

STAT5003

Week 1: Basics of R programming & static graphics

Dr. Justin Wishart



监督学习

regression

classification

无监督学习

clustering

density estimation

在数据中如果每个自变量X都有对应的应变量Y与之对应，我们通过建立 f 的估计去理解其中的X与Y 的关系。

在数据中仅仅只有X自变量而无因变量，我们需要找出自变量之间的内在关系

<https://zhuanlan.zhihu.com/p/109125073>

Review of basic statistical concepts



THE UNIVERSITY OF
SYDNEY

Population

• **Definition:** 一组数据（数字或其他），对应于整个单位的集合，关于这些单位的信息被寻求。

- The set of data (numeric or otherwise) corresponding to the entire collection of units about which information is sought.

• **Examples:**

- Blood pressure – Blood pressure readings of ALL people in Australia.
- The number of languages spoken from ALL currently enrolled students in University of Sydney

Sample

- **Definition:** 在研究过程中实际收集的人口数据的一个子集。
利用样品sample推断人口population
 - A subset of the population data that are actually collected in the course of a study.
- **Examples:**
 - Blood pressure readings of 1000 randomly selected people in Australia.
 - The number of languages spoken from 500 randomly selected students currently enrolled in University of Sydney.

In most studies, it is difficult to obtain information about the whole population. That is why we rely on samples to make estimates and inferences related to the whole population.

Parameters vs statistics

- A **parameter** is a number that describes a population.
 - Notation usually denoted with Greek letters. e.g. μ, σ
- A **statistic** is a number that describes a sample.
 - Sample statistics are usually denoted using Roman letters, e.g. x, s .
- A parameter is a fixed number (usually unknown). A statistic is a variable whose value varies from sample to sample.

参数是一个描述人口的数字。符号通常用希腊字母表示 μ 。统计量是一个描述样本的数字。样本统计量通常用罗马字母 x 表示。一个参数是一个固定的数字（通常是未知的）。统计量是一个变量，其值在不同的样本中会有变化。

Descriptive statistics – numeric and graphics

Many methods are available for summarising data in both **numeric** and **graphical** form

Numeric measures:

- Measure of location – Mean, Median, Mode for numeric data
 - Counts, proportions for categorical data
- Measure of spread – Standard deviation, MAD (median absolute deviation), IQR
- Others: – Min, Max, Quartile, Five number summaries (used later in boxplot)

Basic statistical graphics and



THE UNIVERSITY OF
SYDNEY

Types of graphics covered in this course

- Become familiar with simple graphics using base  and ggplot graphics
- base  use the built in plotting functions
 - typically good for quick plots of simple datasets



- ggplot graphics
 - Name meaning the grammar of graphics
 - Typically better for more complicated datasets



- Not covered but honorable mention  plotly
 - Plotly is a powerful plotting library
 - Can do interactive graphics

Simple example dataframe for plots

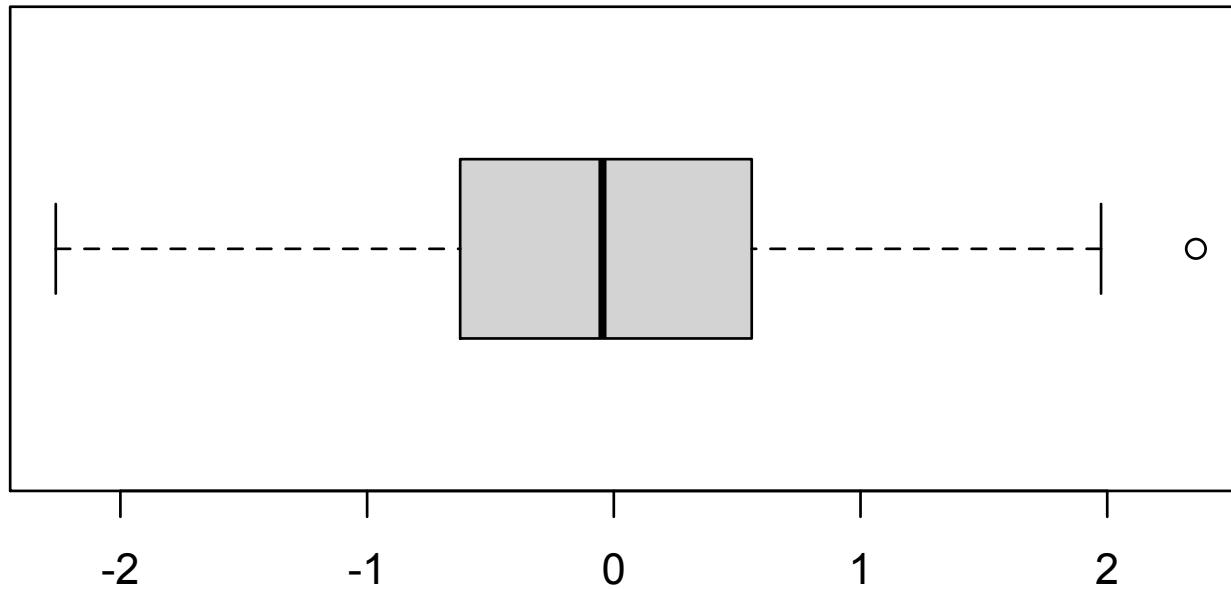
```
example.dat <- data.frame(x = rnorm(100),
                           y = runif(100),
                           cat = sample(LETTERS[1:2], prob = c(1, 3), size = 100, replace = TRUE))
head(example.dat)

##           x         y cat
## 1  0.152735495 0.96224599   B
## 2 -0.002788455 0.06619551   B
## 3  0.653621215 0.23325894   A
## 4  0.746377821 0.43503169   B
## 5 -1.058203297 0.24066605   B
## 6 -0.207491626 0.98604719   A
```

Single numeric variable: Boxplot in base R

```
boxplot(example.dat$x, horizontal = TRUE)
```

25% 50% 75%
出去的点是outlier



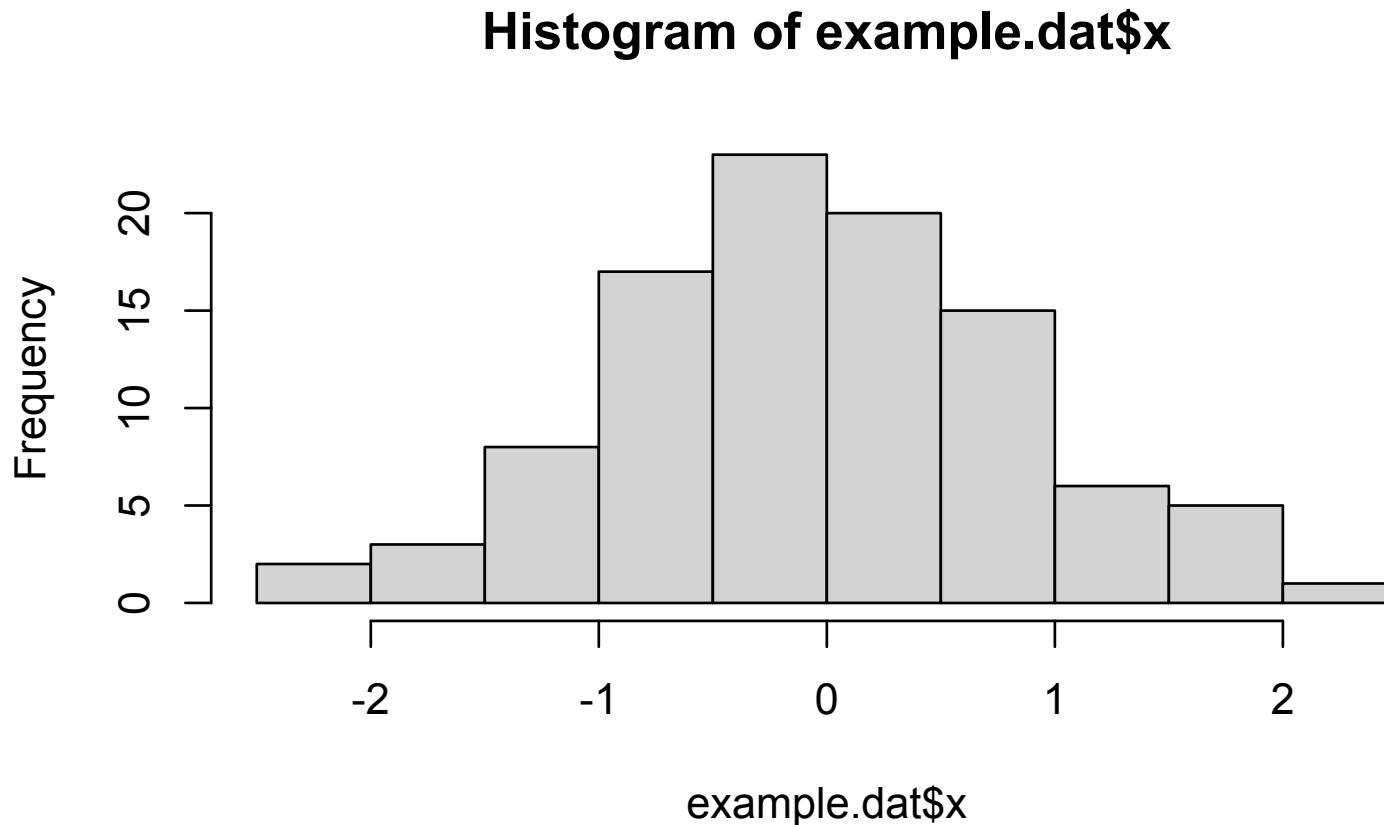
Single numeric variable: Boxplot in



```
library(ggplot2) # Only need to load the library once in an R session  
ggplot(example.dat, aes(x = x)) + geom_boxplot() + theme_minimal()
```

Single numeric variable: Histogram in base R

```
hist(example.dat$x)
```



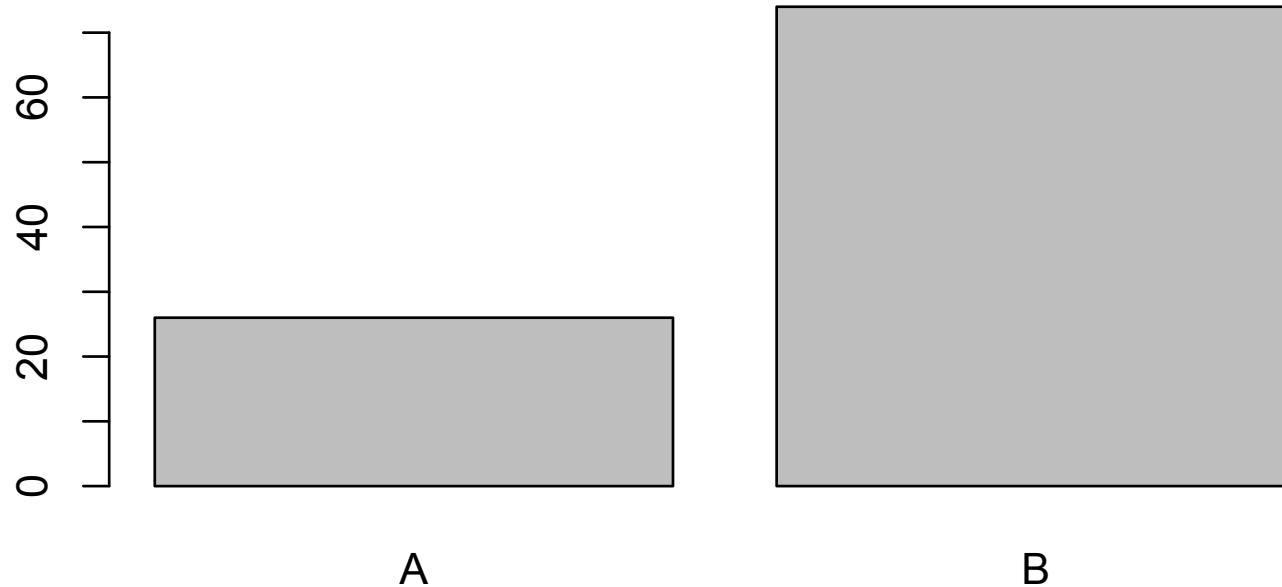
Single numeric variable: Histogram in



```
ggplot(example.dat, aes(x = x)) + geom_histogram() + theme_minimal()
```

Single categorical variable: Bar plot in base

```
barplot(table(example.dat$cat))
```



Single categorical variable: Bar plot



```
ggplot(example.dat, aes(x = cat)) + geom_bar() + theme_bw() # Change the theme
```

Two numeric variables: Scatterplot in base R

```
# These two plot commands are near equivalent
plot(y ~ x, data = example.dat, main = "formula input")
plot(example.dat$x, example.dat$y, main = "argument input")
```

Two numeric variables: Scatterplot in



```
ggplot(example.dat, aes(x = x, y = y)) + geom_point() # default theme here
```

A crash course in

What is ?

- Free, open source software designed for statistical computing
- Runs on Windows, Mac, Linux and other flavours of Unix
- Provides an interactive environment, but it is also an interpreted programming language
- Its power lies in the thousands of contributed packages on CRAN, Bioconductor and github.

Base and the



- The tidyverse popularised by Hadley Wickham and the team at  Studio
 - Has a (somewhat) standardised syntax (pipes `%>%` are king except for `+` in `ggplot2`)
 - Produces more human readable code
 - Not as stable as base, breaking changes occur as tidyverse develops.
 - Good for interactive data analyst
- Core base 
 - Good for production level code
 - Stable
 - Function syntax inconsistent

What is R Studio?

- RStudio is an integrated development environment (IDE) for
- Think of it as the front-end, like a powerful text editor
- R needs to be installed first before RStudio.



Session

- **Working Directory:** Each session runs from a working directory
 - `getwd()` shows the current working directory for R.
 - `setwd(<path>)` to change the working directory where `<path>` is a string
 - E.g. `setwd("C:/Users/usyd-student")` for a windows user.
- **Workspace:** Includes
 - Global environment: Data and variables loaded or defined
 - package environments: Any loaded packages with their functions/data.
- **History:**
 - Can view your recent commands in the History pane
 - Can navigate previous commands using up and down arrows at your prompt

Packages

- Inspect the current environment
 - `sessionInfo()`: shows everything in the current R session.
- To install a new package,
 - `install.packages("cluster")` will install the `cluster` package from the Comprehensive R Archive Network (CRAN) repository
- To load a package to use in the environment `library(cluster)`
 - After loading a package, you will be able to use the functions provided in that package
- If name conflicts arise you can specify the desired function using `::`
 - E.g. `dplyr::filter` will use the `filter` function from the `dplyr` package instead of the `stats::filter` function.

Help

- CRAN requires every exported function in every contributed package to have a help file
 - `help.start()`
 - `help(plot)`
 - `? plot`
- In general, the help file for each function gives a brief description of what the function does, the required inputs, the expected outputs and some examples
 - Some packages also include a "vignette", a short document with guidance on using the package.

Quirks of R syntax

- <- is the symbol for 'assign'
 - Example: x <- 14
 - which is equivalent to: x = 14 when used at the prompt
- Should use = for argument matching in a function
- The period symbol . can be used in variable names
 - Example: new.vector <- c("A", "B", "C")
- Element indexing starts at 1

```
new.vector <- c("A", "B", "C")
new.vector[0]
```

```
## character(0)
```

```
new.vector[1]
```

```
## [1] "A"
```

Basic data types in R

Classical data types

- Numeric
- Integer
- Logical
- Character
- Complex

- Factor : Categorical data type
 - Unique to R (integer with some attributes)

```
data("ToothGrowth")
levels(ToothGrowth$supp)
```

```
## [1] "OJ"  "VC"
```

```
class(ToothGrowth$supp)
```

```
## [1] "factor"
```

```
str(ToothGrowth$supp)
```

```
##  Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
```

Homogeneous vs non-homogenous data types in R

Homogeneous

- Vector 具有相同基本数据类型的数据元素的序列。【2 -8 7】
 - Sequence of data elements of the same basic data type
- Matrix 在一个有行和列的二维数组中的数据元素的集合 row x column
 - Collection of data elements in a 2-dimensional array with rows and columns

Non-homogeneous

- List 包含其他对象（可能包括其他列表）的更一般的结构（字符，数字）
 - More general structure containing other objects (including possibly other lists)
- Data frame 用于存储数据，每一列可以是不同的基本类型.所有列必须有相同的长度
 - Used for storing data, each column can be a different basic type
 - All columns must have the same length

Vectors

```
new.vector <- c(1, 2, 3)
class(new.vector)
```

```
## [1] "numeric"
```

```
length(new.vector)
```

```
## [1] 3
```

```
new.vector[1:2]
```

```
## [1] 1 2
```

```
new.vector <- c(1, 2, "hello")
class(new.vector)
```

```
## [1] "character"
```

Matrix

```
A <- matrix(c(2, 4, 3, 1, 7, 8), nrow = 3)
# Unless specified otherwise, it will fill the matrix by column.
A
```

```
##      [,1] [,2]
## [1,]     2     1
## [2,]     4     7
## [3,]     3     8
```

```
A[2, 1]
```

```
## [1] 4
```

```
A[1, ]
```

```
## [1] 2 1
```

```
A[5]
```

```
## [1] 7
```

List

```
vector.a <- c(1, 2, 3)
vector.b <- c("hello", "world", "!!!")
new.list <- list(c(vector.a, vector.b))
new.list

## [[1]]
## [1] "1"      "2"      "3"      "hello"  "world"  "!!!"
```

```
new.list <- list(vector.a, vector.b)
new.list
```

```
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [1] "hello" "world" "!!!"
```

```
new.list[[1]]
```

```
## [1] 1 2 3
```

Data frames

```
head(warpbreaks)
```

```
##   breaks wool tension
## 1     26    A      L
## 2     30    A      L
## 3     54    A      L
## 4     25    A      L
## 5     70    A      L
## 6     52    A      L
```

```
class(warpbreaks)
```

```
## [1] "data.frame"
```

```
head(warpbreaks$wool)
```

```
## [1] A A A A A A
## Levels: A B
```

```
str(warpbreaks)
```

```
## 'data.frame': 54 obs. of 3 variables:
##   $ breaks : num 26 30 54 25 70 52 51 26 67 18 ...
##   $ wool   : Factor w/ 2 levels "A","B": 1 1 1 1 1
##   $ tension: Factor w/ 3 levels "L","M","H": 1 1 1
```

```
names(warpbreaks)
```

```
## [1] "breaks"  "wool"     "tension"
```

Rprojects

- Each Rstudio project has its own working R session, workspace, history, and source documents
- You can either create an Rproject in a new directory or within and existing directory
- When you open an .Rproj file, the following happens:
 - A new R session (process) is started
 - If it exists, the .Rprofile in the project's main directory is loaded
 - If applicable, the .RData file in the main directory is loaded
 - If set, the .Rhistory file in the main directory is loaded
 - The current working directory is set to the project directory
 - If set, previously edited documents appear in the source editor.

STAT5003

Week 2: Regression and Smoothing

Dr. Justin Wishart



THE UNIVERSITY OF
SYDNEY



Readings



- Introduction to Statistical Learning James, Witten, Hastie, and Tibshirani (2013)
 - Chapter 3 (Linear regression)
 - Chapter 7.4 to 7.6 (Smoothing)

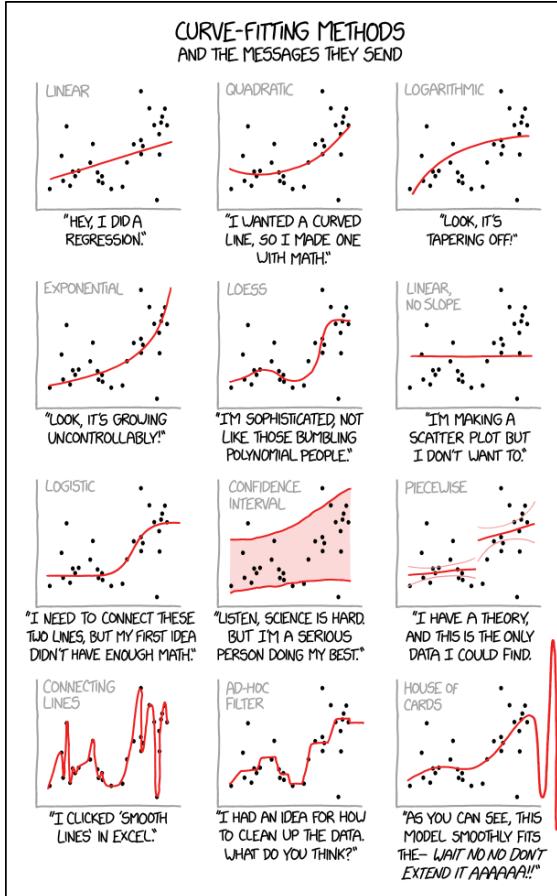
Linear Regression



THE UNIVERSITY OF
SYDNEY

Regression

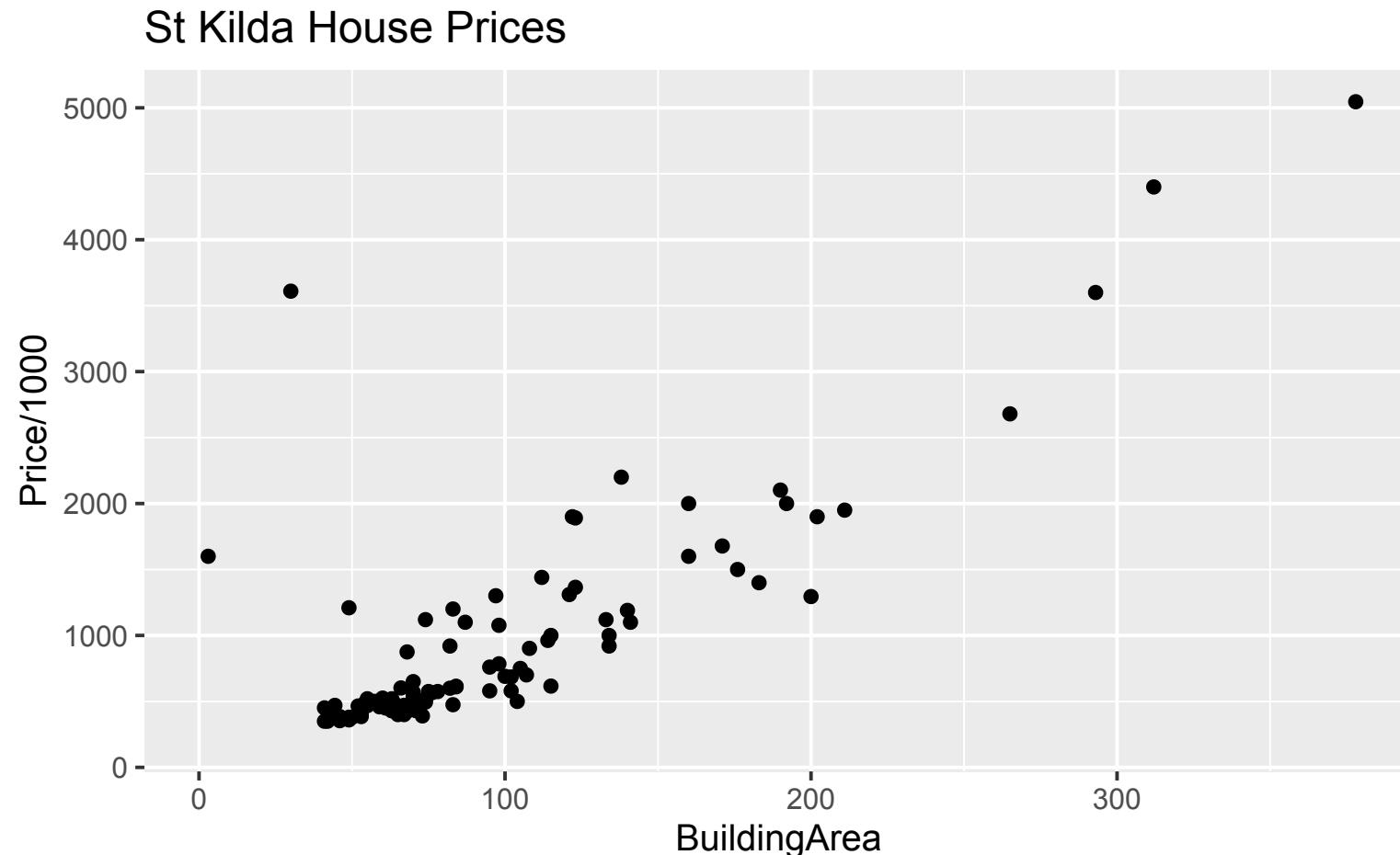
- Numerically fitting the model is easy



- Knowing how to appropriately fit the model is where you add value.

The prediction problem

What is the price of a 100 sqm house in St Kilda?



