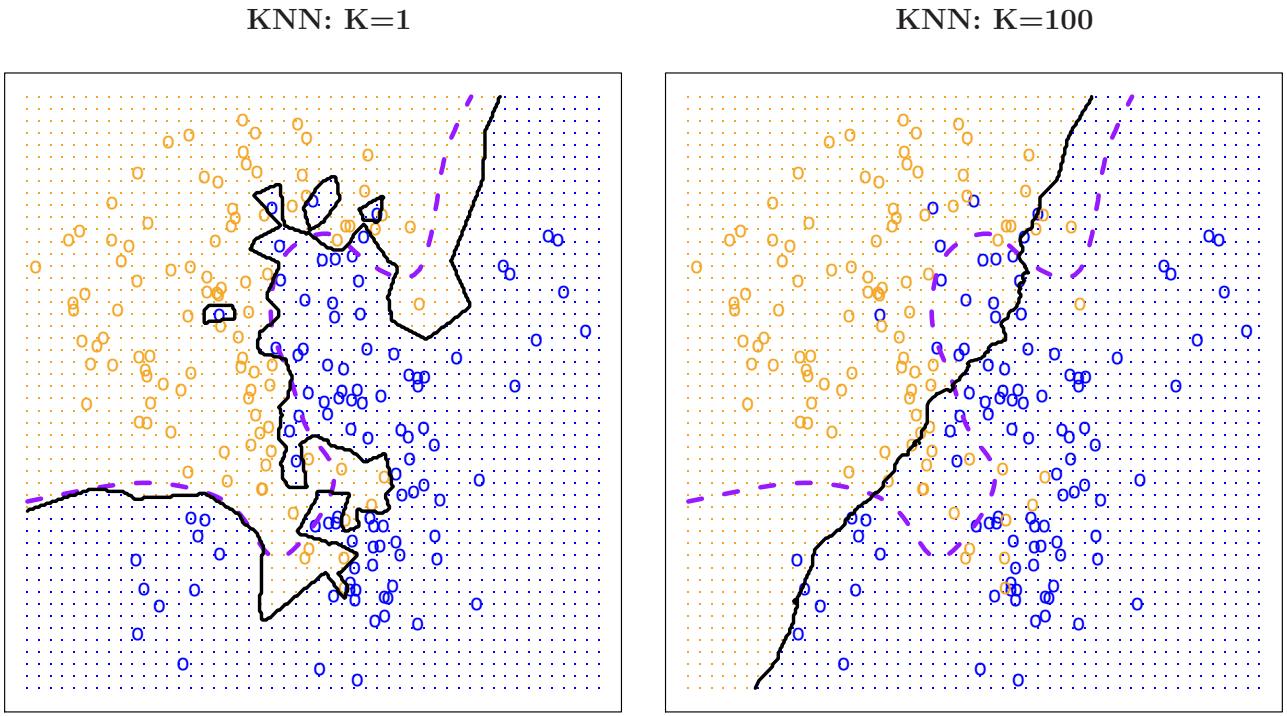
- The KNN approach, using $K = 3$
- The KNN decision boundary for this example is shown in black
- The blue grid indicates the region in which a test observation will be assigned to the blue class,
- And the orange grid indicates the region in which it will be assigned to the orange class.