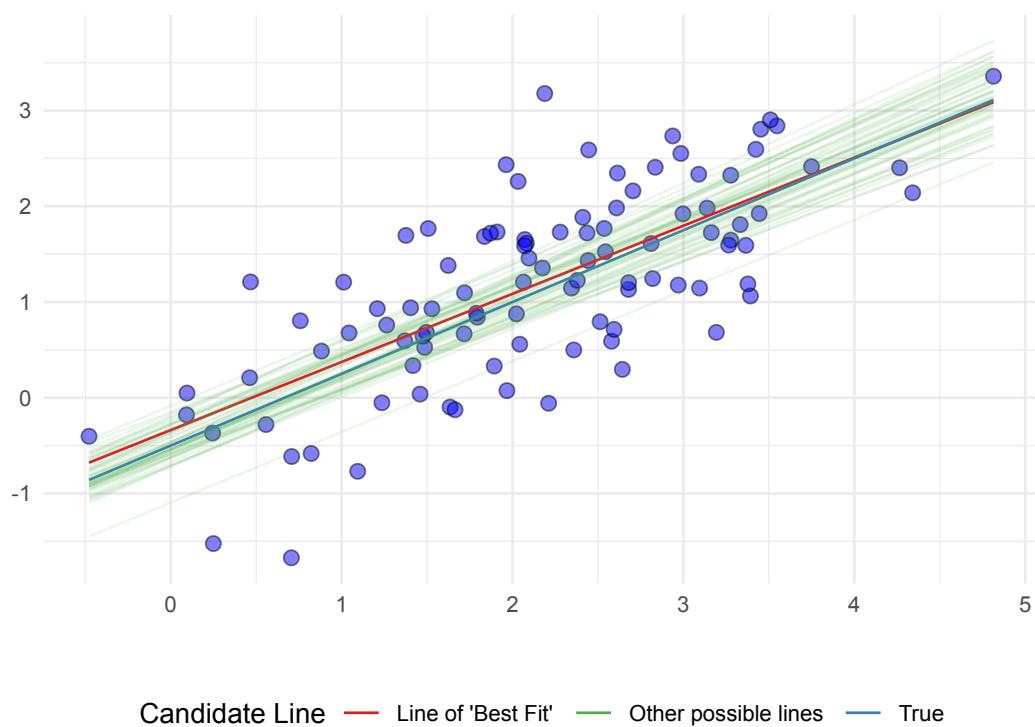
The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\downarrow$$
$$y_i = b_0 + b_1 x_i + e_i$$

- X is the **predictor** (feature or independent variable) 特征或自变量
- Y is the **response** (target or dependent variable) 目标或因变量
- β_0 is the **intercept** of the regression line 截距
 - Expected value of Y when $X = 0$
- β_1 is the **slope** of the regression line 斜率
 - mean increase in Y for a *unit* increase in X
- ε is the **unexplained variation** or random error.
 - Classically assumed to be normally distributed with mean zero and finite variance.

ε 是无法解释unexplained variation的变化或随机误差。经典的假设是正态分布，均值为零，方差为零

Performance of regression estimates

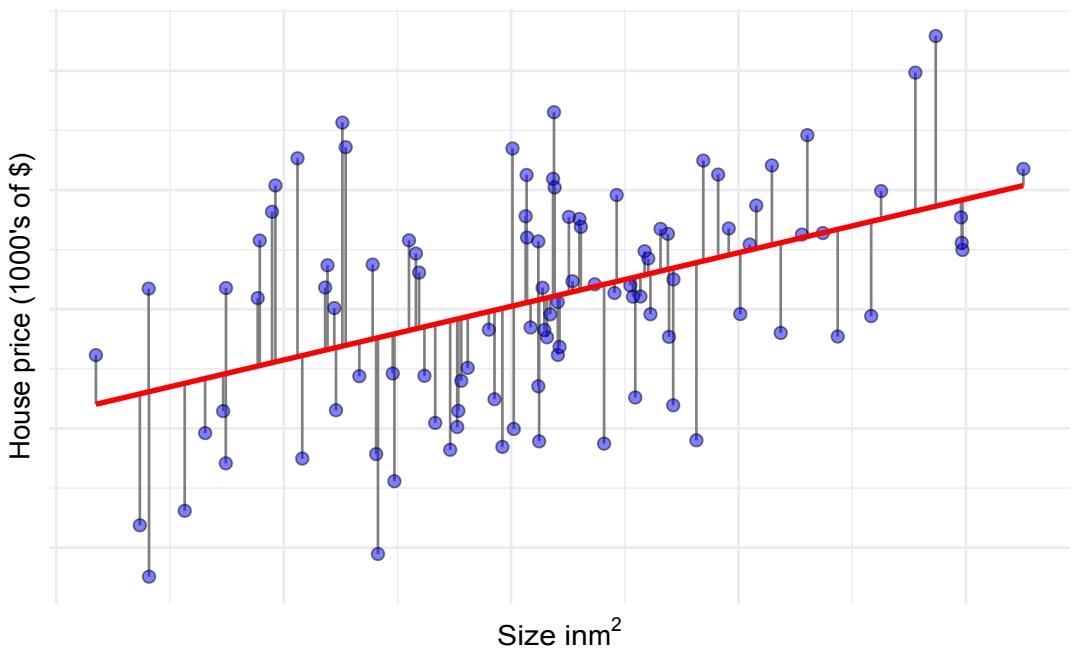


- Data was simulated from model
$$Y = -0.5 + 0.75X + \varepsilon$$
- True line shown in blue
- Standard linear regression fit shown in red
- Why not one of the green lines?

因为有偏差

How to determine the best estimates of $\beta_0 + \beta_1 X$?

- The notion of best needs a **criterion** to measure against. 最小的ss为最好的estimate



- Easiest mathematical solution is the **least squares criterion** 最小二乘法标准

- Minimise the residual sum of squares

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

Rss 是 residuals

DF * residual standard error 的平方

用anova function, 然后找到residuals of sum, 就是RSS

Least squares equations

- Can show by simple calculus the following:

- Regression (slope) coefficient: $b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$

- Intercept: $b_0 = \bar{y} - b_1 \bar{x}$ $Cov(X, Y) = E(XY) - E(X)E(Y)$

- This leads to the estimated regression line: $Var(X) = E(X^2) - E(X)^2$

$$\hat{y} = b_0 + b_1 x$$

- Least squares regression line since it minimises the residual sum of squares.

Prediction using `lm`

```
lm.fit <- lm(Price ~ BuildingArea, data = st.kilda.data)
summary(lm.fit)

## 
## Call:
## lm(formula = Price ~ BuildingArea, data = st.kilda.data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -817415 -201614 -85181   19895  3403199 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -129484.0    91775.9  -1.411   0.161    
## BuildingArea   11209.5     799.8   14.015  <2e-16 ***
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 490300 on 99 degrees of freedom
## Multiple R-squared:  0.6649,    Adjusted R-squared:  0.6615 
## F-statistic: 196.4 on 1 and 99 DF,  p-value: < 2.2e-16
```

Standard error of population mean

标准误差

- Consider single population estimation problem .
 - Wish to estimate some mean, μ , of some random variable Y .
 - If Y_i is sampled then $\hat{\mu} = \bar{Y}$ estimates μ with

SE 是 output 的 standard error

$$Var(\hat{\mu}) = (SE(\hat{\mu}))^2 = \frac{\sigma^2}{n}$$

- σ^2 is the variance of Y_i Variance = (y-yi)的平方相加 再除于n
- n is the sample size.

Standard error of regression coefficient estimates

- Same concept applies to the regression estimates

$$SE(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$
$$SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

where $\sigma^2 = Var(\varepsilon)$

- As $n \rightarrow \infty$, $SE(\hat{\beta}_0) \rightarrow 0$ and $SE(\hat{\beta}_1) \rightarrow 0$
- Interestingly, if the x_i are more spread out, the standard errors will be smaller
 - more leverage to estimate the parameters.

如果这些参数更加分散，标准误差会更小。估计参数的
杠杆作用更大

Using standard errors to compute confidence intervals

```
summary(lm.fit) # Truncated output with coefficient table
```

```
...
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -129484.0    91775.9  -1.411   0.161
## BuildingArea  11209.5     799.8   14.015  <2e-16 ***
## ---
## 
## Residual standard error: 490300 on 99 degrees of freedom
...
```

- We can use the standard error to estimate the 95% confidence interval as:
 - $(\hat{\beta}_1 - t_{n-2,0.975} SE(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,0.975} SE(\hat{\beta}_1)) = b_1 \pm t_{n-2,0.975} SE(b_1) = b_1 \pm t_{99,0.975} SE(b_1)$
- In our housing example, the 95% confidence interval for the coefficient of BuildingArea is [9622.6968, 12796.3032]

$$b_1 \pm t_{n-2,0.975} SE(b_1) = 1.12095 \times 10^4 \pm 1.984 \times 799.8 = (9622.6968, 12796.3032)$$

$$95\% == 0.975(1-(1-0.95)/2) \ qt(0.975, df=99)$$

Confidence intervals of regression coefficients

- More directly in  code.
 - Use the `confint` function.

```
confint(lm.fit)
```

```
##              2.5 %    97.5 %
## (Intercept) -311587.233 52619.18
## BuildingArea   9622.491 12796.50
```

- This is exact and no precision lost to rounding error.
- Easy to change confidence level (99% below)

```
confint(lm.fit, level = 0.99)
```

```
##              0.5 %    99.5 %
## (Intercept) -370524.63 111556.57
## BuildingArea   9108.86 13310.13
```

Is BuildingArea a good predictor of price?

```
summary(lm.fit) # truncated for coefficient table
```

```
...
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -129484.0   91775.9  -1.411   0.161    
## BuildingArea   11209.5     799.8   14.015  <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

- Linear regression assumes $Y = \beta_0 + \beta_1 X + \varepsilon$
- If BuildingArea is not linearly related to Price, then $\beta_1 = 0$.
- Can conduct a test of significance $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$
- Can conduct a hypothesis test by computing the t -statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \stackrel{H_0}{=} \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

判定是否predictor 看最后p-value 小于0.05 说明有关联
R-square 越高， 越拟合
 $Tvalue$ 大于统计值， 该系数显著， 需要保留

Is BuildingArea a good predictor of price?

```
summary(lm.fit) # truncated for coefficient table
```

```
...
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -129484.0    91775.9  -1.411   0.161
## BuildingArea   11209.5     799.8   14.015  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

- The **p-values** for each significance test in the last column.
- Recall, **p-value** gives the probability of observing your test statistic (and other scenarios support H_1) assuming H_0 is true.
- Small p-value here gives very little evidence to support the claim that there is no relationship between Price and BuildingArea

Goodness of fit statistic

- Goodness of fit is measured by the coefficient of determination or R^2

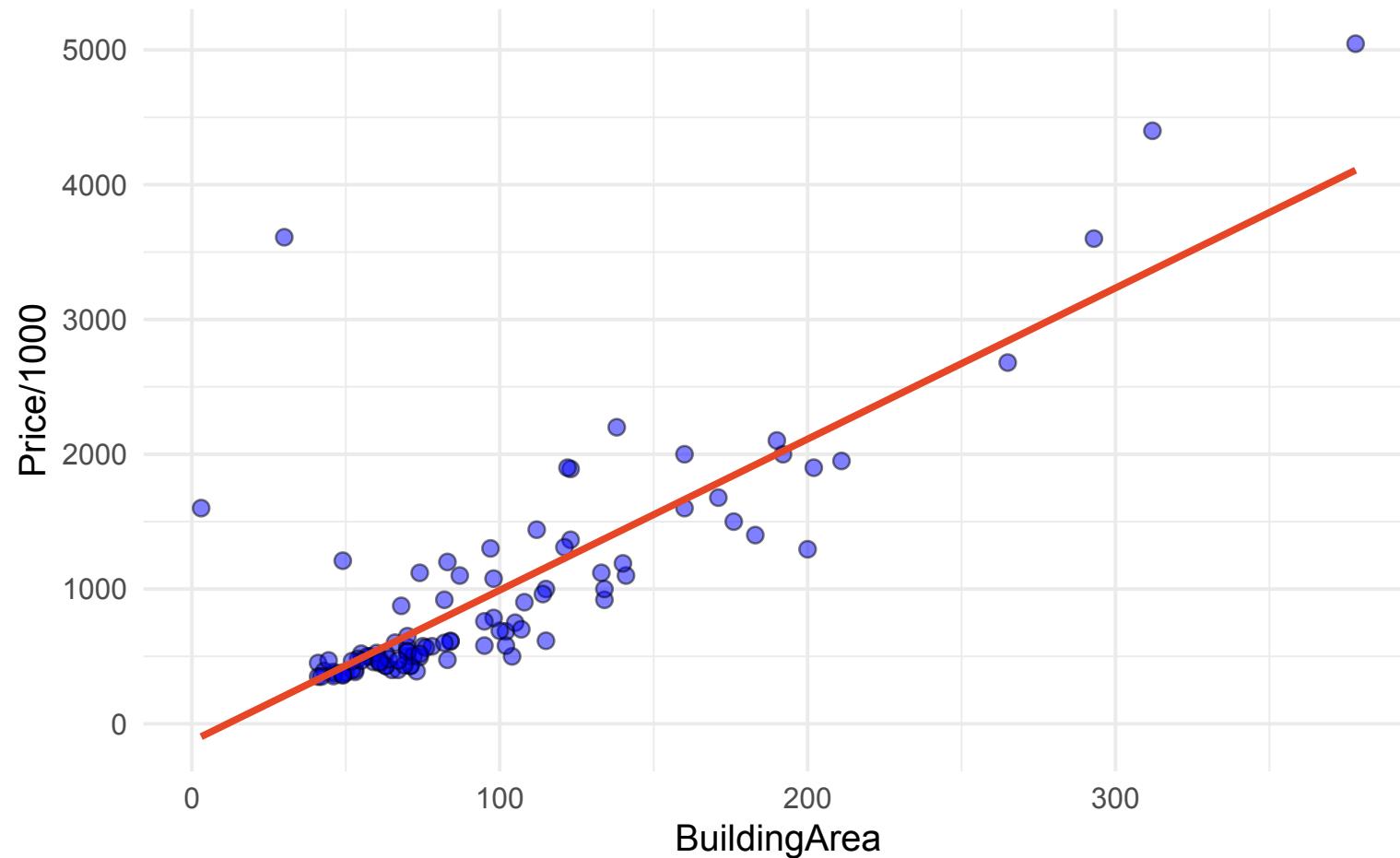
$$\begin{aligned} R^2 &= \frac{\text{Total Sum of Squares} - \text{Residual Sum of Squares}}{\text{Total Sum of Squares}} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

第一个y是实际值，第二个是平均值，第三个是预测线上的值

- R^2 is a measure between 0 and 1
- It measures the proportion of variation in the response Y , explained by the linear regression on X
 - A value of 0 indicates **none** of the variance in Y can be explained linearly by X
 - A value of 1 indicates **all** of the variance in Y can be explained linearly by X

R^2 的值在0-1中间，衡量响应Y的变化比例，由对X的线性回归解释。值为0表示Y的方差没有一个可以用X来线性解释。值为1表示Y的所有方差都可以用来线性解释

Linear regression fit



Estimating the price of a 100 m^2 house in St Kilda

```
new.100 <- data.frame(BuildingArea = 100)
predict(lm.fit, new.100, interval = "confidence")
```

```
##          fit      lwr      upr
## 1 991465.5 894562.7 1088368
```

```
predict(lm.fit, new.100, interval = "prediction")
```

```
##          fit      lwr      upr
## 1 991465.5 13820.26 1969111
```

置信区间估计(confidence interval estimate): 利用估计的回归方程, 对于自变量 x 的一个给定值 x_0 , 求出因变量 y 的平均值的估计区间。置信区间是以样本去估算总体

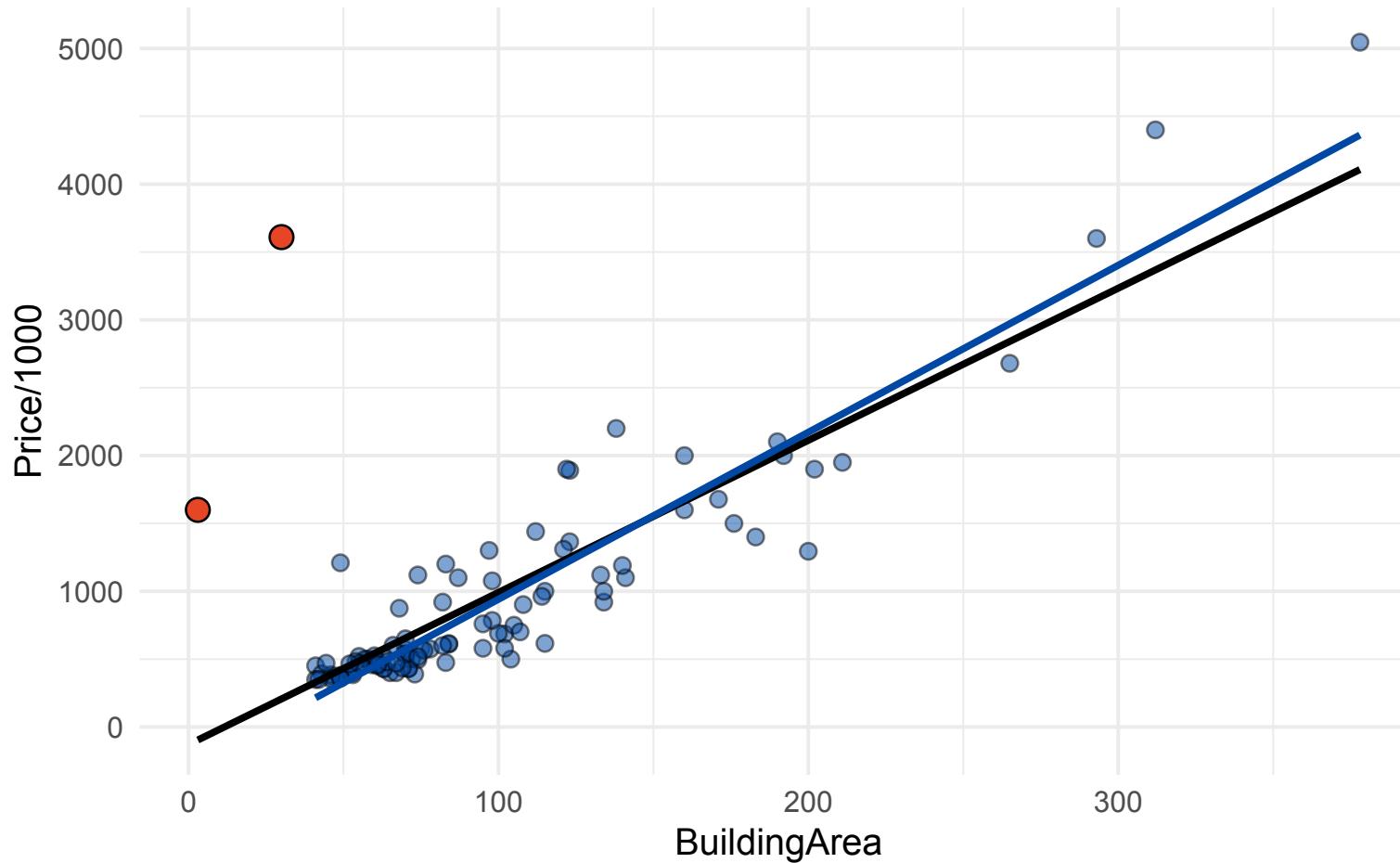
预测区间估计(prediction interval estimate): 利用估计的回归方程, 对于自变量 x 的一个给定值 x_0 , 求出因变量 y 的一个个别值的估计区间。预测区间是指个体的区间

预测区间的范围总是要比置信区间的范围要大的。个别值更容易受一些外界因素影响而有差异性, 而平均值则相对稳定些

Fit improvements

提高拟合度 --- 除去异常值outliers

- Remove outliers: black line gives overall fit, blue line fit only to blue data (without red points)



Linear fit after removing the outliers

```
lm.without.outliers <- lm(Price/1000 ~ BuildingArea, data = st.kilda.data, subset = BuildingArea >= 40)
summary(lm.without.outliers)
```

```
## 
## Call:
## lm(formula = Price/1000 ~ BuildingArea, data = st.kilda.data,
##     subset = BuildingArea >= 40)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -876.75  -137.30   -18.27  109.28  896.31 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -289.254    57.471  -5.033 2.22e-06 ***
## BuildingArea   12.305     0.496   24.807 < 2e-16 ***
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 298.5 on 97 degrees of freedom
## Multiple R-squared:  0.8638,    Adjusted R-squared:  0.8624 
## F-statistic: 615.4 on 1 and 97 DF,  p-value: < 2.2e-16
```

R formulae

- Example formula `Response ~ Predictor1 + Predictor2 + Predictor3`
- Left hand side of `~` is the `response` variable (target to predict)
- Right hand side of `~` are the the `predictor` variables (features)
- Relationship is assumed to be additive
 - I.e. each additional predictor is added to explain the response $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$
- Interaction or multiplicative terms are denoted with `:` and `*` which are beyond the scope fo this course.
 - Would be used to define other relationships
 - E.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_2 + \dots$

Multiple linear regression

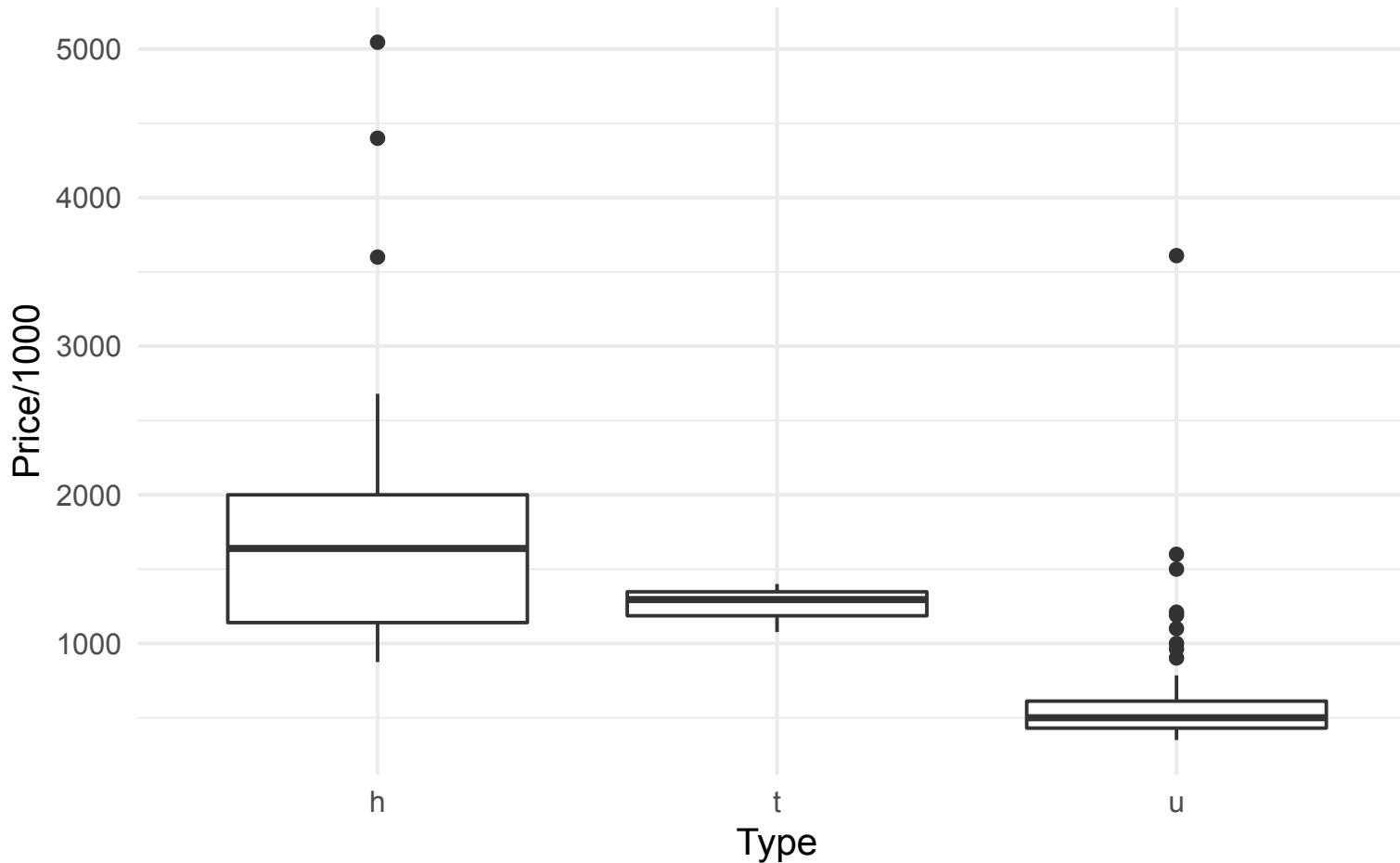
- Real life problems usually have more than one predictor.
 - Simple linear (single variable) regression can be extended to multiple predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_p X_p + \varepsilon$$

- The interpretation is the β_p coefficient denotes the average increase/decrease in Y for each single unit increase in X_p , holding all the other predictors fixed.

Extending the house prediction model to multiple features

- Perhaps 100 m^2 houses cost more than 100 m^2 units?



Multiple regression with `lm`

```
multi.lm <- lm(Price/1000 ~ Type + BuildingArea, data = st.kilda.data)
summary(multi.lm)
```

```
## 
## Call:
## lm(formula = Price/1000 ~ Type + BuildingArea, data = st.kilda.data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -700.3 -173.1  -65.9   18.6 3389.6 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 342.865   186.764   1.836  0.06945 .  
## Typet      -613.953   286.272  -2.145  0.03448 *  
## Typeu      -408.417   139.915  -2.919  0.00436 ** 
## BuildingArea  9.533     1.014   9.398 2.68e-15 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 469.5 on 97 degrees of freedom
## Multiple R-squared:  0.6989,    Adjusted R-squared:  0.6896 
## F-statistic: 75.06 on 3 and 97 DF,  p-value: < 2.2e-16
```

Model interpretation

```
summary(multi.lm)
```

```
...
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  342.865   186.764   1.836  0.06945 .
## Typet       -613.953   286.272  -2.145  0.03448 *
## Typeu       -408.417   139.915  -2.919  0.00436 **
## BuildingArea    9.533     1.014    9.398 2.68e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...

```

```
multi.pred.data <- data.frame(BuildingArea = rep(100, 3), Type = c("u", "t", "h"))
predict(multi.lm, newdata = multi.pred.data)
```

```
##          1          2          3
## 887.7252 682.1894 1296.1427
```

Nonparametric regression or Smoothing



THE UNIVERSITY OF
SYDNEY

Parametrics vs non-parametric methods

- Parametric methods involve selecting a statistical model (e.g. linear regression model) and fitting the parameters of the model (e.g. slope, intercept) using the training data
- Nonparametric methods don't require selecting a strict model. The data is allowed to *speak for itself*. However, don't have easily interpretable parameters. They are generally intended for description rather than formal inference (e.g. k -nearest neighbor smoothing)

parametric参数化方法包括选择一个统计模型（如线性回归模型），并利用训练数据确定模型的参数（如斜率、截距）。

Nonparametric非参数方法不需要选择一个严格的模型。数据被允许为自己说话。然而，没有容易解释的参数。它们通常是为了描述。而不是正式推断（例如，K-近邻平滑KNN）。

Data smoothing

With **predictor-response** data, the random response variable Y is assumed to be a **non-linear** function of the predictor variable X .

$$Y_i = f(X_i) + \varepsilon_i$$

- f is some fixed, non-linear smooth function.
- ε_i is a zero-mean random variable.
- Smoothing is a **non-parametric method to estimate** f .

Local averaging

大多数平滑器（平滑函数）依赖于局部平均的概念。相比之下，简单的线性回归试图拟合出最佳的全局线

- Most smoothers (smoothing functions) rely on the concept of *local averaging*
 - In contrast, simple linear regression attempts to fit the best global line.
- E.g. Suppose you want to determine the response Y conditional on x .
 - The Y_i whose corresponding x_i are near x should be averaged with higher weight to attempt to estimate $f(x)$.
- A generic local-averaging smoother can be written as

$$\hat{f}(x) = \text{average}(Y_i | x_i \in N(x))$$

- where average is some generalised averaging operation.
- $N(x)$ is some neighbourhood of x .

Constant-Span Running Mean, k -nearest neighbours

- A simple smoother takes the sample mean of k nearby points
- We define $N(x_i)$ as x_i itself, the $(k - 1)/2$ points whose predictor values are nearest below x_i , and the $(k - 1)/2$ points whose predictor values are nearest above x_i
- This neighbourhood is termed the *symmetric nearest neighbourhood*, and the smoother is called a **moving average** or a **k -nearest neighbours (kNN) smoother**.
- The constant-span running-mean smoother can be written as:

$$\hat{f}(x_i) = \text{mean} \left[Y_j \text{ such that } \max \left(i - \frac{k-1}{2}, 1 \right) \leq j \leq \min \left(i + \frac{k-1}{2}, n \right) \right]$$

Regression splines

- Fit piece-wise functions, where each function can be a d -dimensional polynomial function
- Constrain the function to be smooth and continuous
- Cubic spline fits cubic polynomial functions, with the constraints:
 - continuous at each knot, continuity of the 1st derivative and continuity of the 2nd derivative
- Advantage of the cubic spline is that the curve looks smooth to the eye, and can be used to fit almost any function

拟合分片函数，其中每个函数可以是d维的多项式函数.约束函数是平滑和连续的.

立体花键拟合立体多项式函数，约束条件是。每个结点都是连续的，第一次导数的连续性和第二次导数的连续性

立体花键cubic spline的优点是曲线看起来很平滑，而且可以用来处理几乎所有的函数划分多个区间

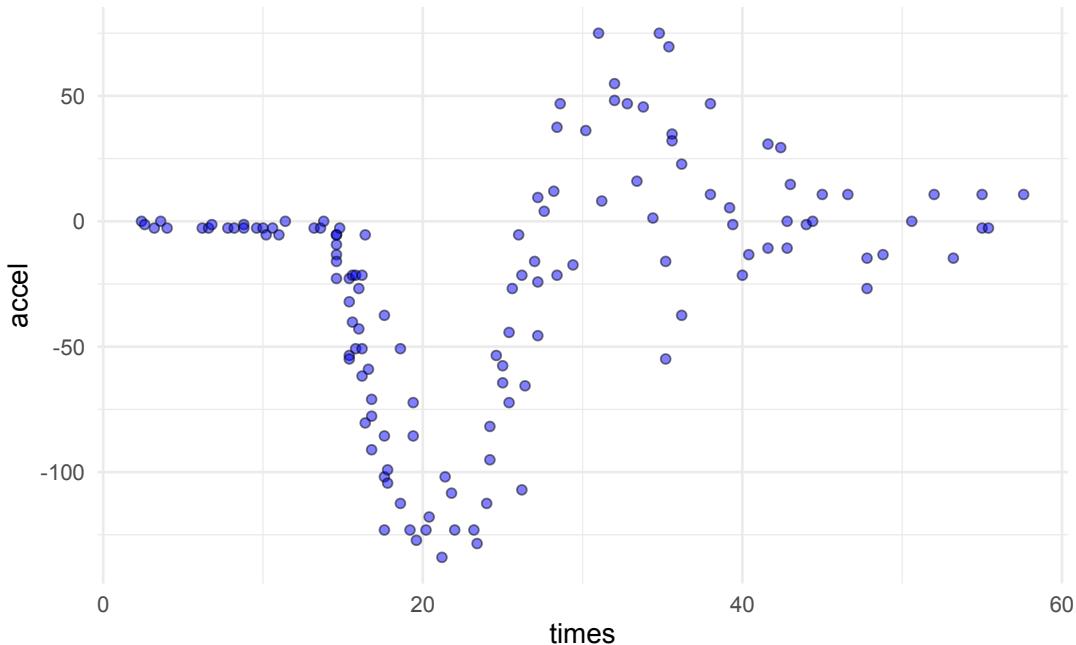
Loess

- Loess is a Locally weighted scatterplot smoothing method
- The loess (**L**Ocal regr**E**SSion) smoother is a widely used method with good robustness properties.
- It is essentially a weighted running-line smoother, except that each local line is fitted using a robust method rather than least squares.
- As a result, the smoother is nonlinear.
- Loess is computationally intensive and require densely sampled data.

一种局部加权的散点图平滑方法,是一种广泛使用的方法，具有良好的稳健性。
它本质上是一种加权流水线平滑法，只是每个局部的流水线是用一种稳健的方法而不是最小二乘法来计算的。因此，该平滑器是非线性的。loess是计算密集型的，需要密集的采样数据。
最近的点权重为0，最近的点权重为1

Local regression

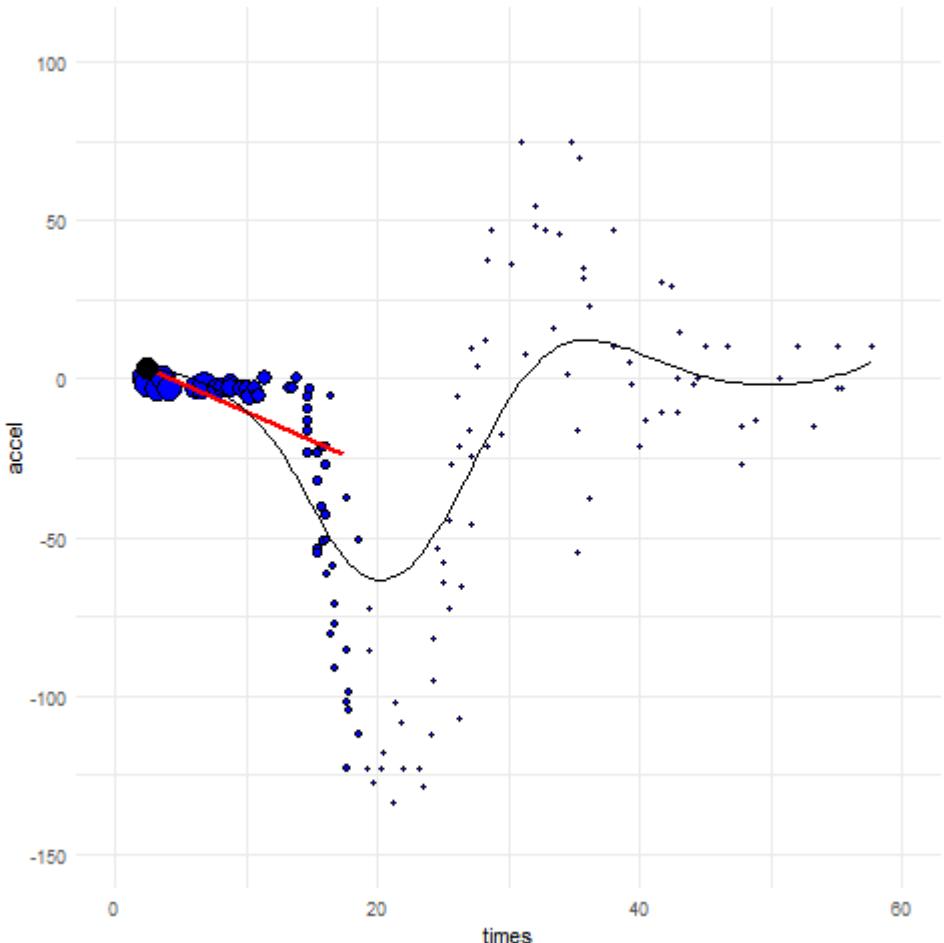
- Fitting local linear fits that are weighted.
- Formula for local constant fit is $\hat{f}(x) = \frac{\sum_{i=1}^n Y_i K((X_i-x)/h)}{\sum_{i=1}^n K((X_i-x)/h)}$



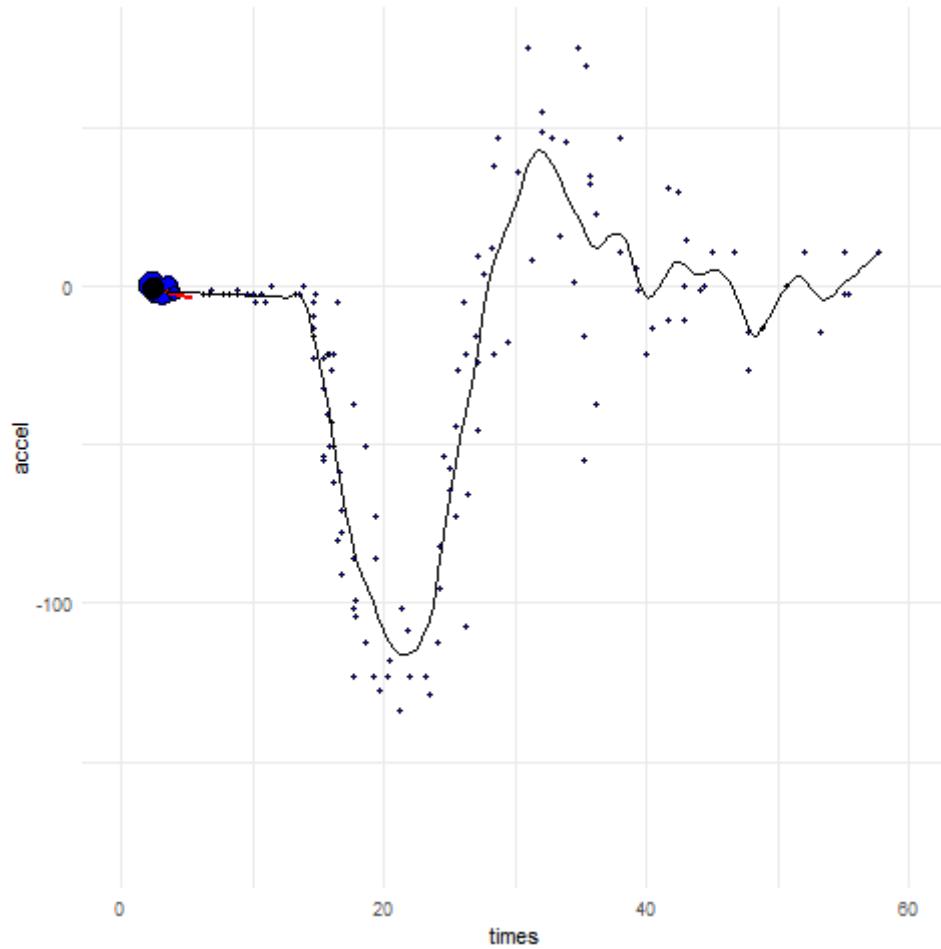
- Sharp changes in behaviour from the acceleration on a crash test dummy.

Local regression animation

- Using a large averaging window



- Using a smaller averaging window



Nonparametric smoothing vs linear regression

- Advantages of non-parametric smoothing
 - Can model non-linear functions (e.g. splines, loess)
 - Does not make any assumption about the functional form of the data
- Advantages of linear regression
 - Computationally efficient, even for multivariate linear regression
 - Model is interpretable, i.e. one can know the statistical meaning of the estimated slope parameters.

非参数平滑法的优点

可以对非线性函数进行建模（例如，splines, loess）。

对数据的函数形式不做任何假设

线性回归的优点

计算效率高，甚至对多变量线性回归也是如此

模型是可解释的，即人们可以知道估计的斜率参数的统计意义。

Reference list

James, G, D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

STAT5003

Week 3 : Density Estimation

Dr. Justin Wishart



THE UNIVERSITY OF
SYDNEY



Readings



- For the bias variance tradeoff see Section 2.2 James, Witten, Hastie, and Tibshirani (2013)

Review on probability distribution functions



THE UNIVERSITY OF
SYDNEY

Discrete distributions 离散分布

For any random variable X with a discrete distribution, there is a sample space Ω with finite number of possible values (outcomes) $x = \{x_1, x_2, \dots\}$ and associated probabilities $\{p_1, p_2, \dots\}$.

The point probabilities for each value of x are denoted $f(x)$ and the cumulative distribution function denoted $F(x)$ where

$$f(x) = P(X = x), \quad F(x) = P(X \leq x) \quad (\text{dnorm})$$

Properties:

- There is a *countable* number of possible values;
- $\sum_{i=1}^{\infty} p_i = 1$
- $p_i \geq 0$

概率密度的总体形状被称为概率分布 (probability distribution)，常见的概率分布有均匀分布、正态分布、指数分布等名称。对随机变量特定结果的概率计算是通过概率密度函数来完成的，简称为 PDF (Probability Dense Function)

Binomial distribution (dbinom)

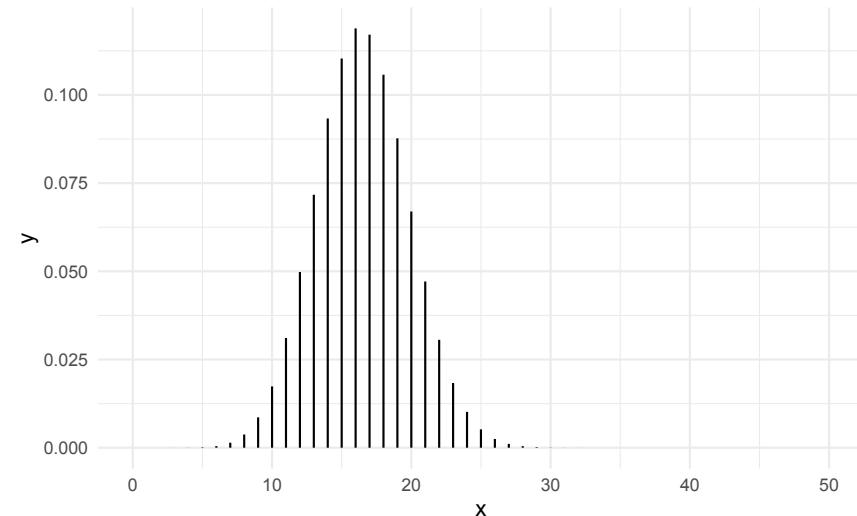
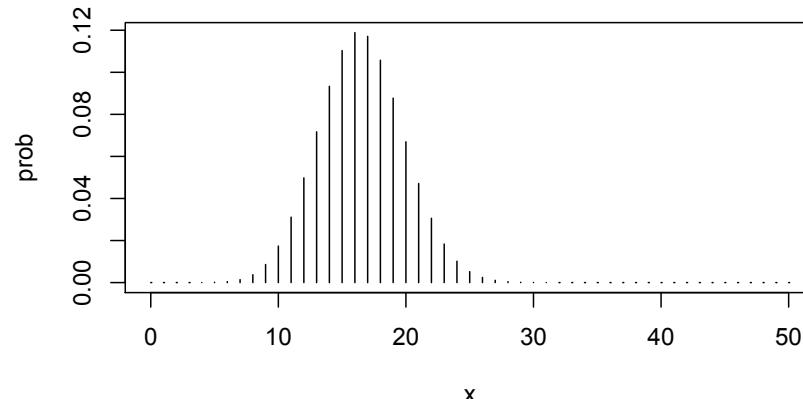
二项分布

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

The $\binom{n}{x}$ are known as the binomial coefficients.

The parameter p is the probability of success.

```
x <- 0:50
prob <- dbinom(x, size = 50, prob = 0.33)
# Base R graphics
plot(x, prob, type = "h")
dat <- data.frame(x = x, y = prob)
# ggplot2 version
ggplot(dat, aes(x = x, y = y, xend = x, yend = 0))
  geom_segment() + theme_minimal()
```



Continuous distributions

连续分布

- A continuous random variable X is where the outcome can take an infinite (uncountable) number of possible values.
 - These values may be within a fixed or unbounded interval.
- For example, the height of male in cm may be within the range of [50, 300].

The point probabilities for each value of x is $P(X = x) = 0$ and the cumulative distribution function

$$F(x) = \int_{-\infty}^x f(t) dt = P(X \leq x)$$

Properties:

- There are an infinite (uncountable) number of possible values;
- $f(x)$ is called the density function
- $f(x) \geq 0$ (non-negative)
- $\int_{-\infty}^{\infty} f(x) dx = 1$ (unit measure)

$$\text{Normal(Gaussian) distribution: } f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

正态高斯分布

```
geom_line() + theme_minimal()
```

- The most famous continuous distribution
- Fully specified by two parameters
 - μ the location parameter (mean)
 - σ the scale parameter (sd)
- Notation $X \sim N(\mu, \sigma)$,

```
mu <- 0; sig <- 1
x <- seq(from = mu - 4 * sig, to = mu + 4 * sig,
          length.out = 128)
dens <- dnorm(x, mean = mu, sd = sig)
# Base R graphics
plot(x, dens, type = "l")
dat <- data.frame(x = x, y = dens)
# ggplot2 version
ggplot(dat, aes(x = x, y = y)) +
```

Density estimation - Likelihood approach



THE UNIVERSITY OF
SYDNEY

Density estimation

In exploratory data analysis, an estimate of the density function can be used

- to assess multimodality, skew, tail behaviour, etc.
- in decision making, classification, and summarizing Bayesian posteriors
- as a useful visualisation tool (a simple summary of a distribution)

Suppose random variables X_1, X_2, \dots, X_n have been observed and assumed to be sampled independently from the distribution with density f .

Goal: The estimation of the density function f .

在探索性数据分析中，密度函数的估计可用于评估多模态、偏斜、尾部行为等。

在决策、分类和总结贝叶斯后验中

作为一个有用的可视化工具（一个分布的简单总结）

假设随机变量 X_1, X_2, X_n 已经被观察到，并假定从密度为 f 的分布中独立取样

Parametric density estimation

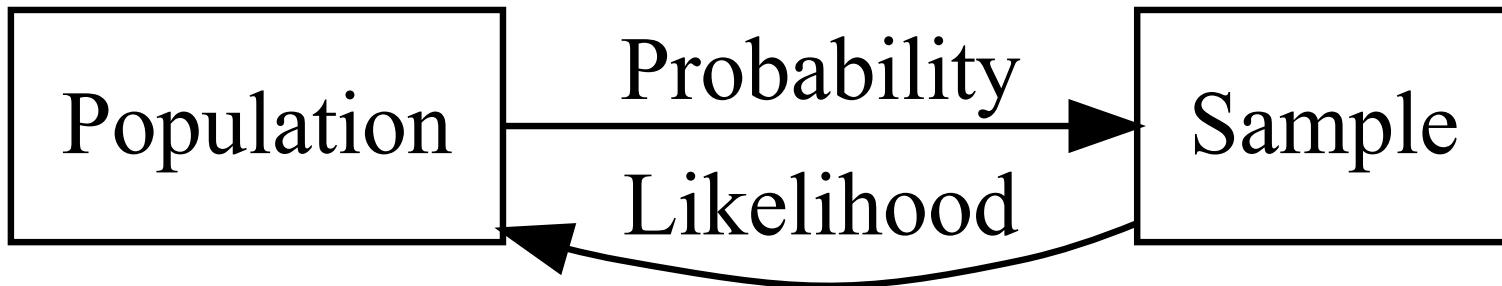
- The **parametric** approach to density estimation assumed a **parametric** model.
- That is, $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} f_{\theta}$ where θ is a parameter vector.
 - For example, $\theta = (\mu, \sigma)$ when $X \sim N(\mu, \sigma)$
- Typically the parameter θ is estimated using the method of **maximum likelihood**.
- Density function is then estimated as $f(x | \hat{\theta})$

Maximum likelihood the best value for the parameters is the one for which the probability of obtaining the observed samples is the largest.

密度估计的参数化方法假定了一个参数化模型。

通常，参数 θ 是用最大似然法 maximum likelihood 估计的。是获得观察样本的概率最大的一个。

What is a likelihood?



Simple example:

- Population has girl:boy ratio of 2:1 (100 girls for 50 boys)
- If I draw a sample of 50 people, what is the **probability** of picking 10 boys
- If I draw a sample of 50 people, and picked 10 boys, what is the **likelihood** that the girl:boy ratio is 2:1

Probability 已知 μ 和 σ 求概率 用pnorm, likelihood 已知具体的数据求参数 用dnorm

Normal distribution example

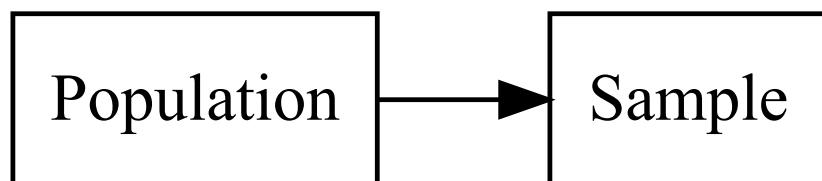
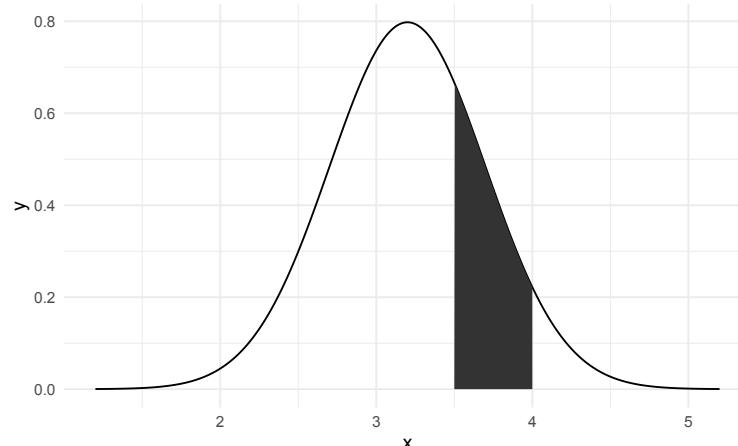
- Consider a random variable $X \sim N(3.5, 0.2)$
- What is the probability that X is between 3.5 and 4?