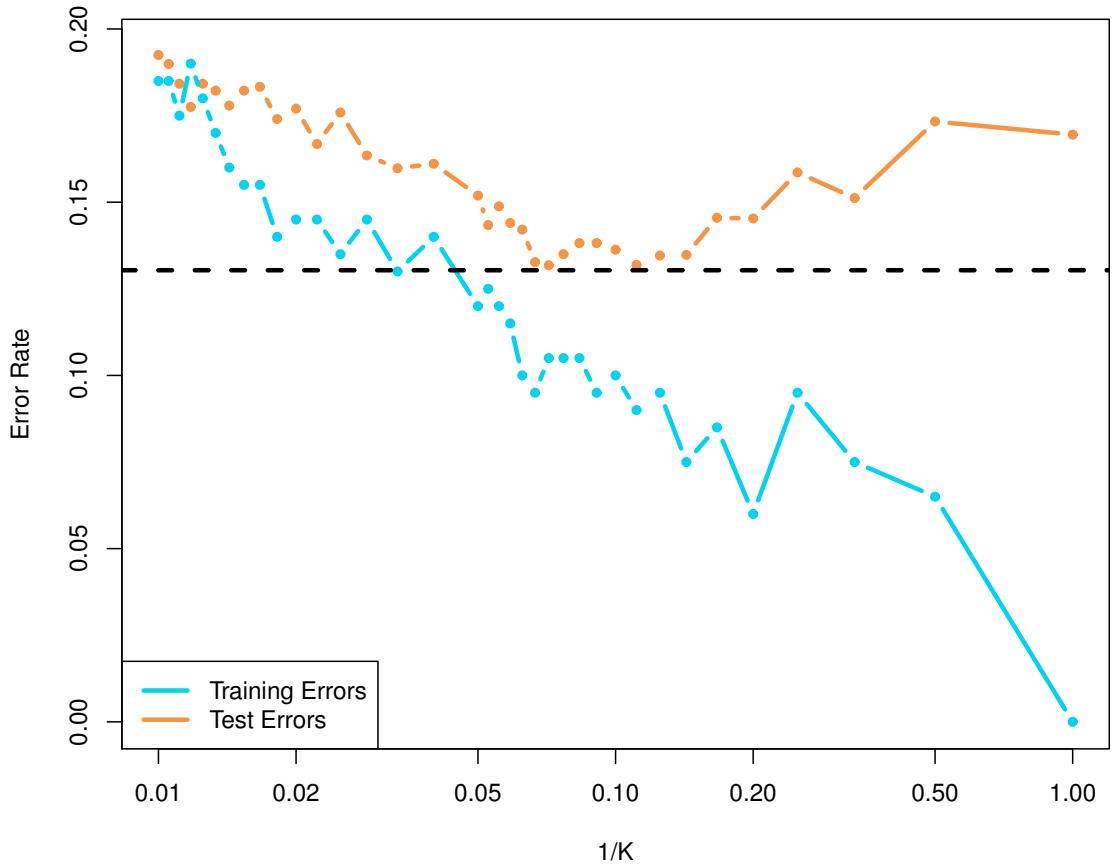
- The choice of K has a drastic effect on the KNN classifier obtained
- Purple curve for Bayes Classifier
- Black curve $K = 10$
- The KNN and Bayes decision boundaries are very similar



- A comparison of the KNN decision boundaries
- Solid black curves obtained using $K = 1$ and $K = 100$
- Purple curve for Bayes Classifier
- With $K = 1$, the decision boundary is overly flexible
- K = 100 it is not sufficiently flexible



- The KNN training error rate (blue, 200 observations)
- Test error rate (orange, 5,000 observations)
- As the level of flexibility (assessed using 1/K) increases, or equivalently as the number of neighbours K decreases
- The black dashed line indicates the Bayes error rate
- The jumpiness of the curves is due to the small size of the training data set



- 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.
- Later we return to this topic and discuss various methods for estimating test error rates
- And thereby choosing the optimal level of flexibility for a given statistical learning method.

Summary

- What we learned
  - supervised vs unsupervised learning
  - Inference vs prediction
  - Prediction accuracy and interpretability
  - bias-variance trade-off
- To do
  - **Read** Chapters 1 and 2 in "An introduction to statistical learning"
  - **Run** the Lab in Chapter 2
  - **Attempt** the exercises questions in Chapter 2, and understand them.
  - **Set up** the programming environment if you choose to use your own computer.
- Acknowledgement
  - Figures used in this slides are from the book "Introduction to Statistical Learning with Applications in R"
  - Some slides were adapted from Prof Mark John's slides on "Introduction to machine learning"