- Compute the area under the density. $P(3.5 \leq X \leq 4) = \int_{3.5}^4 f(t) dt$

```
mu = 3.2; sig = 0.5
pnorm(4, mean = mu, sd = sig) -
  pnorm(3.5, mean = mu, sd = sig)
```

```
## [1] 0.2194538
```

```
# Or in one line
## diff(pnorm(c(3.5, 4), mean = mu, sd = sig))
```



Likelihood

- Consider a single value is observed from $X \sim N(\mu, 0.2)$, say $x = 3.7$
- Determine the likelihood of drawing this value. Flip the perspective $f(x | \theta) \rightsquigarrow L(\theta | x)$

```
dnorm(3.7, mean = 3.5, sd = 0.2)
```

```
## [1] 1.209854
```

```
dnorm(3.7, mean = 3.6, sd = 0.2)
```

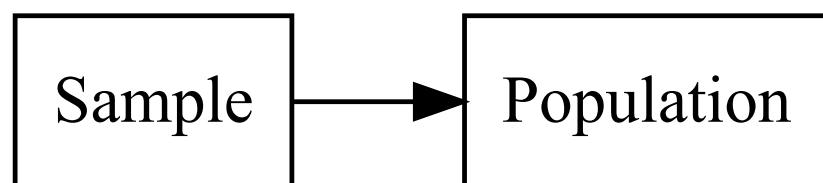
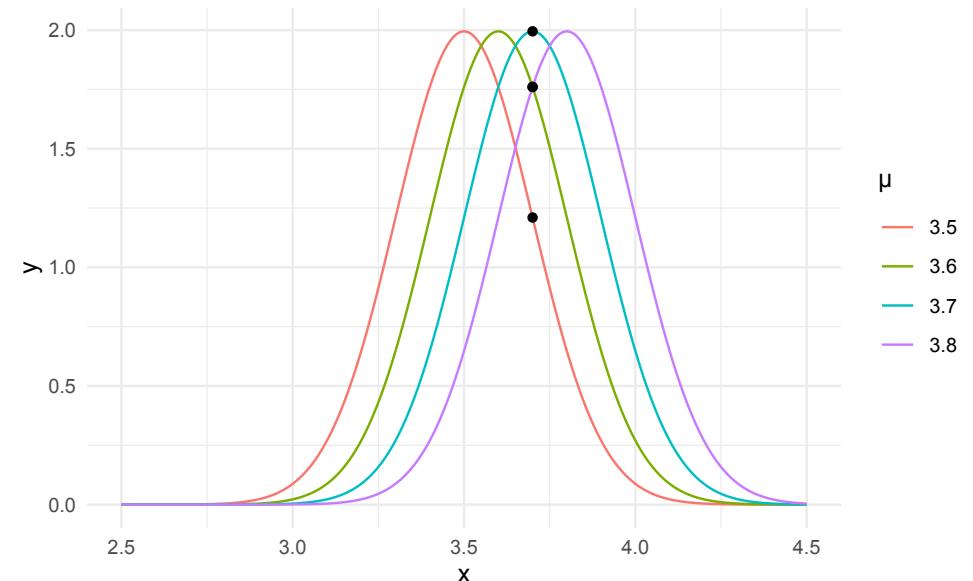
```
## [1] 1.760327
```

```
dnorm(3.7, mean = 3.7, sd = 0.2)
```

```
## [1] 1.994711
```

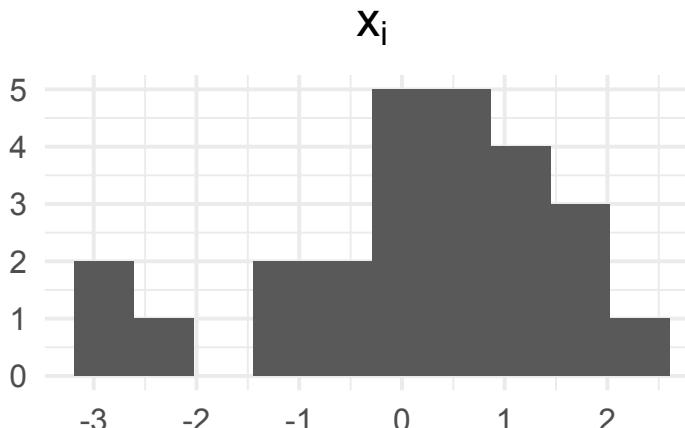
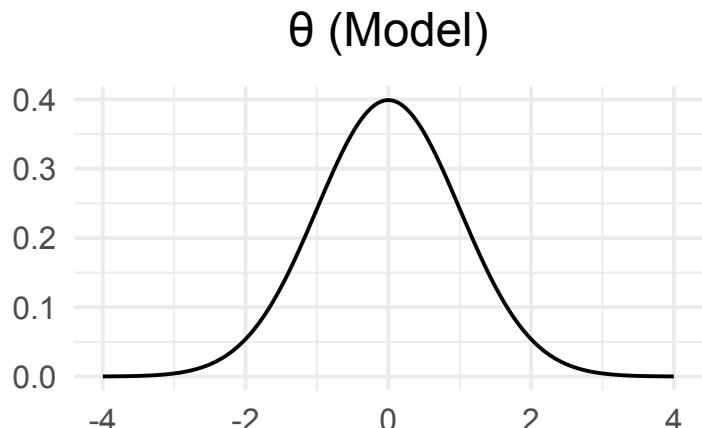
```
dnorm(3.7, mean = 3.8, sd = 0.2)
```

```
## [1] 1.760327
```



Maximum likelihood approach

- $f(x_1, x_2, \dots, x_n | \theta)$ is the probability of observing x_1, x_2, \dots, x_n given the parameter θ .



利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值

- Assuming independent and identically distributed variables $f(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$

Maximising the log-likelihood is often easier so it is common to maximise

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n f(x_i | \theta) \rightsquigarrow L(\theta | \mathbf{x}) = \ln L(\theta | \mathbf{x}) = \sum_{i=1}^n \ln f(x_i | \theta)$$

Density estimation - Non-parametric approach



THE UNIVERSITY OF
SYDNEY

Non-parametric density estimation

- Danger of misspecification with parametric approach
 - If the assumed f_θ is incorrect.
 - Serious danger of inferential errors.
- Non-parametric approaches to density estimations
 - Assume little about the structure of f
 - use *local information* to estimate f at a point x
- Histograms are
 - one type of nonparametric density estimators
 - piecewise constant density estimators
 - produced automatically by most software packages

参数化方法的错误描述的危险
如果假定的 f_θ 是不正确的。

有推断错误的严重危险。

密度估计的非参数方法

对 f 的结构假设很少

使用局部信息来估计 x 点的 f

柱状图是非参数密度估计的一种类型

非参数密度估计的一种类型

平行常数密度估计器

由大多数软件包自动生成

Histograms

- Very simple visualization
- Sensitive to the number of bins chosen and bin width

Kernel functions

- A kernel is a special type of probability density function (PDF) having the properties.
 - non-negative $K(x) \geq 0$, symmetric $K(-x) = K(x)$, unit measure $\int K(x) dx = 1$

只要K的积分等于1，就能保证估计出来的密度函数积分等于1。

Kernel functions must be continuous, symmetric

Kernel density estimation

- Kernel density estimation is a non-parametric approach estimating densities
 - Knowledge of the structure of f is not required
- Essentially, at every data point, a kernel function is created with the point at its centre.
- The PDF is estimated by adding all of these kernel functions and dividing by the number of data to ensure that it satisfies
 - every possible value of the PDF is non-negative.
 - the definite integral of the PDF over its support set equals 1

核心密度估计是一种估计密度的非参数方法

不需要对 f 的结构进行了解

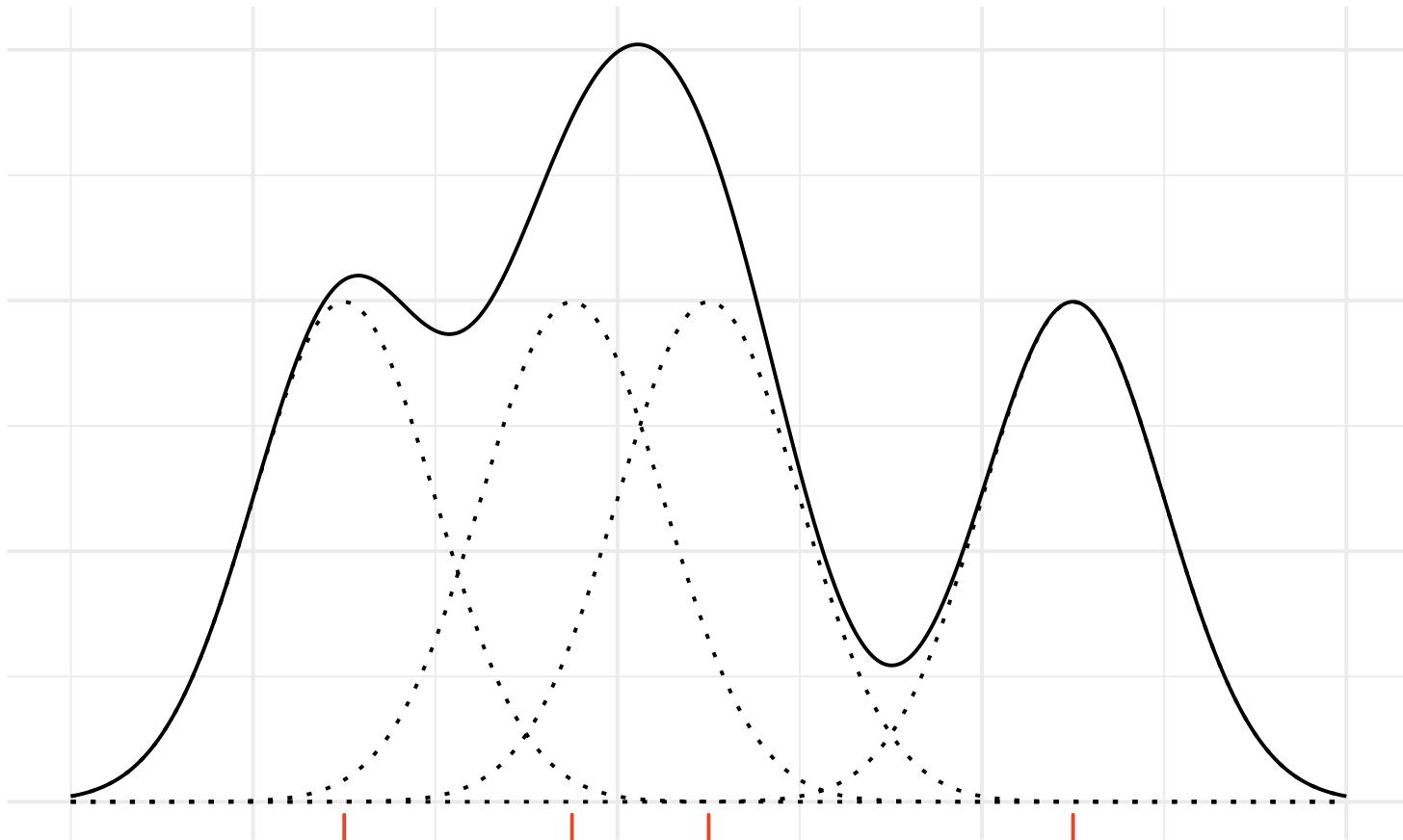
基本上，在每个数据点上，都会以该点为中心创建一个核函数。

通过将所有这些核函数相加并除以数据的数量来估计PDF。

确保它能满足PDF的每个可能值都是非负的和PDF在其支持集上的定积分等于1

Overall density estimate 等于全部weight相加

Normal kernel density estimate



- E.g. Four sampled variables marked in red with Gaussian weights sum together to give the overall density estimate

Kernel density estimator (KDE)

- A simple one weights all points within a window h of x equally

选择不同的bandwidth 选择不同的
k

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^n 1 \left\{ |X_i - x| < h \right\}$$

- More generally a univariate kernel density estimator has a general weight function (Kernel)

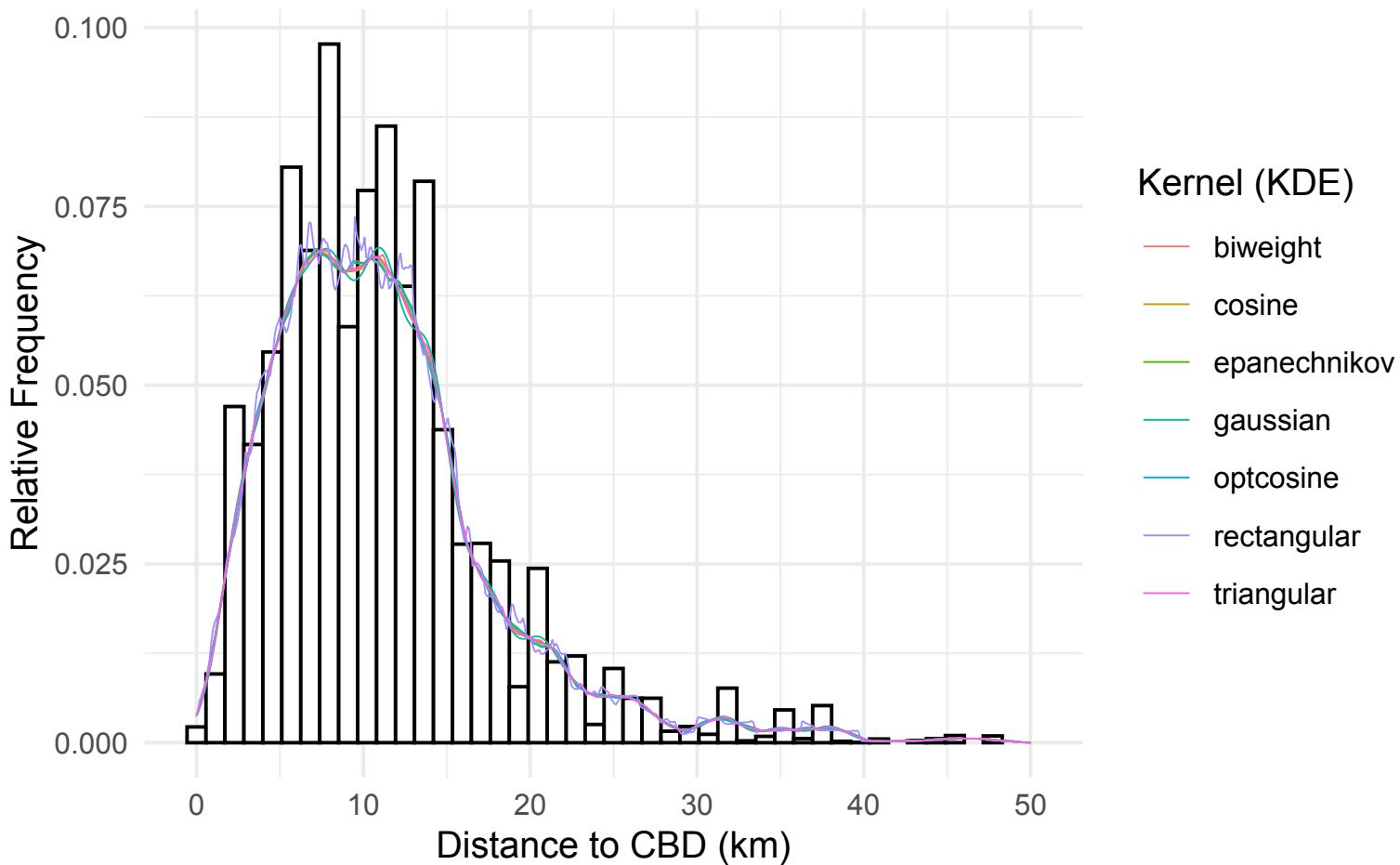
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$$

- K is a Kernel function
- h is a bandwidth parameter (possibly fixed or varying)
- Consider only h fixed for this course.

Tuning the Kernel density estimator (KDE)

- There are two main components for the KDE $\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$
 - The choice of K
 - The choice of h
- The choice of Kernel is less important and generally gives similar results
- The choice of bandwidth is important and can vary the result greatly.
- Some standard kernels

Different choices of Kernel function with same bandwidth



Computing density in

- Base  there is `density`
 - `density` computes the KDE
 - Can wrap in `plot`, i.e. `plot(density(x))`, to visualize
 - Can inspect details in `summary`
- For plotting `ggplot` there is `geom_density`
 - Can specify the bandwidth with `bw` argument

Choosing the bandwidth

- The density estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$$

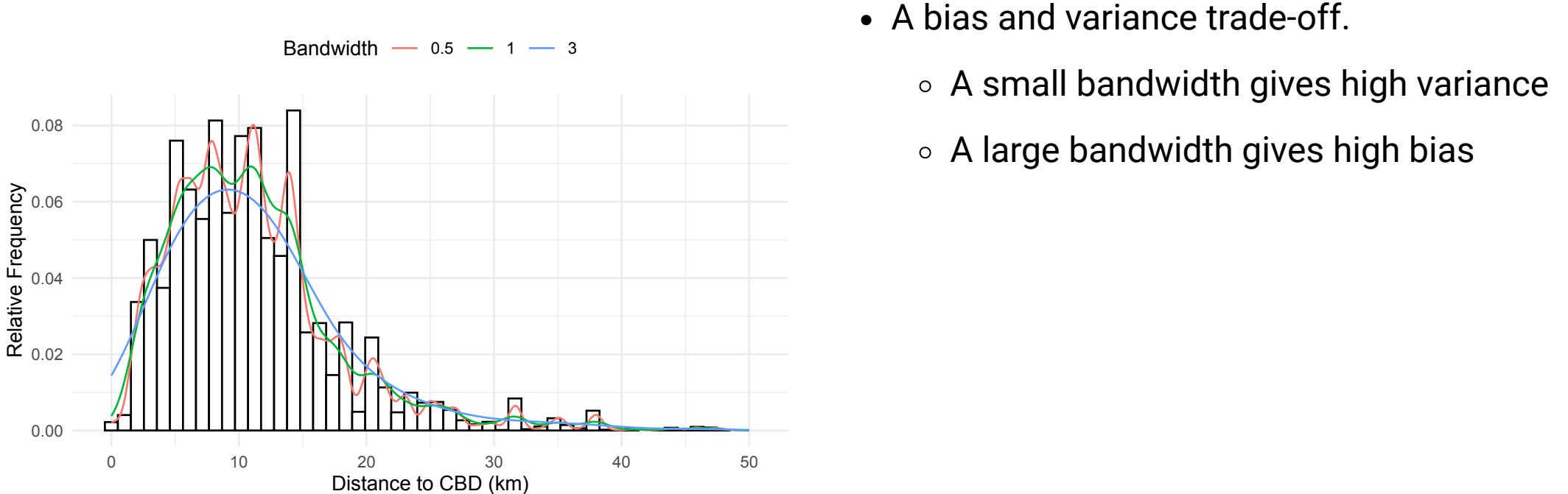
- is a fixed-bandwidth kernel density estimator since h is constant.
- If h is too small, the density estimator will tend to assign probability density too locally near observed data
 - a wiggly estimated density function with many false modes.
- If h is too large, the density estimator will spread probability density contributions too diffusely
 - smooths away important features of f

如果bandwidth太小，密度估计器将倾向于在观察到的数据附近分配太多概率密度。会造成一个具有许多错误模式的摇摆不定的估计密度函数。小的bandwidth会产生高的变异

如果bandwidth太大，密度估计器会使概率密度贡献分布得过于分散使得f的重要特征变得平滑
小的bandwidth会产生高的变异variation，较大的bandwidth会产生较高的偏差bias

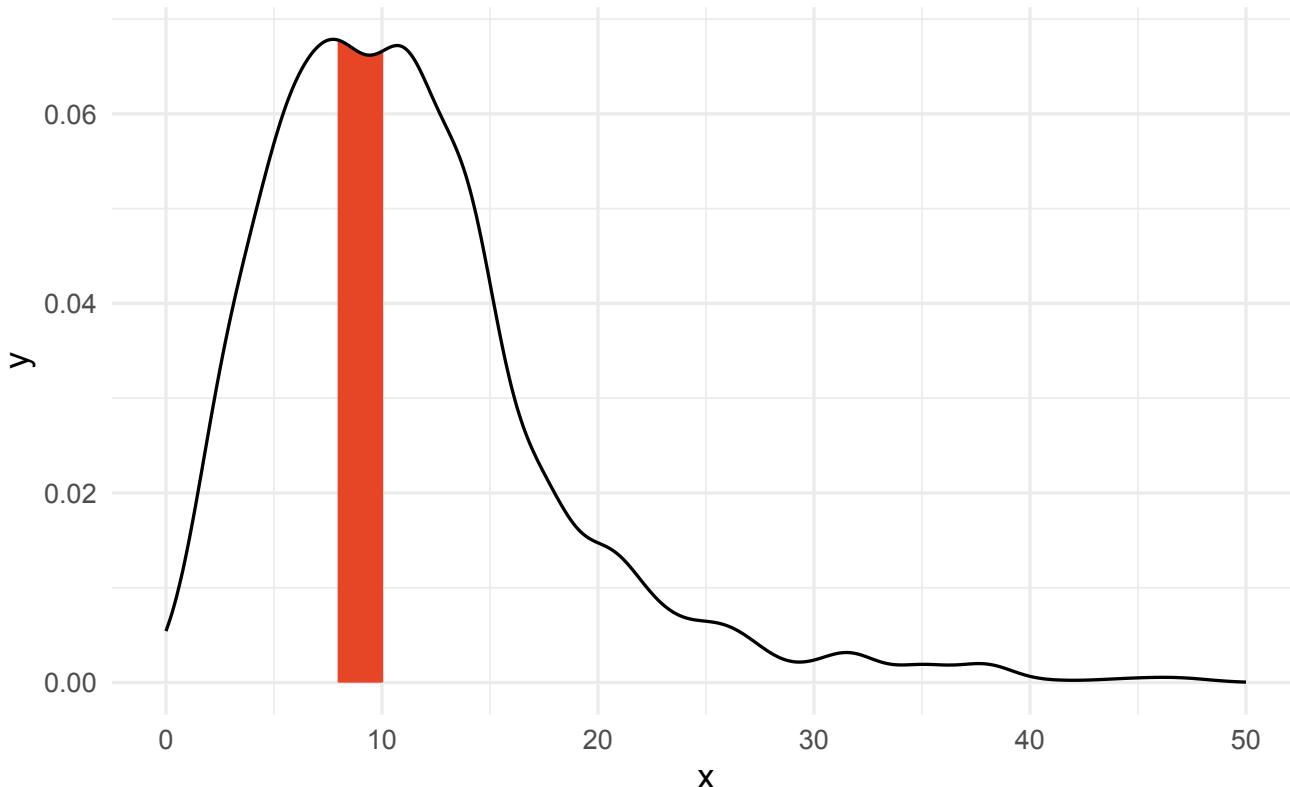
Choice of bandwidth

- Consider the distance from CBD variable again with three bandwidths



Uses of the density estimate

- **Compute probabilities:** Consider the probability a property is between 8-10km of CBD
- Integrate the density function between 8 and 10 yields $p = 0.13 \rightsquigarrow 13\%$ chance of finding a property between 8-10km of CBD



Mean squared error, Bias and Variance 平均平方误差，偏差和方差

We can decompose the mean squared error (MSE) into the sum of three quantities: The **variance**, the **squared bias** and the **variance of the error**.

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

- Variance here denoting how much would $\hat{f}(x)$ change if we estimate using a different training set.
- Bias
 - Error introduced by approximating the data using a model.

Kernel density estimation type equivalent

$$Var(\hat{f}(x)) = O\left(\frac{1}{nh}\right)$$

$$Bias(\hat{f}(x)) = O(h)$$

References

James, G, D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

STAT5003

Week 4 : High dimensional visualization and analytics

Dr. Justin Wishart



THE UNIVERSITY OF
SYDNEY



Readings

- In James, Witten, Hastie, and Tibshirani (2013)
 - PCA Dimension reduction, see Section 10.2
 - Clustering, see Section 10.3
- In Hastie, Tibshirani, and Friedman (2017)
 - MDS, see Section 14.8

Clustering



THE UNIVERSITY OF
SYDNEY

Clustering basics

- Group observations that are similar based on predefined criteria.
- Requires a similarity or dissimilarity measure
- Goals of clustering:
 - We want clusters to be compact.
 - Small distance between observations within a cluster
 - Large distance between observations between different clusters
- Example algorithms:
 - Hierarchical clustering
 - k -means clustering
 - Gaussian mixture model

根据预先设定的标准对相似的观察结果进行分组。

需要一个相似性或不相似性的衡量标准

聚类的目标goal。

我们希望聚类是紧凑compact的。

同一个聚类内的观测值之间的距离要小

不同聚类之间的观测值之间的距离大
算法。

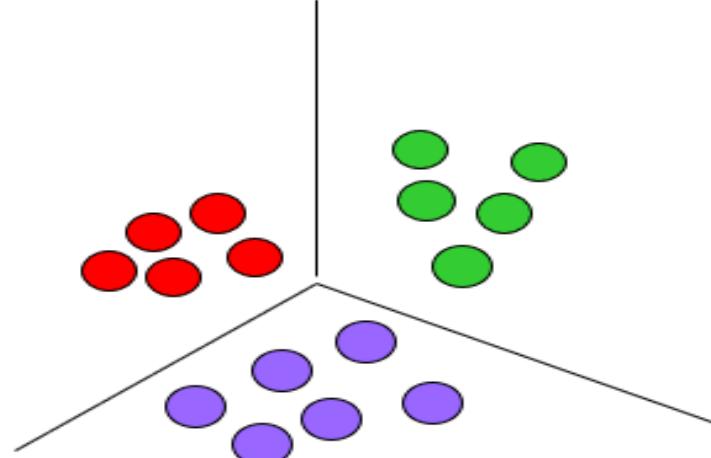
-分层聚类Hierarchical clustering

-均值聚类 k-means clustering

高斯混合模型 Gaussian clustering

Typical methods

Partitioning



- Partitioning

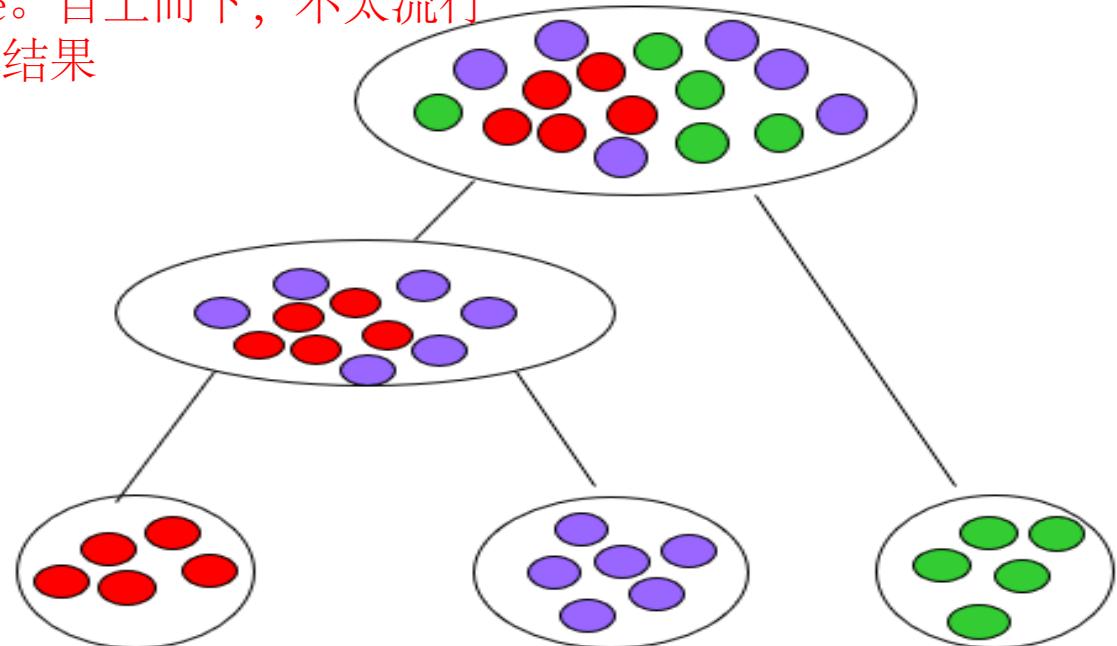
- Pre-specified number K of mutually exclusive and exhaustive groups.
- Iterate until criteria is met.

预先指定数量的互斥和穷举组。
迭代直到满足标准。

聚合式Agglomerative。自下而上，更受
欢迎

分裂性Divisive。自上而下，不太流行
用树状图显示结果

Hierarchical



- Hierarchical methods. Two paradigms
 - Agglomerative: Bottom up, more popular
 - Divisive: Top down, less popular
 - Display results with dendrogram

***k*-means approach**

欧式距离，认为两个目标的距离越近，相似度越大

将每个观测值随机地初始化为一个簇。迭代以下内容，直到收敛。

- Initialize each observation at random to a cluster.
- Iterate the following until convergence.

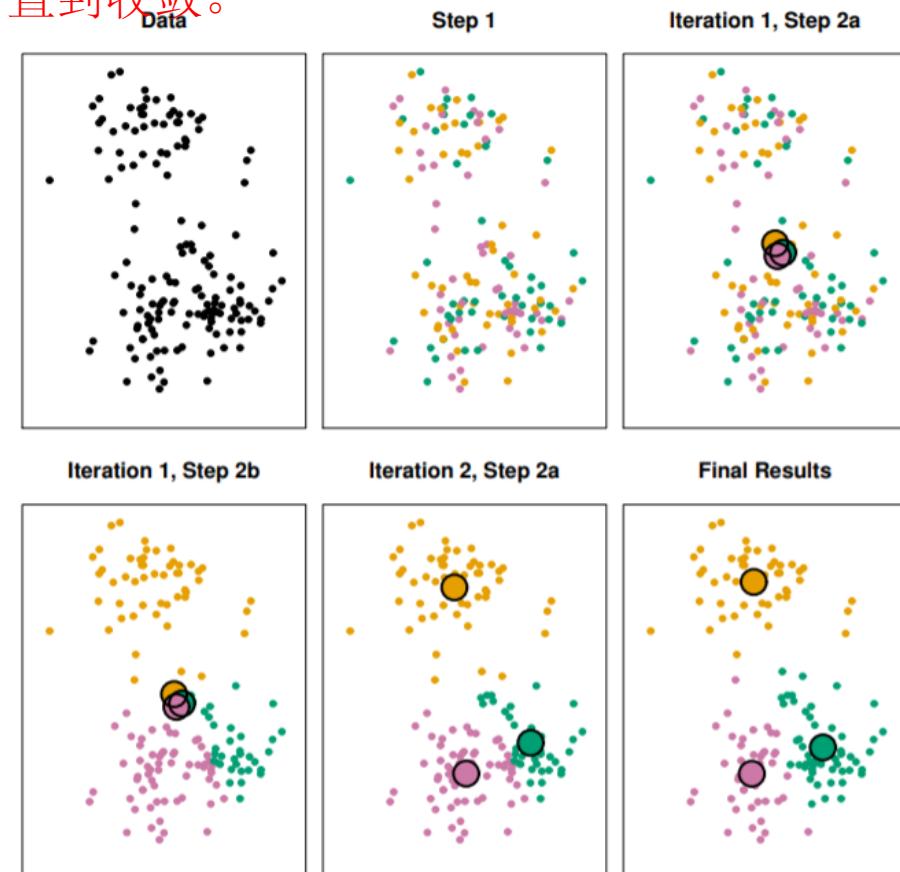
1. Find cluster means with cluster memberships fixed 找到群集的平均值cluster means, 群集成员关系固定

$$\widehat{\bar{x}}_j = \operatorname{argmin}_m \sum_{\text{cluster}(i)=j} \|x_i - m\|^2$$

2. Find cluster memberships with cluster means fixed

$$\widehat{\text{cluster}}(i) = \operatorname{argmin}_k \|x_i - \widehat{\bar{x}}_k\|^2$$

在集群平均值cluster means固定的情况下，寻找集群成员cluster memberships



k-means properties

- The number of clusters K needs to be specified.
- Local solution and not necessarily global solution.
- Depends on starting values (the random starting values).
- Best for compact, spherical clusters.
- Does not work well when cluster sizes are different.

需要指定群集clusters K的数量。

局部解决方案，不一定是全局解决方案。

取决于起始值（随机起始值）。

对紧凑的、球形的聚类最好。

当集群大小不同时，效果不好

不适合太离散的分类、样本类别不平衡的分类、非凸形状的分类

Choosing K

方差之和

- For cluster C_k can define within-group sum of squares as:

$$WSS_k = \frac{1}{|C_k|} \sum_{i,j \in C_k} \|x_i - x_j\|^2$$

这是Euclidean distances的总和，除以聚类中观测值的总数。

- This is the sum of all the pairwise squared Euclidean distances between observations in the k^{th} cluster, divided by total number of observations in the k^{th} cluster.
- The total within sum of squares criterion aggregates this metric across

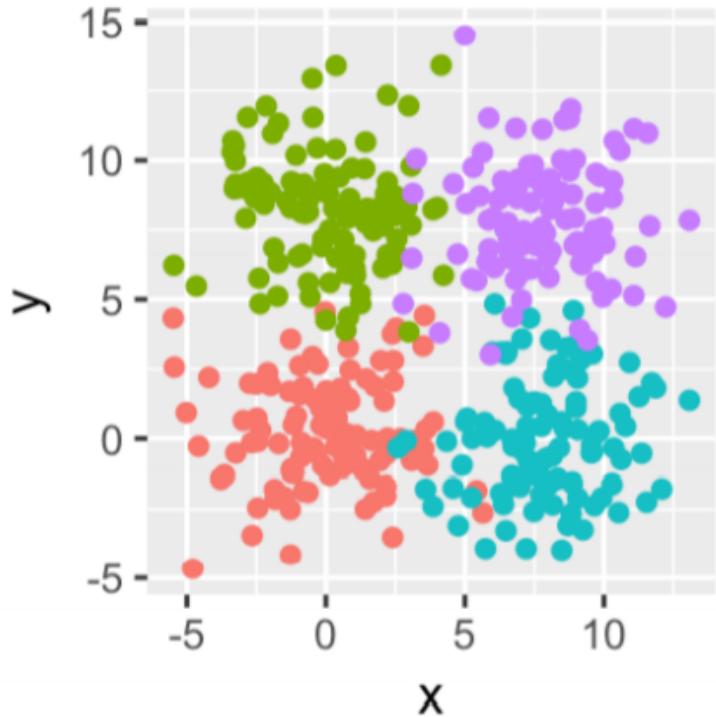
$$WSS_{Total} = \sum_{k=1}^K WSS_k$$

WSS为其它点到中心点的距离

- The total within sum of square criterion will decrease as k increases.
- Rule of thumb: Look for the elbow

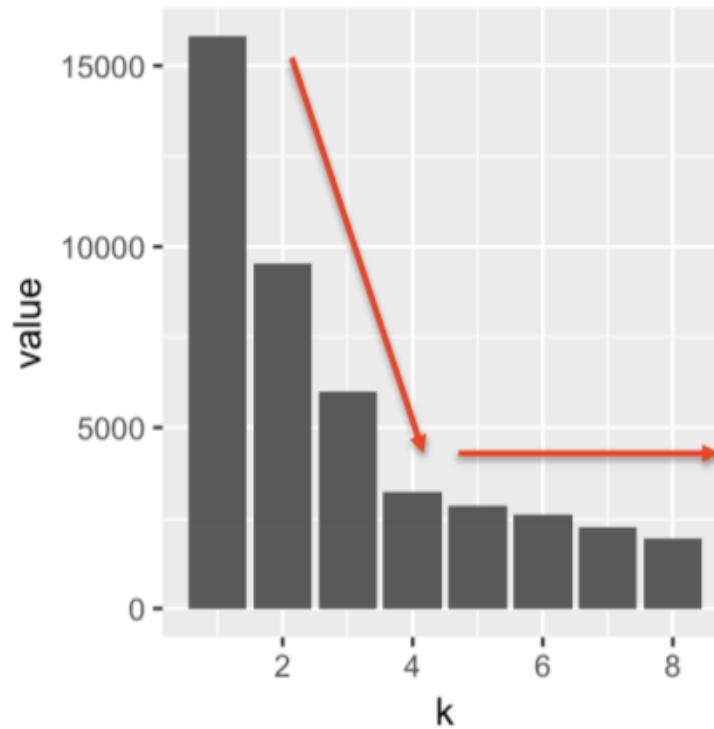
K值的选择 看Elbow plot, 下降点越明显然后平缓
(拐弯处) 的则为最适合的k值

Elbow plot



class

- 0:0
- 0:8
- 8:0
- 8:8



Hierarchical Clustering

- Begin with every observation representing a single cluster.
- At each iteration, merge the two closest clusters into one cluster.
 - Needs a measure of similarity/dissimilarity between two clusters
 - These measures are called linkages.
- Linkages - Measure of dissimilarity between two sets of objects that determine how two set of objects are merged.
 - Single linkage.
 - Complete linkage.
 - Average Linkage.



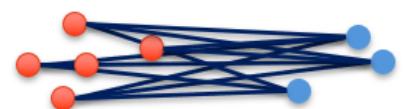
Single (minimum)



Complete (maximum)



Distance between centroids



Average (mean) linkage

开始时，每个观测值都代表一个聚类。在每次迭代中，将两个最接近的聚类合并为一个聚类。需要对两个聚类之间的相似性/不相似性进行衡量。这些措施被称为联系Linkages。

联系Linkages - 衡量两组对象之间的不相似性。

衡量两组对象之间的差异，决定两组对象如何

Dimension reduction: Principal Components Analysis (PCA)



THE UNIVERSITY OF
SYDNEY

High dimensional data

多维数据是指特征feature多余观察值observation

- High-dimensional data refers to data set with more features p than observations n
 - Examples: in genetic data, we can easily measure 500k individual DNA mutations (human genome have \sim 3 billion base pairs of DNA), but experiments generally have < 1000 people e.g. $p \sim 500k, n \sim 1000$
- It is very hard to visualize high-dimensional data
 - Only have 2 (sometimes 3 or 4) dimensional canvas to create plots.
- Many algorithms and methods have been designed for low dimensional data and would not work well for high-dimensional data
- To build a linear regression model data with $500k$ features will result in $500k$ parameters. This problem is underdetermined if we only have 1000 observations.

Dimension reduction

- Dimension reduction can be a pre-processing step, do it before applying clustering, classification and/or regression
- Data with small number of dimensions are easier to visualize and plot
- Dimension reduction can be a useful exploratory data analysis tool.

降维可以是一个预处理步骤，在应用聚类、分类和/或回归之前进行。
维数少的数据更容易被可视化和绘制。降维可以是一个有用的探索性
数据分析exploratory data analysis工具

Dimension reduction strategies

- Eliminate or remove features
 - Need to decide which features to be eliminated? Keep ones with high variance?
- Select features
 - E.g. Lasso and ridge regression (coming soon in later module)
- Build or construct new features from existing ones
 - Replace many existing features with a single one.
 - PCA and t -SNE

消除或剔除特征：需要决定哪些特征要被消除？保留那些高变异的特征？

选择特征：例如，套索和岭回归Lasso and ridge regression。

从现有的特征中建立或构建新的特征：用一个特征取代许多现有特征。PCA和t-SNE

PCA

- Suppose we have a data matrix X with n observations and p features.
 - Can we plot the data in a 2-dimensional plot?
- Naively, we can do all pairwise combinations i.e. 1 vs 2, 1 vs 3, ..., (p vs $(p - 1)$)
 - $\binom{p}{2} = \frac{p(p-1)}{2} = \mathcal{O}(p^2)$ different plots!
- Principal components analysis (PCA) finds a way to represent the data in a different space
 - It is still p dimensional, albeit a different coordinate space.
 - Aims to explain most variation in the first few dimensions.

总结一下 PCA 的算法步骤：

设有 m 条 n 维数据：

将原始数据按列组成 n 行 m 列矩阵 X ；

将 X 的每一行进行零均值化，即减去这一行的均值；

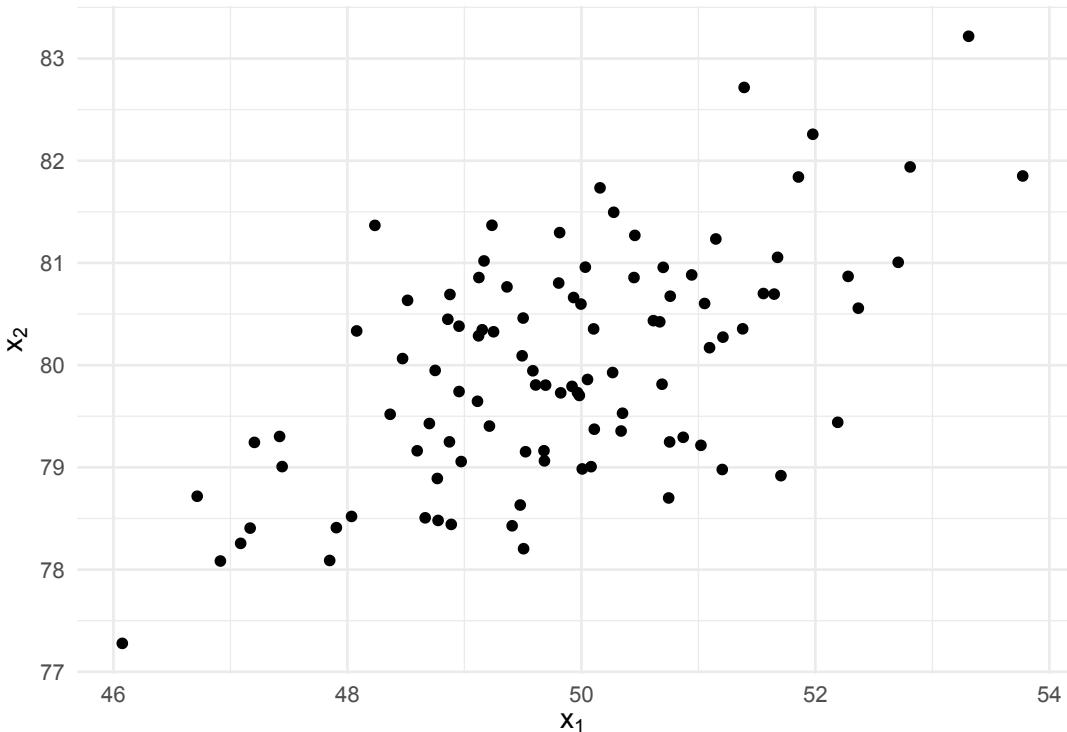
求出协方差矩阵 $C = \frac{1}{m} X^T X$ ；

求出协方差矩阵的特征值及对应的特征向量；

将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 K 行组成矩阵 P ；

求得矩阵 $Y = PX$ ，即为降维到 K 维后的数据。

Best way to represent 2d in 1d?



- Could use a single variable? x_1 say?
- Or could remap x_1 and x_2 to a single variable
 - $z = \phi_1 x_1 + \phi_2 x_2$
- Can generalize this to many dimensions.
 - $z = \sum_{i=1}^p \phi_i x_i$
- Pick the transformation that maximises the variance!

Principal Components

Start with a data matrix \mathbf{X} , assume it has mean zero.

$$\mathbf{X} = (X_1 \quad X_2 \quad \dots \quad X_p)$$

The first principal component is the **normalised** linear combination of the features that **maximises the variance** in the new component. 第一个主成分是特征的归一化线性组合，使新成分的方差达到最大。

$$Z = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p = \sum_{i=1}^p \phi_{i1}X_i = \boldsymbol{\phi}_1^T \mathbf{X}$$

The elements ϕ_{i1} are known as the **loadings** of the first principal component

- By normalised, we mean the squared loadings have to sum to 1, i.e. $\sum_{i=1}^p \phi_{i1}^2 = 1 \Leftrightarrow \boldsymbol{\phi}_1^T \boldsymbol{\phi}_1 = 1$
- Also, it is desired to maximise

$$\text{Var}(Z_1) = \text{Var}(\boldsymbol{\phi}_1^T \mathbf{X}) = \sum_{i=1}^p \phi_{i1}^2 \text{Var}(X_i) + \sum_{i \neq j} \phi_{i1}\phi_{j1} \text{Cov}(X_i, C_j)$$

Solving the first principal component (not assessable)

- To find the first principal component, solve the following optimization problem

$$\max_{\phi_{11}, \dots, \phi_{p1}} \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \quad \text{such that} \quad \sum_{i=1} \phi_{i1}^2 = 1 \quad (1)$$

- Define the Covariance matrix

$$\boldsymbol{\Sigma} = \mathbb{V}\text{ar}(\mathbf{X}) = \begin{pmatrix} \mathbb{V}\text{ar}(X_1) & \mathbb{C}\text{ov}(X_1, X_2) & \dots & \mathbb{C}\text{ov}(X_1, X_p) \\ \mathbb{C}\text{ov}(X_1, X_2) & \mathbb{V}\text{ar}(X_2) & \dots & \mathbb{C}\text{ov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}(X_p, X_1) & \dots & \dots & \mathbb{V}\text{ar}(X_p) \end{pmatrix}$$

- Also, $\mathbb{V}\text{ar}(Z_1) = \mathbb{V}\text{ar}(\boldsymbol{\phi}_1^T \mathbf{X}) = \boldsymbol{\phi}_1^T \boldsymbol{\Sigma} \boldsymbol{\phi}_1$ and the above optimization is equivalent to

$$\max_{\boldsymbol{\phi}_1} \boldsymbol{\phi}_1^T \boldsymbol{\Sigma} \boldsymbol{\phi}_1 \quad \text{such that} \quad \boldsymbol{\phi}_1^T \boldsymbol{\phi}_1 = 1$$

Solving via multivariable calculus (not assessable)

- Can solve this with multivariable calculus! The Lagrangian.

$$L(\boldsymbol{\phi}, \lambda) = \boldsymbol{\phi}_1^T \boldsymbol{\Sigma} \boldsymbol{\phi}_1 + \lambda(1 - \boldsymbol{\phi}_1^T \boldsymbol{\phi}_1)$$

- Computing partial derivatives and solving

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\phi}_1} &= 2\boldsymbol{\Sigma}\boldsymbol{\phi}_1 - 2\lambda\boldsymbol{\phi}_1 = \mathbf{0}, & \frac{\partial L}{\partial \lambda} &= 1 - \boldsymbol{\phi}_1^T \boldsymbol{\phi}_1 = 0 \\ &\Updownarrow & &\Updownarrow \\ \boldsymbol{\Sigma}\boldsymbol{\phi}_1 &= \lambda\boldsymbol{\phi}_1 & \boldsymbol{\phi}_1^T \boldsymbol{\phi}_1 &= 1\end{aligned}$$

This is the eigenvalue equation. The **eigenvector** of $\boldsymbol{\Sigma}$ gives the loadings.

Solving the second principal component (not assessable)

- Can repeat the process to get the next principal component.
- Find ϕ_2 to optimise

$$\max_{\phi_{12}, \dots, \phi_{p2}} \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j2} x_{ij} \right)^2 \quad \text{such that} \quad \sum_{i=1} \phi_{i2}^2 = 1 \quad \text{and} \quad \sum_{i=1} \phi_{i1} \phi_{i2} = 0$$

- Or using vector notation

$$\max_{\phi_2} \phi_2^T \Sigma \phi_2 \quad \text{such that} \quad \phi_2^T \phi_2 = 1 \quad \text{and} \quad \phi_2^T \phi_1 = 0$$

Principal Component Scores

- Given the principal component loadings, we can project our data matrix \mathbf{X} onto the principal component space.
 - The projection is a linear combination of the sample feature values:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \dots + \phi_{p1}x_{ip}$$

This is known as the principal component **score**.

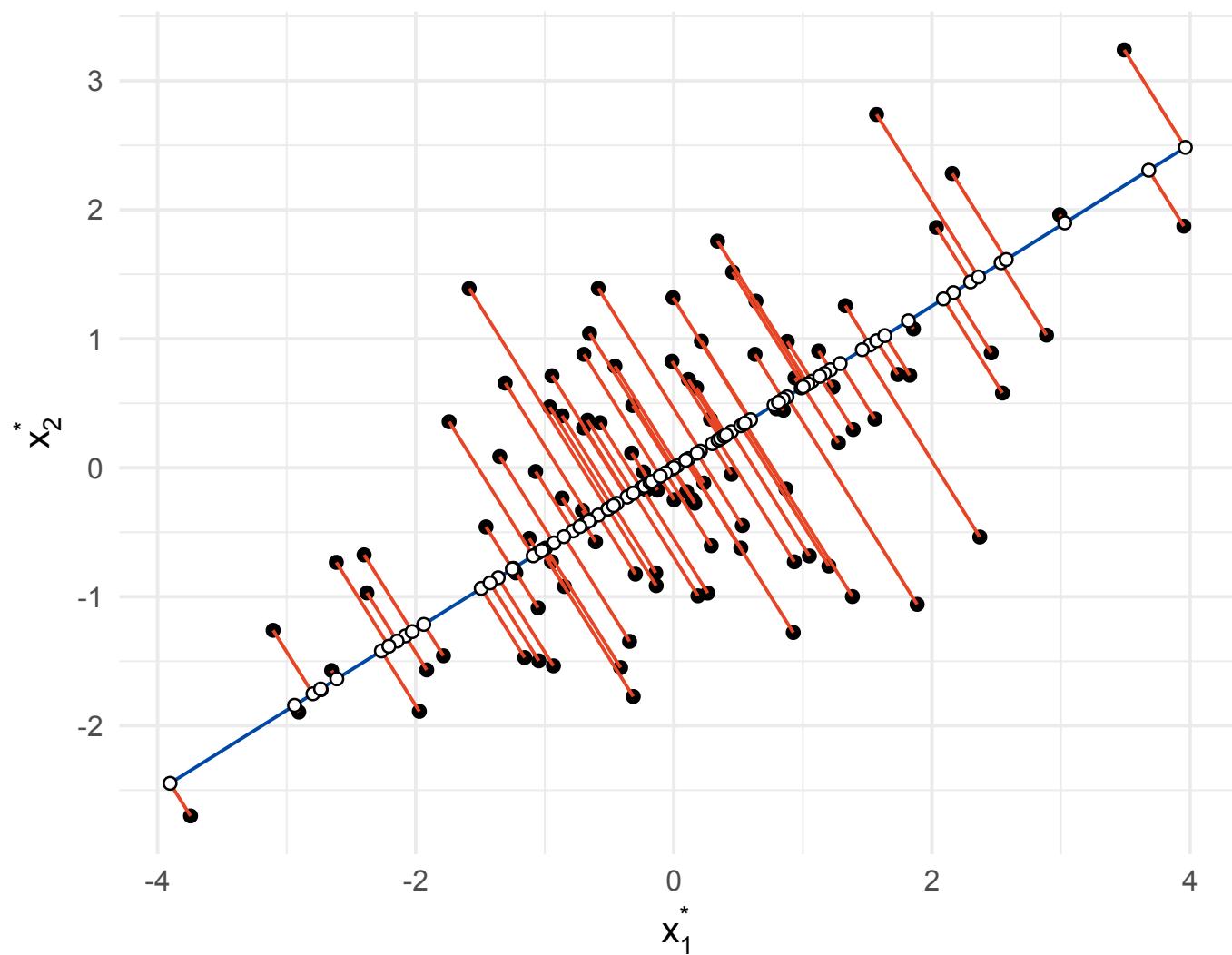
- The first principal component score vector is

$$\mathbf{Z}_1 = (z_{11}, z_{21}, \dots, z_{n1})$$

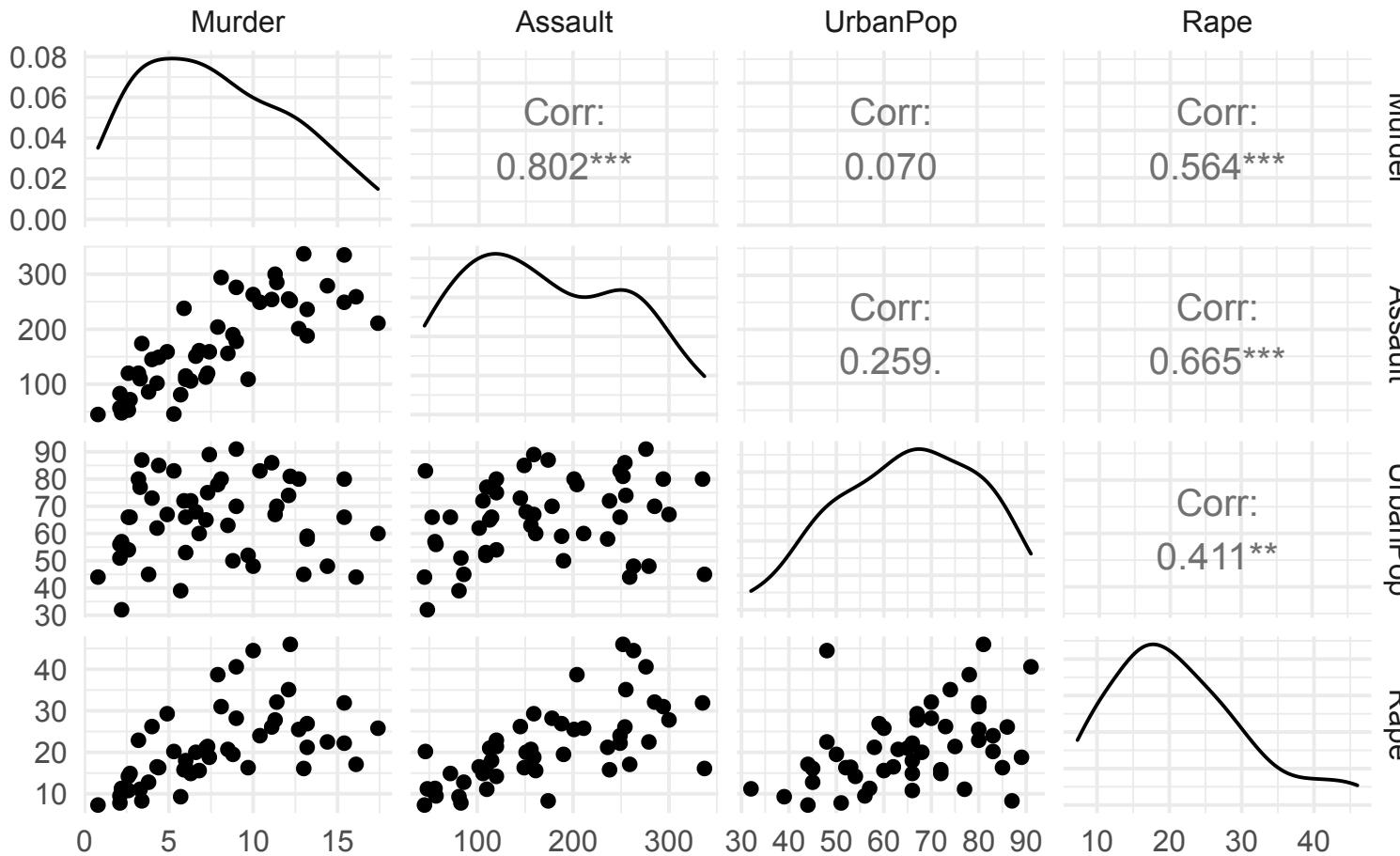
The principal component score vectors are **uncorrelated**.

主成分得分向量是不相关的。

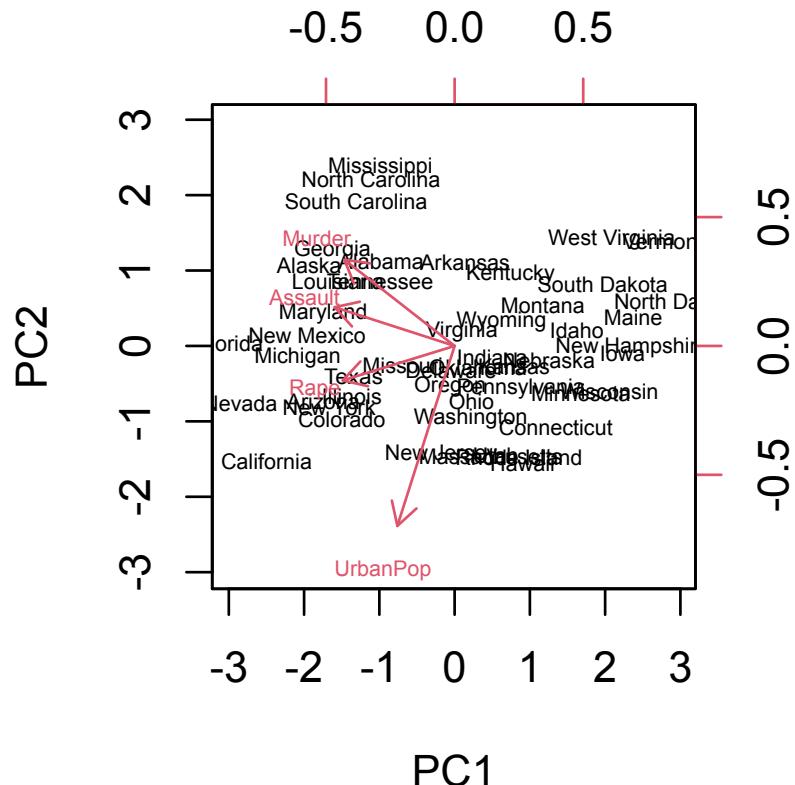
Geometric interpretation



USArests Example



Biplot of the USArrests

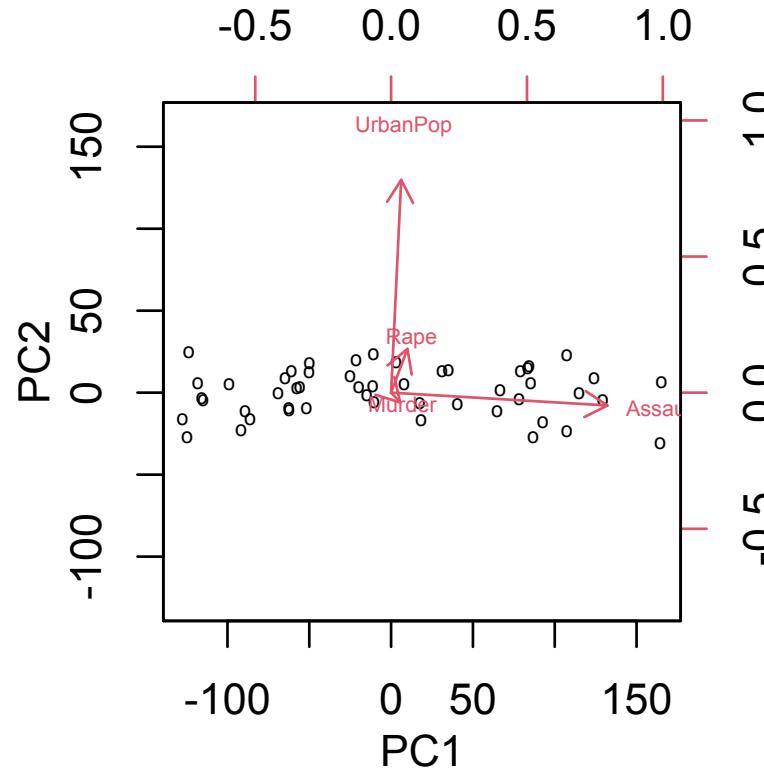
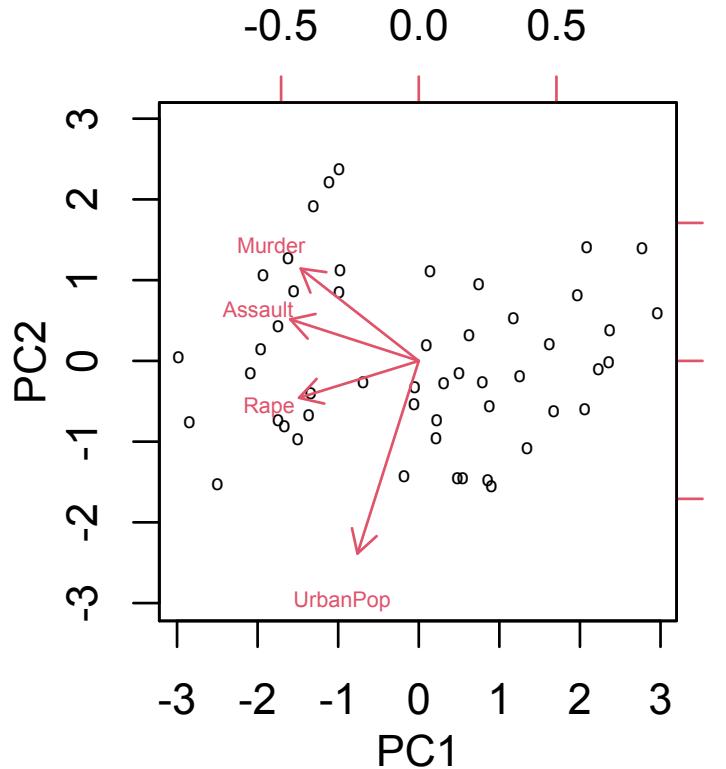


Scaling the variables

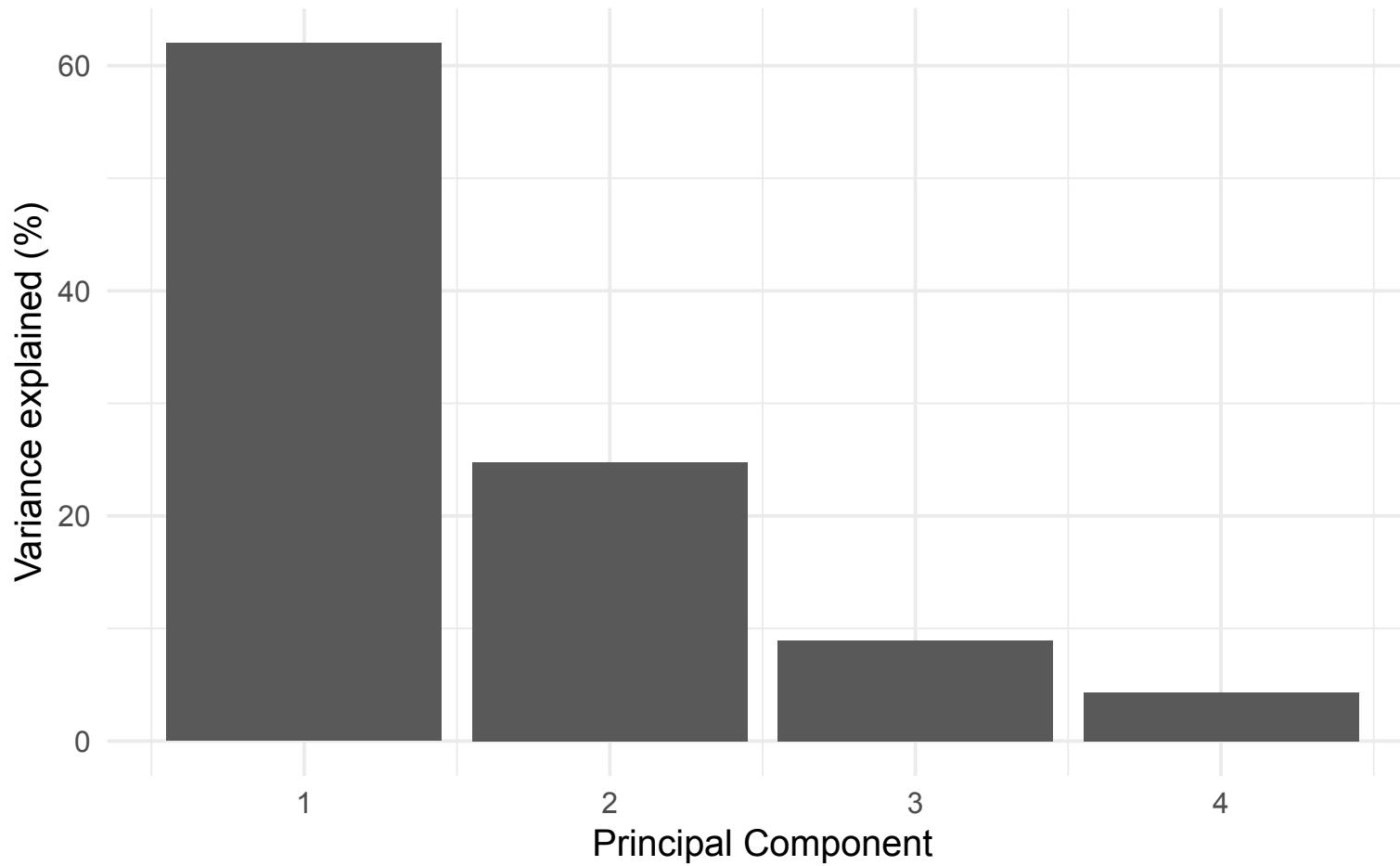
- In a PCA analysis, it is common to centre the variable by removing the mean
- You can also standardise the data to make all the variables have a standard deviation of 1
- If the variables have different units (e.g. in the USArrests dataset, Murder is measured as number per 100,000 people, but UrbanPop is the percentage of population that lives in urban area), the variance would be very different
- The loadings will put more weight on variables with higher variance
 - this may not be what you want!
- However, if all the variables share the same unit, then standardisation may not be necessary

在PCA分析中，通常通过去除平均值来对变量进行中心化处理。你也可以将数据标准化，使所有变量的标准差为1。然而，如果所有的变量都有相同的单位，那么标准化可能就没有必要了

Effect of scaling (left) vs unscaled (right)



Scree plot



Properties of PCA

- Unique and Global solution!
- Ordered components
- Best low rank approximation to the data

$$\min_{\widehat{\mathbf{X}}} \|\mathbf{X} - \widehat{\mathbf{X}}\|_F^2 \quad \text{such that } \text{rank}(\widehat{\mathbf{X}}) = p$$

- Best linear dimension reduction possible
- Is not the best for non-linear relationships

可能的最佳线性降维
对非线性关系不是最好的

PCA 的一些性质

缓解维度灾难：PCA 算法通过舍去一部分信息之后能使得样本的采样密度增大（因为维数降低了），这是缓解维度灾难的重要手段；

降噪：当数据受到噪声影响时，最小特征值对应的特征向量往往与噪声有关，将它们舍弃能在一定程度上起到降噪的效果；

过拟合：PCA 保留了主要信息，但这个主要信息只是针对训练集的，而且这个主要信息未必是重要信息。有可能舍弃了一些看似无用的信息，但是这些看似无用的信息恰好是重要信息，只是在训练集上没有很大的表现，所以 PCA 也可能加剧了过拟合；

特征独立：PCA 不仅将数据压缩到低维，它也使得降维之后的数据各特征相互独立；

PCA with K-means

- Very common approach to deal with high dimensional data
- Use the first M principal component scores as inputs into the kmeans algorithm ($M \ll p$)
- Can help improve the clustering model if the signal in the data can be captured in a few principal components

处理高维数据的非常普遍的方法

使用前 M 个主成分分数作为Kmeans算法的输入($M < p$)

如果数据中的信号能被几个主成分所捕捉，则有助于改善聚类模型组成部分

PCA with regression

- Use the first M principal component scores as the predictors in a linear regression model
- We are assuming that a small number of principal components can explain most of the variability in the data as well as the response
- PCR is useful when variables in the data are highly correlated (i.e. collinear)

在线性回归模型中使用前 M 个主成分的分数作为预测因子
我们假设少量的主成分可以解释数据中的大部分变异性variability和响应response。
当数据中的变量高度相关highly correlated时， PCA很有用。

Dimension reduction t -SNE



THE UNIVERSITY OF
SYDNEY

t-SNE

欧氏距离转化为条件概率来表征点间相似度使用梯度下降算法来使低维分布学习/拟合高维分布

- Non-linear technique developed for visualizing high dimensional datasets
- Uses local structure in the data to find a low dimensional representation
- Applications include computer security research, music analysis, cancer research, bioinformatics, and biomedical signal processing.

为实现高维数据集的可视化而开发的非线性技术.利用数据中的局部结构，形成低维表示法

MNIST Example

- 28 by 28 pixel images of handwritten digits
- 784 features

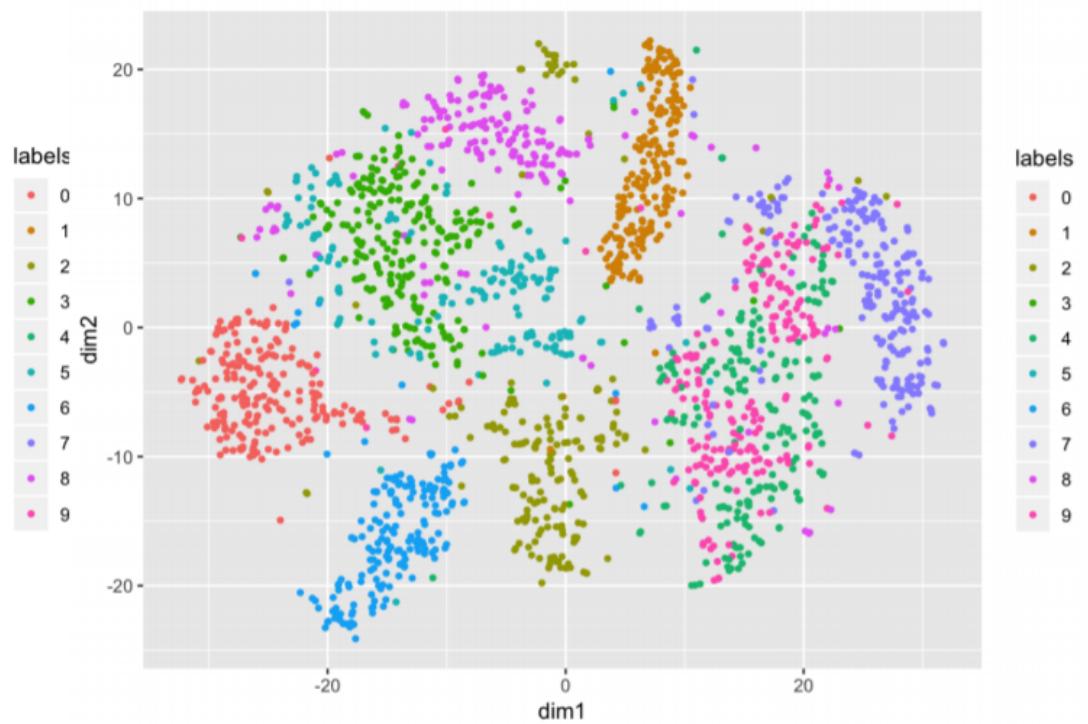
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

MNIST Example

PCA



tSNE



t-SNE vs PCA

- *t*-SNE is a probabilistic method – it will give you a different representation every time you run it.
- PCA is defined by a mathematical formula
- *t*-SNE is mostly a visualization method. The PCs from PCA can be interpreted whereas *t*-SNE representation cannot be used for inference.
- *t*-SNE is more computationally intensive than PCA
- PCA is a linear method so can only capture linear relationships whereas *t*-SNE can find more complicated non-linear relationships

PCA VS T-SNE

T-SNE是一种概率Probability方法--每次运行它都会给你一个不同的表示。PCA是由一个数学公式定义的
T-SNE主要是一种可视化的方法。PCA可以被解释，而T-SNE表示不能用于推理。

T-SNE的计算量比PCA要大。PCA是一种线性方法，所以只能捕捉线性关系，而T-SNE可以发现更复杂的非线性关系。

PCA是一种线性降维技术。它试图保留数据的全局结构Global Structure。与 t-SNE 相比，它 poor effect。它不涉及超参数Hyperparameters。它受到异常值outlier的影响很大。PCA 是一种确定性算法Deterministic algorithms。它的工作原理是旋转向量以保持方差Rotate vectors to maintain variance。我们可以找到决定使用特征值保留多少方差variance to retain using feature

t-SNE是一种非线性降维Dimensionality reduction技术。它试图保留数据的本地结构local (集群)。它是最好的降维技术之一。它涉及超参数。它可以处理异常值。它是一种非确定性或随机化算法Non-deterministic or randomized algorithms。它通过最小化 Gaussian 中点之间的距离来工作。我们不能保留方差，而是可以使用超参数来保留距离。

Three steps in t -SNE

1. Constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked while dissimilar points have an extremely small probability of being picked.
2. Defines a similar probability distribution over the points in the low-dimensional map.
3. Minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map.

Recommended to view the guide at <https://distill.pub/2016/misread-tsne/> for more information on using the t-SNE framework.

1. 在高维物体对上构建一个概率分布，使相似的物体被选中的概率很高，而不相似的点被选中的概率极小。物体有很高的概率被选中，而不相似的点有极小的概率被选中。
2. 在低维地图中的点上定义一个类似的概率分布。
3. 使两个分布之间的Kullback-Leibler分歧最小化，即关于地图中各点的位置

More details: Step 1 (Not assessable)

Given a set of n high dimensional objects x_1, x_2, \dots, x_n in p -dimensional space, t-SNE first computes probabilities p_{ij} that are proportional to the similarities of objects x_i and x_j as follows:

$$p_{j|i}(\sigma_i^2) = \frac{\phi(x_j; x_i, \sigma_i^2)}{\sum_{k \neq i} \phi(x_k; x_i, \sigma_i^2)}$$

- $\phi(x; \mu, \sigma^2)$ denotes the Gaussian density.
- Think of $p_{j|i}(\sigma_i^2)$ as a conditional probability that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i and with variance σ_i^2
- The conditional probability can be made symmetric with

$$pij = \frac{p_{j|i} + p_{i|j}}{2n}$$

More details: Step 2 (Not assessable)

- Learn a lower dimensional representation of the data: y_1, y_2, \dots, y_n that preserves the similarities p_{ij} as much as possible
- In the low dimensional space, use the heavy-tailed Student t -distribution with one degree of freedom to define similarities.
- Hence define similarities q_{ij} in low dimensional space as:

$$q_{ij} = \frac{(1 + \|y_j - y_i\|_2^2)^{-1}}{\sum_{k \neq i} (1 + \|y_j - y_k\|_2^2)^{-1}}$$

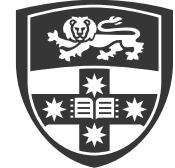
More details: Step 3 - Kullback-Leibler divergence (Not assessable)

- Find the low-dimensional representation y_1, y_2, \dots, y_n that minimizes the Kullback-Leibler (KL) divergence of the distribution q_{ij} from p_{ij} .
- KL divergence is a non-symmetric measure of the difference between two probability distributions
- KL divergence is defined as:

$$KL(p||q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

Think of KL divergence as a measure of how many bits of information is lost when we use q to approximate p

Dimension reduction: Multidimensional Scaling (MDS)



THE UNIVERSITY OF
SYDNEY

Multidimensional Scaling (MDS)

- Visually represent proximities (similarities or distances) between objects in a lower dimensional space. (usually 2 or 3d space)
- The objective of MDS is to take a Matrix of similarities or dissimilarities, D , and find projections z_1, \dots, z_k where k is the desired lower dimension.
- The distances are near preserved by optimizing a stress function
- Full data not required

视觉上表示低维空间中物体之间的接近性（相似性或距离）。(通常是2或3维空间)

MDS的目的是取一个相似性或不相似性的矩阵，以及投影，其中是所需的低维。

通过优化一个压力函数来保留距离。

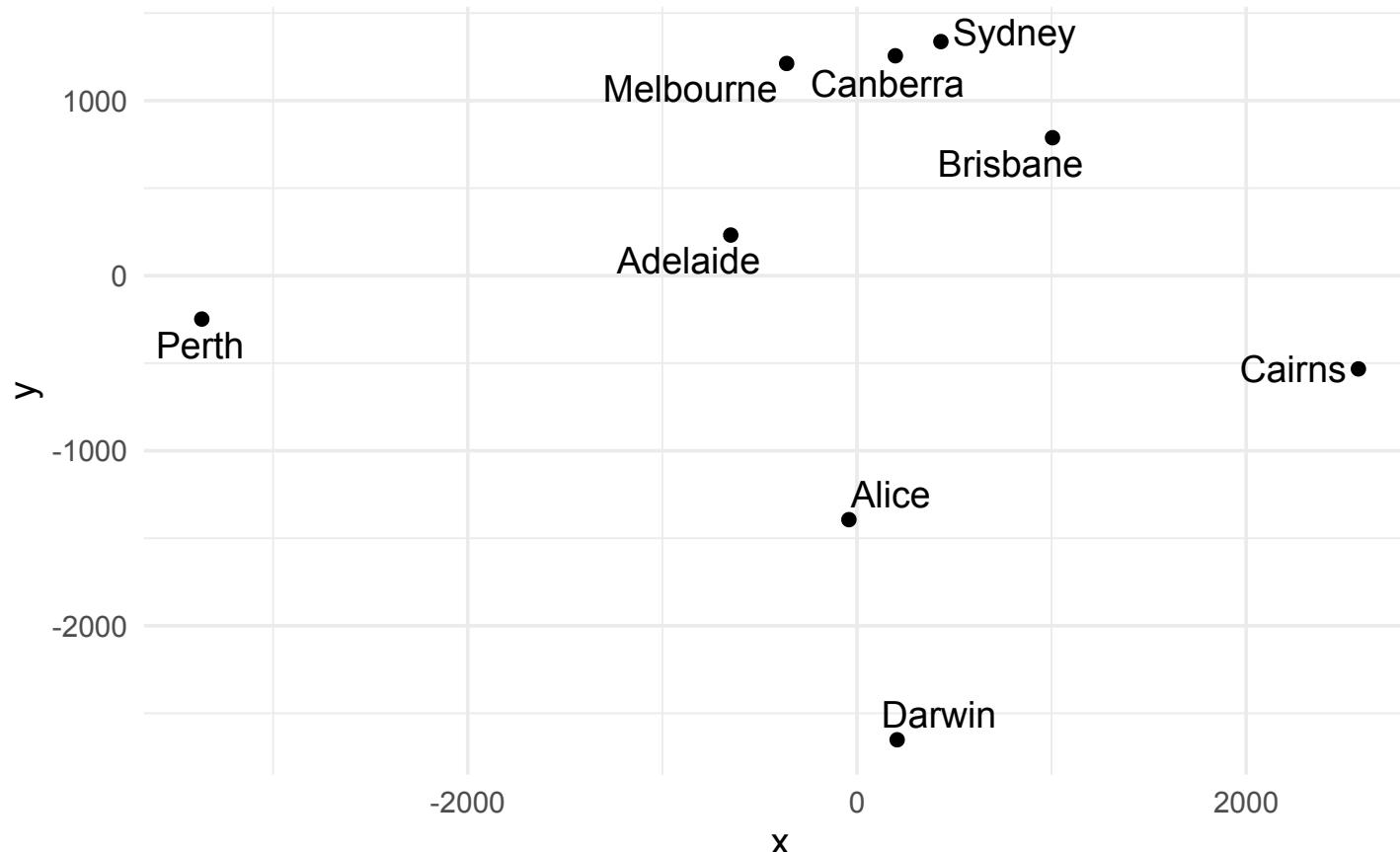
不需要完整的数据

Multidimensional Scaling (MDS) Example

	Adelaide	Alice	Brisbane	Cairns	Canberra	Darwin	Melbourne	Perth	Sydney
Adelaide	0	1533	2044	3143	1204	3042	728	2725	1427
Alice	1533	0	3100	2500	2680	1489	2270	3630	2850
Brisbane	2044	3100	0	1718	1268	3415	1669	4384	1010
Cairns	3143	2500	1718	0	2922	3100	3387	5954	2730
Canberra	1204	2680	1268	2922	0	3917	647	3911	288
Darwin	3042	1489	3415	3100	3917	0	4045	4250	3991
Melbourne	728	2270	1669	3387	647	4045	0	3430	963
Perth	2725	3630	4384	5954	3911	4250	3430	0	4110
Sydney	1427	2850	1010	2730	288	3991	963	4110	0

Multidimensional Scaling (MDS) Example

```
mds <- cmdscale(city.dist, k = 2); colnames(mds) <- c("x", "y")
mds <- data.frame(mds, City = colnames(city.dist))
ggplot(mds, aes(x = x, y = y, label = City)) + geom_point() + ggrepel::geom_text_repel() + theme_minimal()
```



MDS Stress functions

- These functions attempt to force the lower dimensional projections to preserve the distances in the original data.
这些函数试图迫使低维投影保留原始数据中的距离
- Common stress functions

- Least squares $S_{LS}(z_1, z_2) = \sqrt{\sum_{i \neq j} (d_{ij} - \|z_i - z_j\|)^2}$

MDS Benefits/Drawbacks

- Full data not required, only its distance or dissimilarity matrix
- Need to choose K (could use the same elbow in Scree plot technique)
- Can be used as a visualization technique for non-linear data.

不需要完整的数据，只需要它的距离或异质性矩阵

需要选择K

可以作为非线性数据的可视化技术来使用

Interpreting MDS outputs

- Interpreting MDS maps:
 - Can be rotated (axes and orientation are somewhat arbitrary).
 - Only relative locations important.
 - Typically look for objects close in the MDS map

可以旋转（坐标轴和方向有点随意）。
只有相对位置重要。
通常在MDS地图中寻找接近的物体

References

- Hastie, T, R. Tibshirani, and J. Friedman (2017). *The elements of statistical learning: data mining, inference, and prediction*. Second Edition, 12th printing. Springer Science & Business Media.
- James, G, D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

STAT5003

Week 5 : Introduction to classification techniques

Dr. Justin Wishart



Readings



- Classification covered in Chapter 4 in James, Witten, Hastie, and Tibshirani (2013)
- Support Vector Machines covered in Chapter 9 in James, Witten, Hastie, et al. (2013)
- **Optional** for SVMs
 - Section 4.5.2 and Sections 14.1-14.3 in Hastie, Tibshirani, and Friedman (2017)

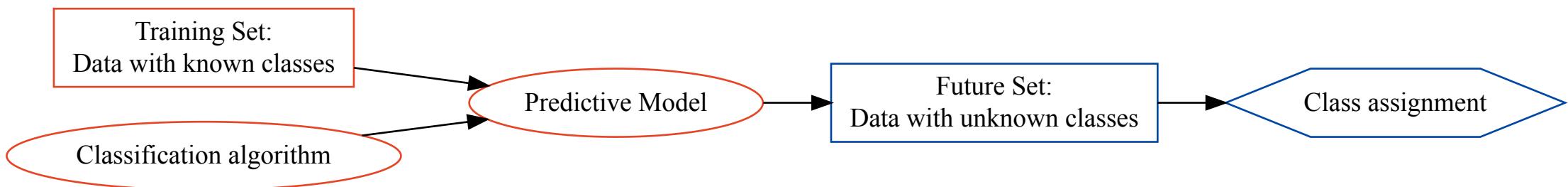
Classification



THE UNIVERSITY OF
SYDNEY

Basic principles of classification

- Each observation has two properties
 - A class label or response, y
 - A feature vector (vector of predictor variables), $\mathbf{x} = (x_1, x_2, \dots, x_p)$
- Goal is to classify y using \mathbf{x}



training set 和 classification model 结合生成 predictive model 然后用 future set 生成 class assignment

Classification vs Clustering 聚类是无监督， 分类是有监督

Clustering: classes are unknown, want to discover them from the data (unsupervised)

Classification: classes are predefined, want to use a (training or learning) set of labeled objects to form a classifier for classification of future observations (supervised)

聚类：类是未知的，想从数据中发现它们（无监督的unsupervised）。

分类：类是预先设定好的，想用一个（训练或学习）标记的对象集来形成一个分类器来对未来的观察进行分类（有监督(supervised)

Classification vs Regression

Regression: no class definition, the response variable is a continuous value. Model the relationship between explanatory variables and the response variable.

Classification: samples are predefined to be from a given class. Classification models produce a continuous valued prediction, which is usually in the form of a probability (i.e. the predicted values of class membership for any individual sample are between 0 and 1 and sum to 1). A predicted class is required in order to make a decision.

回归：没有类别定义，响应变量是一个连续值。对解释变量和响应变量之间的关系进行建模 解释变量和响应变量之间的关系。

分类：样本被预先定义为来自一个特定的类别。分类模型产生一个连续值的预测，它通常是以概率的形式出现的（即任何一个样本的类属预测值都介于类成员资格的预测值在0和1之间，总和为1） 预测的类别是为了做出决定，需要一个预测的类别

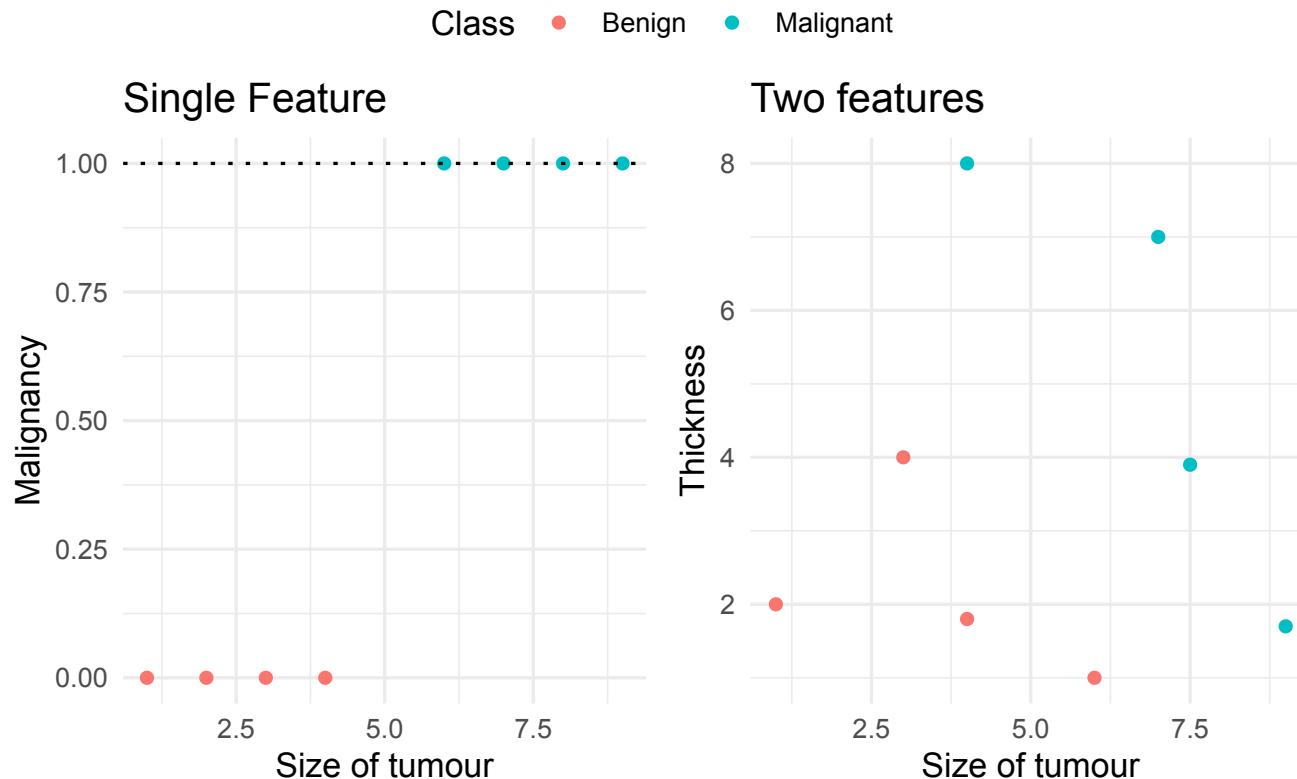
Classification algorithms to discuss

- Logistic Regression
- Linear discriminant analysis (LDA)
- k -nearest neighbours
- Support vector machines (SVM)

Binary or Two class classification

- Binary in there are two possible values (0 or 1, TRUE or FALSE)
- Examples of binary classification:
 - Email: Spam / Not Spam
 - Tumour: Malignant /Benign
- Labels are similarly described, $y \in \{0, 1\}$
 - 0: "negative class"
 - 1: "positive class"

Problem setup

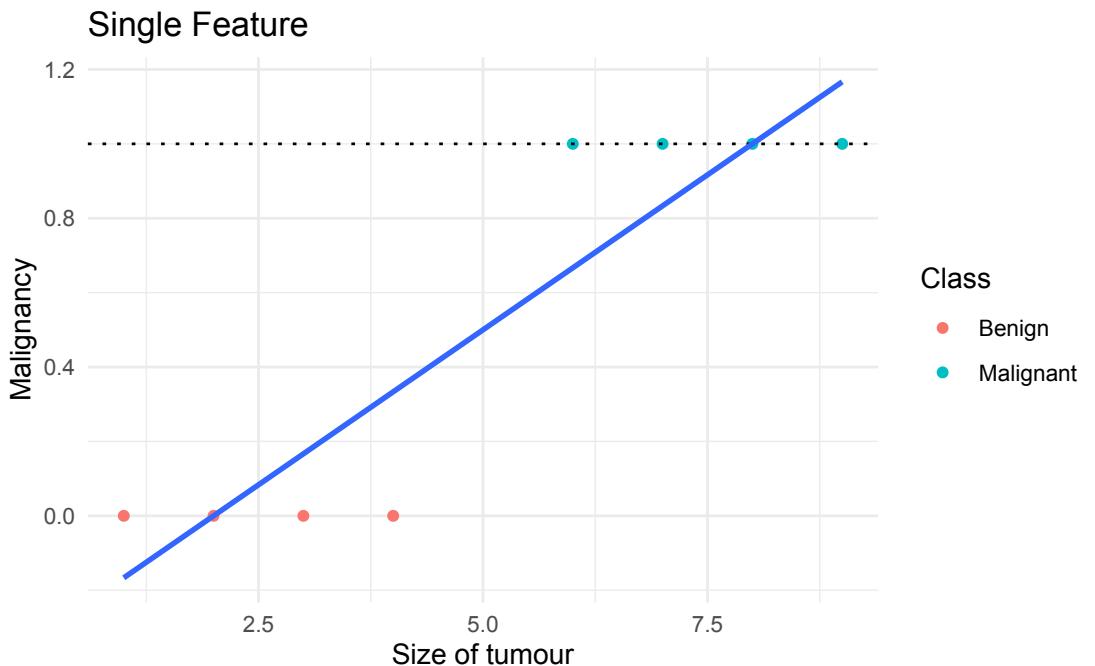


- Threshold classifier output $h_\theta(x)$ at 0.5:

- if $h_\theta(x) > 0.5$, predict $y = 1$
 - if $h_\theta(x) < 0.5$, predict $y = 0$

Why not use simple linear regression?

- Y is the target value is a *binary* outcome.



线性回归不能使用因为不是0或1

- Linear regression is not constrained to $0 < y < 1$ for all x
 - What is the interpretation when $\hat{y} > 1$ (or < 0)

Linear regression misspecifications here

- The regression line $\beta_0 + \beta_1 x$ can span the entire real line
 - all values between $-\infty$ to ∞
- In the tumour diagnosis problem, the target variable y only takes two values: 0 or 1.
- The linear regression model is not well specified for this purpose.

参数机器学习算法包括: 逻辑回归, 线性成分分析 (PCA)

优点:

简洁: 理论容易理解和解释结果, 快速: 参数模型学习和训练的速度都很快

数据更少: 通常不需要大量的数据, 在对数据的拟合不很好时表现也不错

局限性:

约束: 以选定函数形式的方式来学习本身就限制了模型, 有限的复杂度: 通常只能应对简单的问题,

拟合度小: 实际中通常无法和潜在的目标函数吻合

非参数机器学习算法decision tree, Bayes, SVM

优势:

可变性: 可以拟合许多不同的函数形式。模型强大: 对于目标函数不作假设或者作微小的假设, 表现良好:

对于预测表现可以非常好。

局限性:

需要更多数据: 对于拟合目标函数需要更多的训练数据, 速度慢: 因为需要训练更多的参数, 训练过程通常比较慢。过拟合Overfitting: 有更高的风险发生过拟合, 对于预测也比较难以解释

Logistic regression



THE UNIVERSITY OF
SYDNEY

Logistic regression (需要数值型变量)

- Previously we had the multiple regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

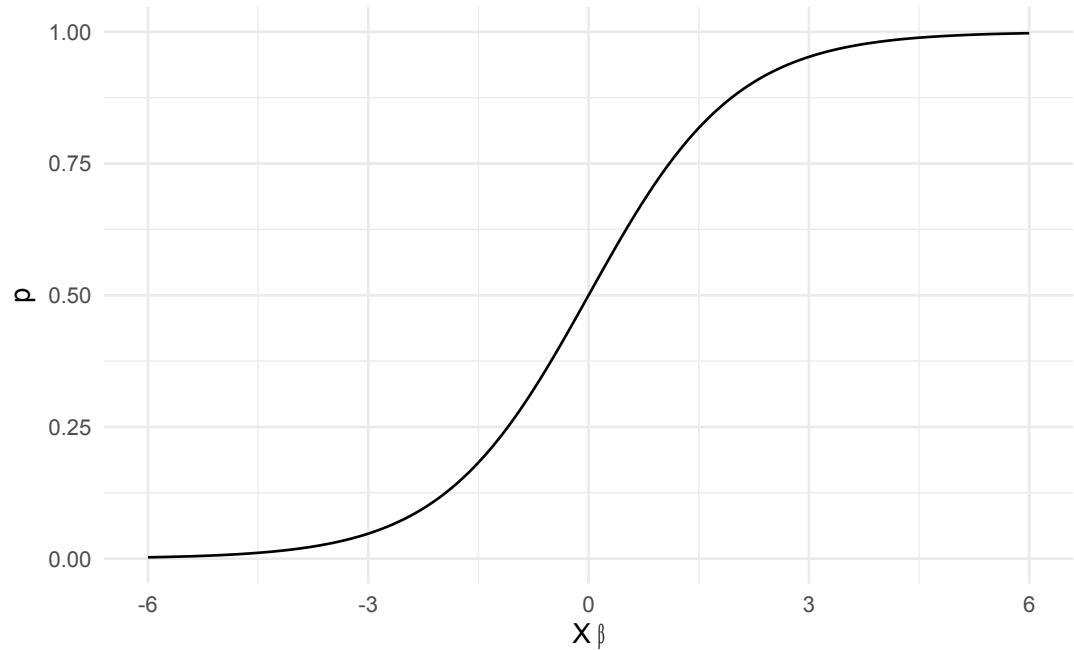
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Define $\boldsymbol{\theta}^T \mathbf{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

- Could write this as $\mathbb{E}Y = \boldsymbol{\theta}^T \mathbf{x} = \mu$
- Can **generalise** this to $g(\mathbb{E}Y) = \boldsymbol{\theta}^T \mathbf{x} = g(\mu)$
- Logistic regression is a special case of one of these generalised linear models.
- $\log\left(\frac{p}{1-p}\right) = \boldsymbol{\theta}^T \mathbf{x}$
- Solve for p gives

o

$$p = P(Y = 1 | \mathbf{x}) = g^{-1}(\boldsymbol{\theta}^T \mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x})}$$



Logistic regression terminology

- Logistic function $\frac{1}{1+e^{-\theta^T x}}$
 - Responsible from mapping the features from $(-\infty, \infty) = \mathbb{R}$ to $(0, 1)$
- Odds ratio: $\frac{p}{1-p}$ 赔率
 - Maps the probability from $(0, 1)$ to $(0, \infty)$
- Log-odds or logit: $\log\left(\frac{p}{1-p}\right)$
- In logistic regression we want the values in the logit space to be linear in X

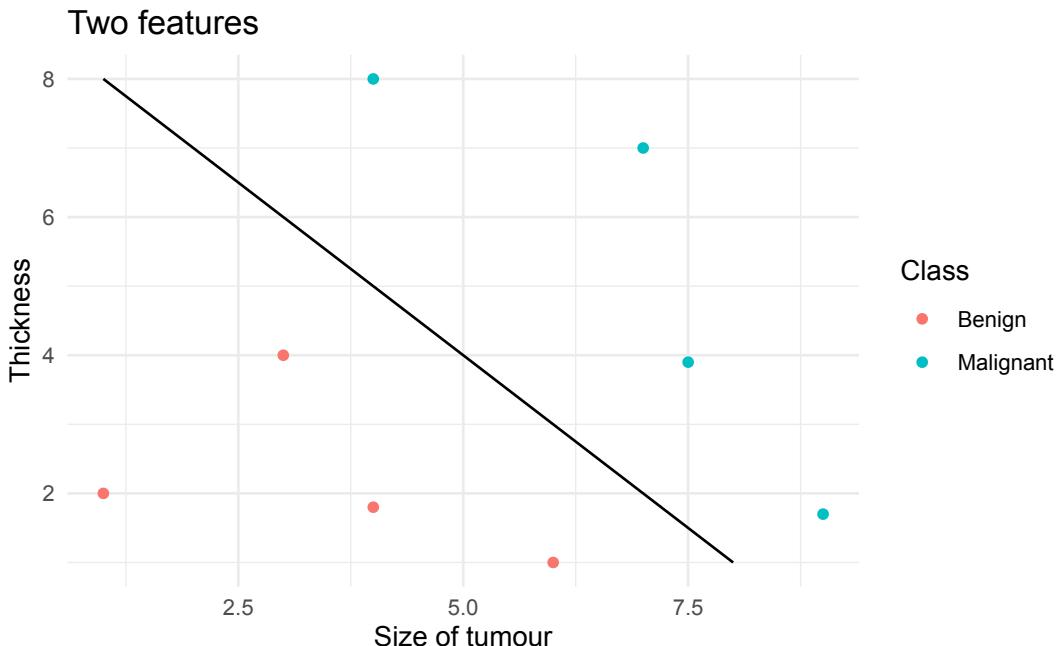
Logistic regression: decision boundary

- Decision boundary

$$P(Y = 1 | \mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x})}$$

决策边界，也称为决策面，是用于在N维空间，将不同类别样本分开的平面或曲面。

- Predict $Y = 1$ if $\boldsymbol{\theta}^T \mathbf{x} \geq 0$



Linear Discriminant Analysis (LDA)



THE UNIVERSITY OF
SYDNEY

Linear Discriminant Analysis (LDA)

LDA undertakes the same task as Logistic Regression. It classifies data based on categorical variables

- Malignant or benign
- Making profit or not
- Buy a product or not
- Satisfied customer or not

Bayes' Theorem in the classification context

$$p_k(x) = P(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^K \pi_\ell f_\ell(x)}$$

Posterior: The probability of classifying observation to group k given it has features x

Prior: The prior probability of an observation in general belonging to group k

- $f_k(x) = P(X = x | Y = k)$ is the density function for feature x given it's in group k

Logistic Regression vs LDA formulations

- With Logistic Regression we modeled the probability of Y being from the k^{th} class as

$$p_k(x) = P(Y = k|X = x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- Bayes' Theorem states

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^K \pi_\ell f_\ell(x)}$$

π_k : Probability of coming from class k (prior probability)

$f_k(x)$: Density function for X given that X is an observation from class k .

LDA estimates of π_k and $f_k(x)$

- We can estimate π_k and $f_k(x)$ to compute $p_k(x)$
- The most common model for $f_k(x)$ is the Normal Density (LDA)

$$f_k(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma^2}\right)$$

- Using the above density, we only need to estimate three quantities to compute $p_k(x)$
 - That is, μ_k , σ_k^2 and π_k
- For simplicity, assume common variance.

Use training data set for estimation

- The mean $\widehat{\mu}_k$ could be estimated by the average of all training observations from the k^{th} class.
- The variance σ^2 could be estimated as the weighted average of variances of all k classes.
- The proportion π_k is estimated as the proportion of the training observations that belong to the k^{th} class.

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\sigma^2 = \frac{1}{n - K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \widehat{\mu}_k)^2$$

Simple example with one predictor

- Suppose we have only one predictor
- Two normal density function $f_1(x)$ and $f_2(x)$, represent two distinct classes
- The two density functions overlap, so there is some uncertainty about the class to which an observation with an unknown class belongs
- The dashed vertical line represents Bayes' decision boundary

Deriving LDA for one predictor

- Assuming one predictor (and common variance)

$$f_k(x) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

$$p_k(x) = P(Y = k|x) = \frac{\frac{\pi_k}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)}{\sum_{l=0}^K \frac{\pi_l}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)}$$

- Find the class k which we maximize:

$$x \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

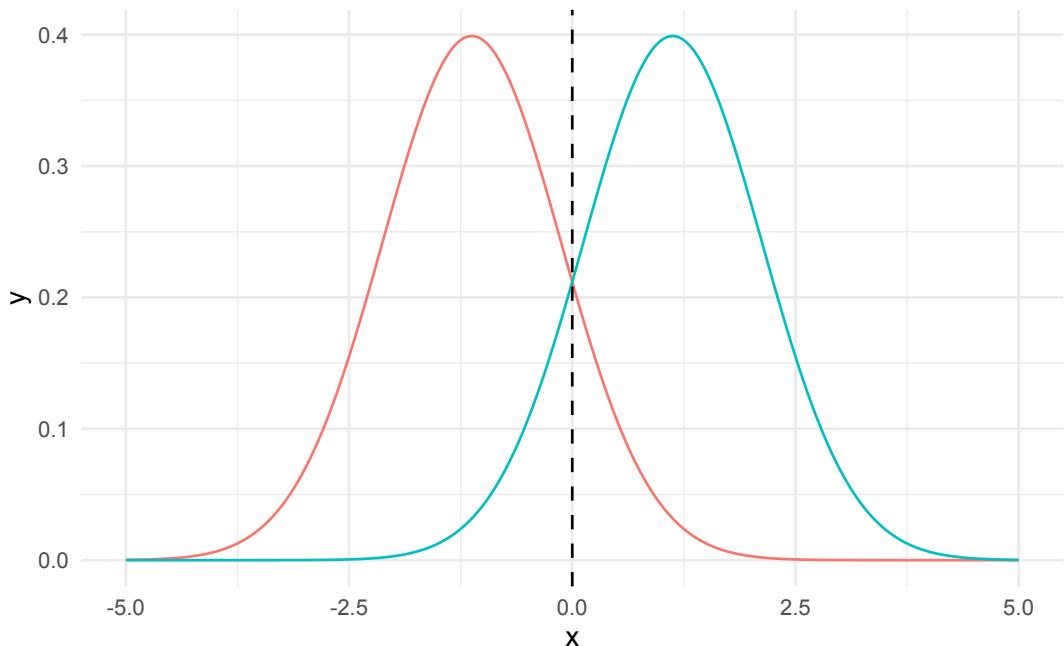
LDA Decision boundary

If $K = 2$ and $\pi_1 = \pi_2$, then assigns an observation to class 1 if $\log p_1(x) > \log p_2(x) \rightsquigarrow \log\left(\frac{p_1(x)}{p_2(x)}\right) > 0$

Substituting in the previous equation (assuming $\sigma_i = \sigma$) we have,

$$\log\left(\frac{p_1(x)}{p_2(x)}\right) > 0$$

$$\log(\pi_1) - \log(\pi_2) + \frac{x\mu_1}{2\sigma^2} - \frac{x\mu_2}{2\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \frac{\mu_2^2}{2\sigma^2} > 0$$
$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$$



- Decision boundary at

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

Why not logistic regression?

- In the case where n is small, and the distribution of predictors X is approximately normal, then LDA is more stable than Logistic Regression
- LDA is more popular when we have more than two response classes. More intuitive to predict class assignment.
- When the classes are well separated, the parameter estimates for logistic regression are unstable. However, LDA doesn't suffer any stability issues in this case.

LDA 比 logistic 的优势

在 n 很小的情况下，预测因子 X 的分布近似于正态，那么 LDA 是比 Logistic 回归更稳定

当我们有两个以上的响应类时，LDA 更受欢迎。更加直观地预测类分配。

当类被很好地分开时，逻辑回归的参数估计是不稳定的。然而，LDA 在这种情况下不会出现任何稳定性问题。逻辑回归 VS LDA

Logistic Regression vs LDA

Similarity:

- Both Logistic Regression and LDA produce linear boundaries

Differences:

- LDA assumes that the observations are drawn from the normal distribution with common variance in each class, while logistic regression does not have this assumption.
- LDA would do better than Logistic Regression if the assumption of normality hold, otherwise logistic regression may outperform LDA

相似性。

Logistic回归和LDA都能产生线性边界linear boundaries

不同之处。

LDA假设观测值来自于正态分布normal distribution，每一类都有共同的方差。

而逻辑回归则没有这个假设assumption。

如果正态性假设成立，LDA会比Logistic Regression做得更好，否则Logistic回归可能优于LDA

k -Nearest Neighbours (kNN)



THE UNIVERSITY OF
SYDNEY

k -Nearest Neighbours

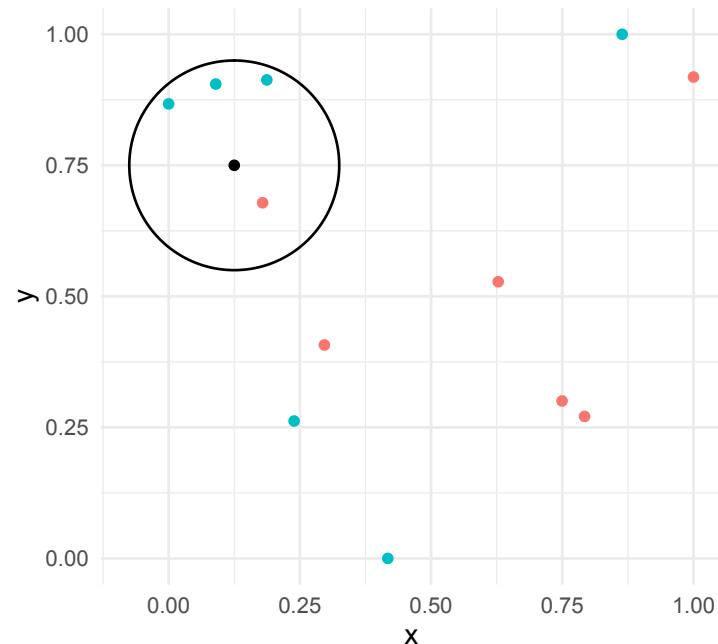
- kNN model is probability of an observation with features \mathbf{x} belonging to group ℓ depends on the membership of the nearest points to \mathbf{x}

$$P(Y = \ell | \mathbf{x}) = \frac{1}{k} \sum_{N_x^k} 1_{\{y=\ell\}} = \frac{1}{k} \times \text{Count of the closest } k \text{ points that belong to group } \ell$$

- Suppose $k = 4$ is chosen. At the candidate black point. The four nearest neighbours are inspected. There is probability 3/4 of being in the green group and 1/4 for being in the orange group.

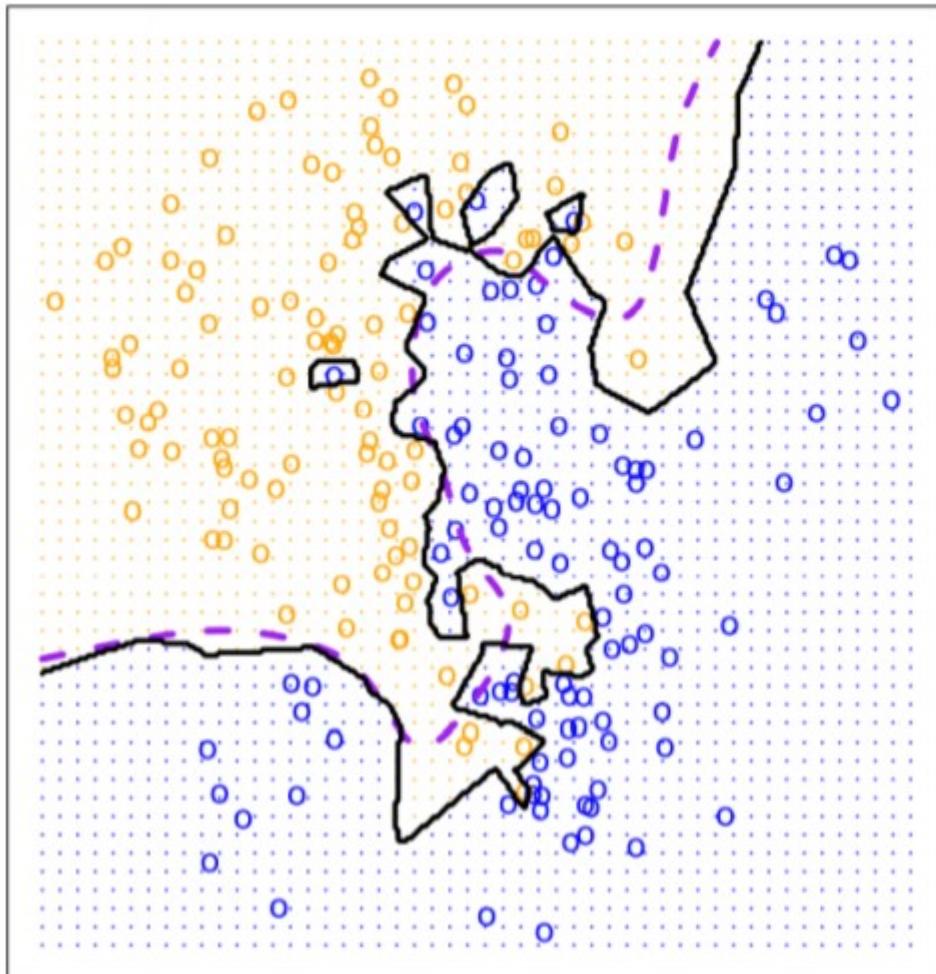
如果 $k=4$ 就看与 x 点最近4个点的比例来判断分类

kNN模型是指一个具有特征的观察值属于组的概率取决于与 x 最近的点的成员资格

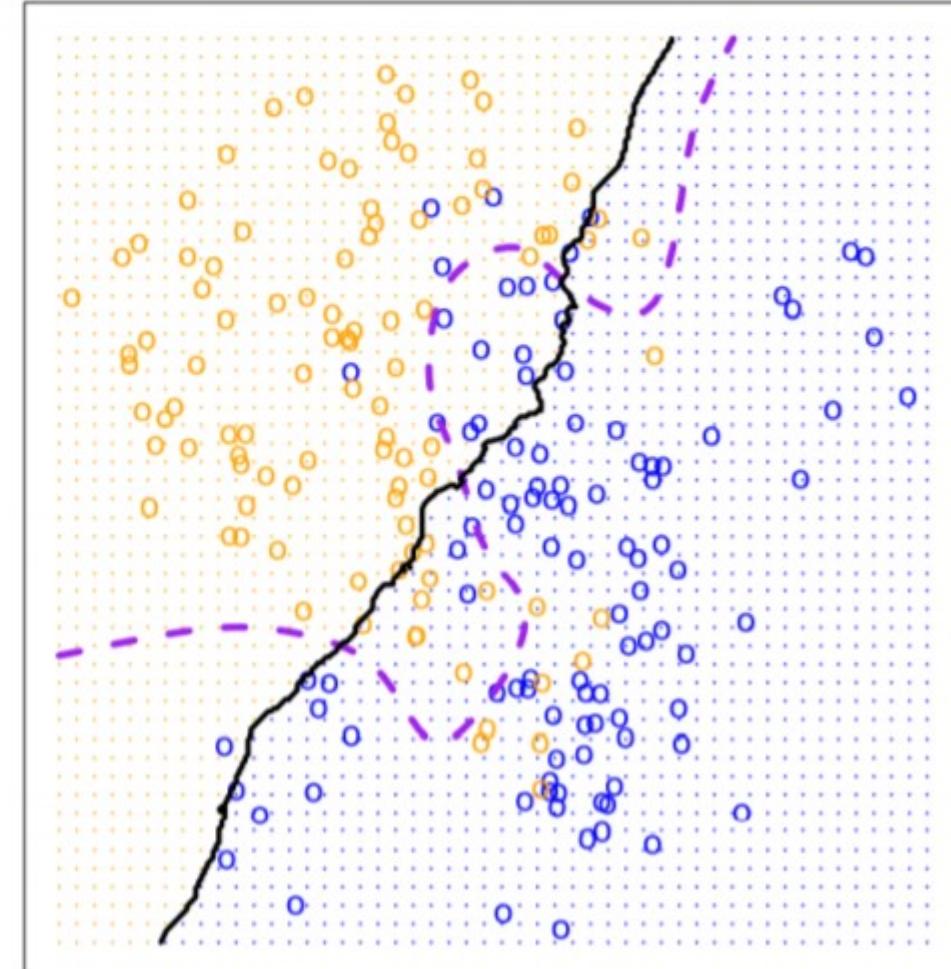


k -Nearest Neighbours

KNN: K=1



KNN: K=100



kNN vs (LDA and Logistic Regression)

- kNN takes a completely different approach
- kNN is completely non-parametric: No assumptions are made about the shape of the decision boundary
- Advantage of kNN: We can expect kNN to dominate both LDA and Logistic Regression when the decision boundary is highly non-linear
- Disadvantage of kNN: kNN does not tell us which predictors are important (no table of coefficients)

kNN是完全非参数化parametric的。对决策边界的形状不做任何假设。

kNN的优势。当决策边界高度非线性时，我们可以预期kNN会在LDA和Logistic Regression中占优势。决策边界是高度非线性的

kNN的缺点：kNN不能告诉我们哪些预测因子是重要的（没有协方系数表）

Support Vector Machines (SVM)



THE UNIVERSITY OF
SYDNEY

Support Vector Machines (SVM)

- Basic idea behind SVM

Find a plane that separates the classes in the feature space.

- If a basic mathematical plane is not possible due to overlap
 - Relax the idea of complete separation (allow points to violate the boundary)
 - Enrich and enlarge the feature space so that separation is possible
 - Think dimension expansion

找到一个能在特征空间中分离出各个类别的平面。

如果由于重叠而不可能有一个基本的数学平面

放松完全分离的想法（允许点违反边界）。

丰富和扩大特征空间，使分离成为可能

考虑维度扩展

What is a hyperplane?

- In p dimensions it is a flat affine subspace of dimension $p - 1$
- General equation has the form

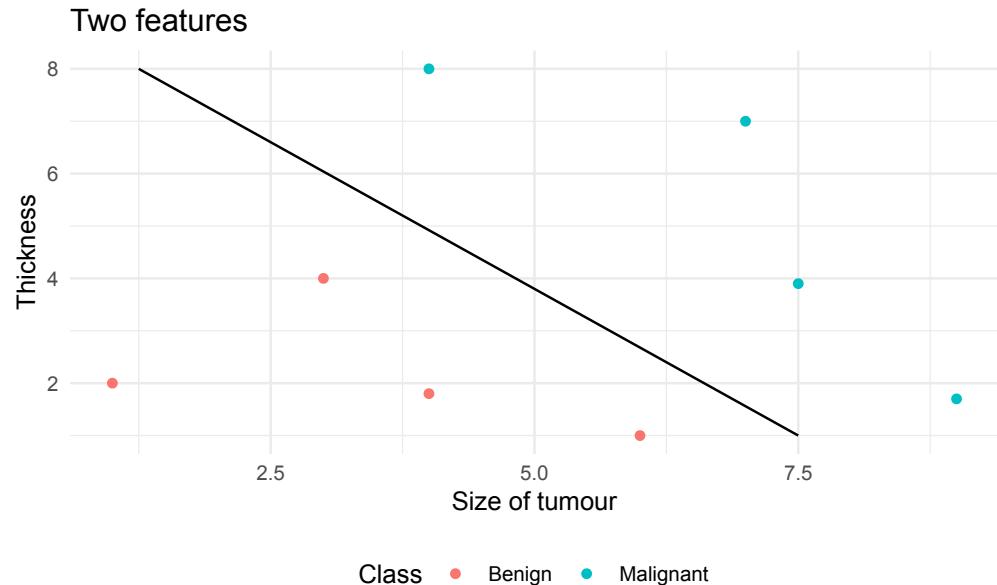
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

- In $p = 2$ dimensions, the hyperplane is a line.
- If $\beta_0 = 0$, the hyperplane passes through the origin, otherwise it does not.
- The vector $(\beta_1, \beta_2, \dots, \beta_p)$ is called the normal vector
 - It points in a direction orthogonal to the surface of the hyperplane

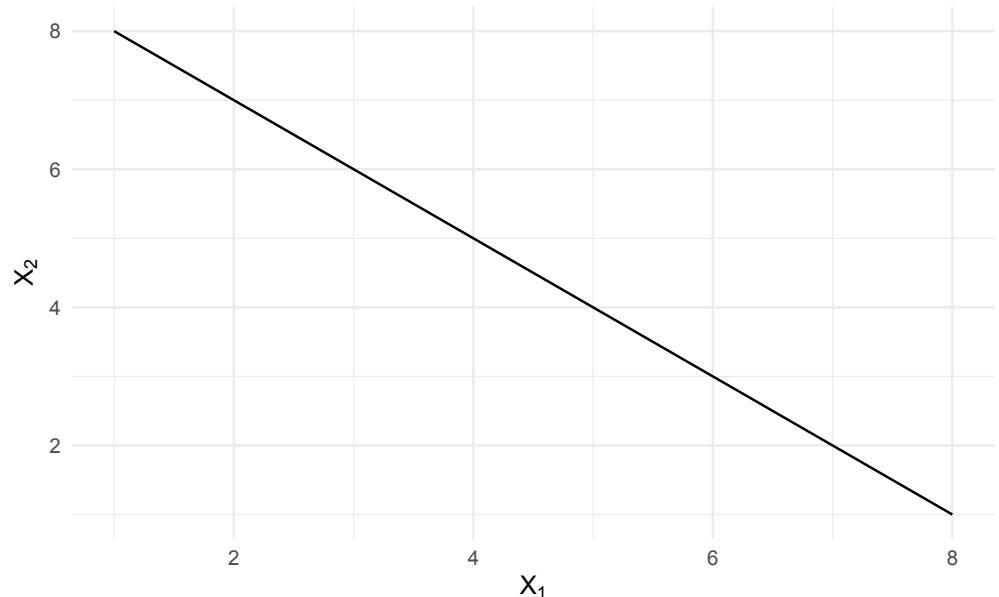
在维数为 p 时，它是一个维数为 $p-1$ 的仿生子空间 当 $p=2$ 时，是一条line
 $Y_i = -1$ 和 $y_i = 1, f(x) > 0$ or $f(x) < 0$

Hyperplane example

- Earlier hypothetical example

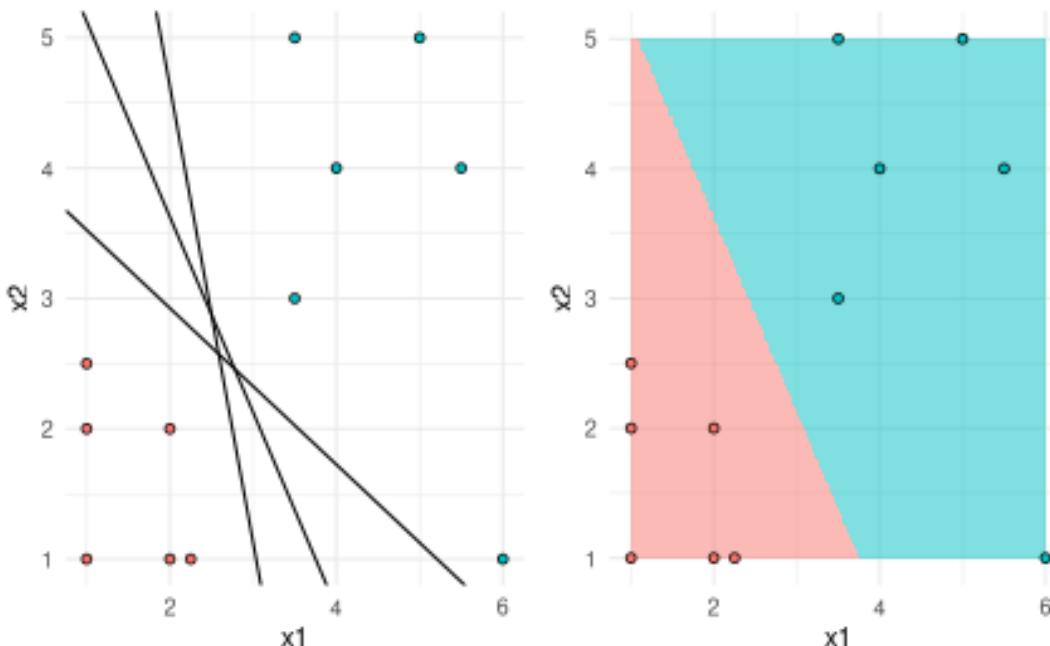


- Consider just the line (hyperplane)



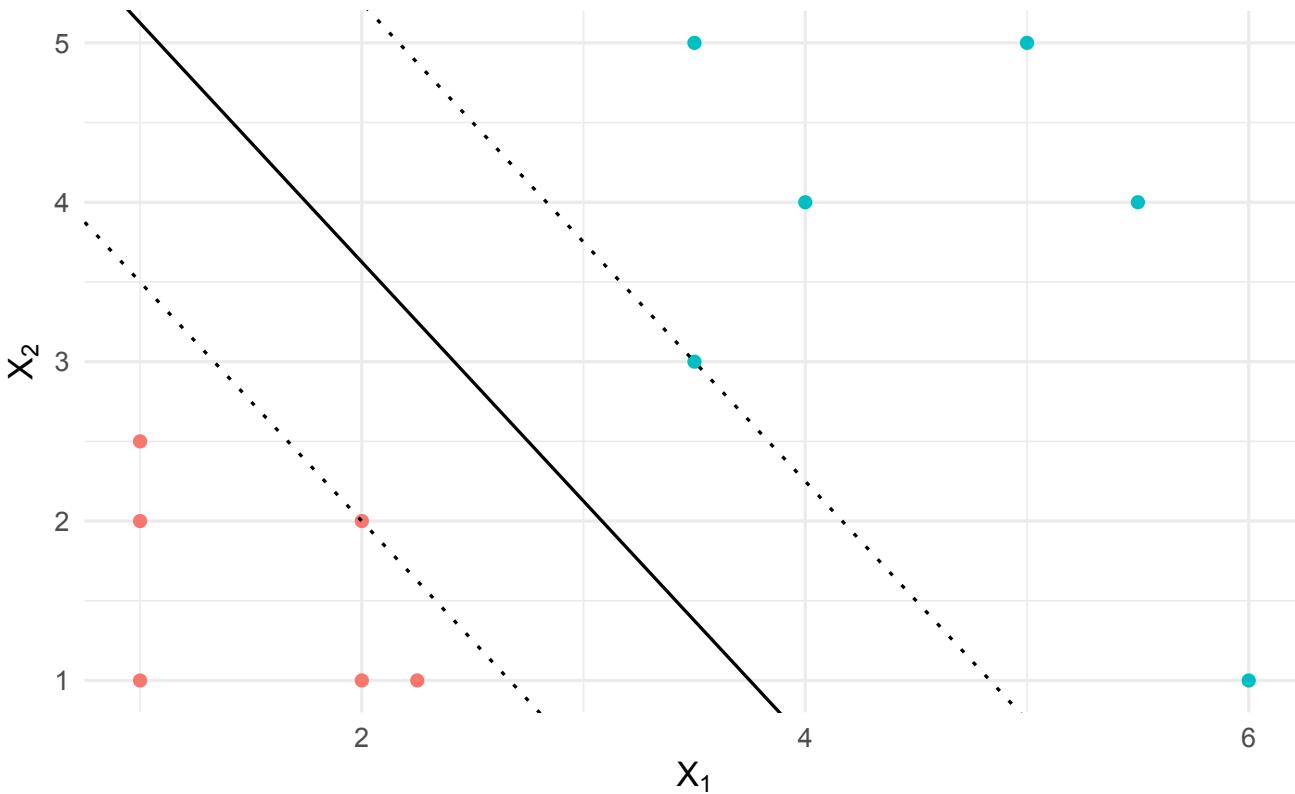
- Equation of the hyperplane here is $-9 + X_1 + X_2 = 0$

Separating hyperplanes



- Consider coding Benign (red?) as $y_i = -1$ and malignant (blue?) as $y_i = 1$
- Then $y_i f(x_i) > 0$ for all i , $f(x_i)$ defines a separating hyperplane.
- If $f(x_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ defines a hyperplane
 - $f(x) > 0$ defines a region on one side of the hyperplane
 - $f(x) < 0$ defines a region on one other side of the hyperplane

Maximal Margin Classifier

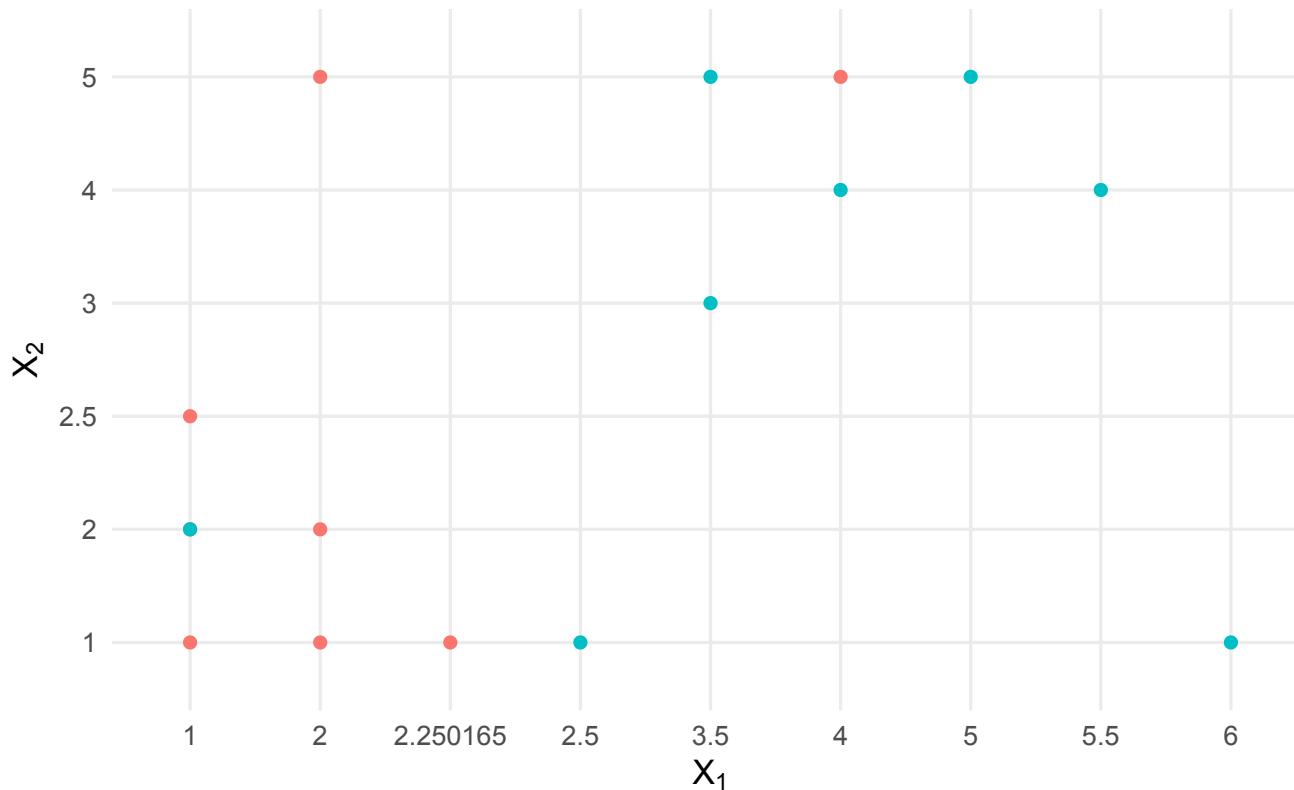


最大化margin， margin
为两线之间的距离

$$\max_{\beta_0, \beta_1, \beta_2, \dots, \beta_p} M \quad \text{such that} \quad \sum_{j=1}^p \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M$$

Non-perfect separation

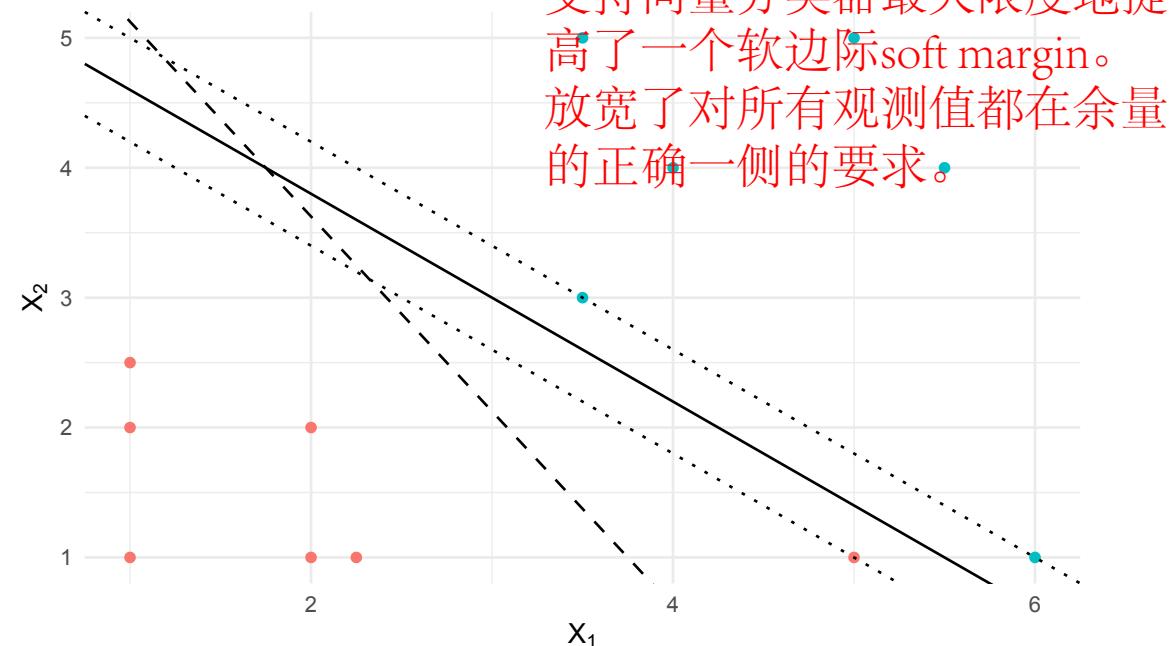
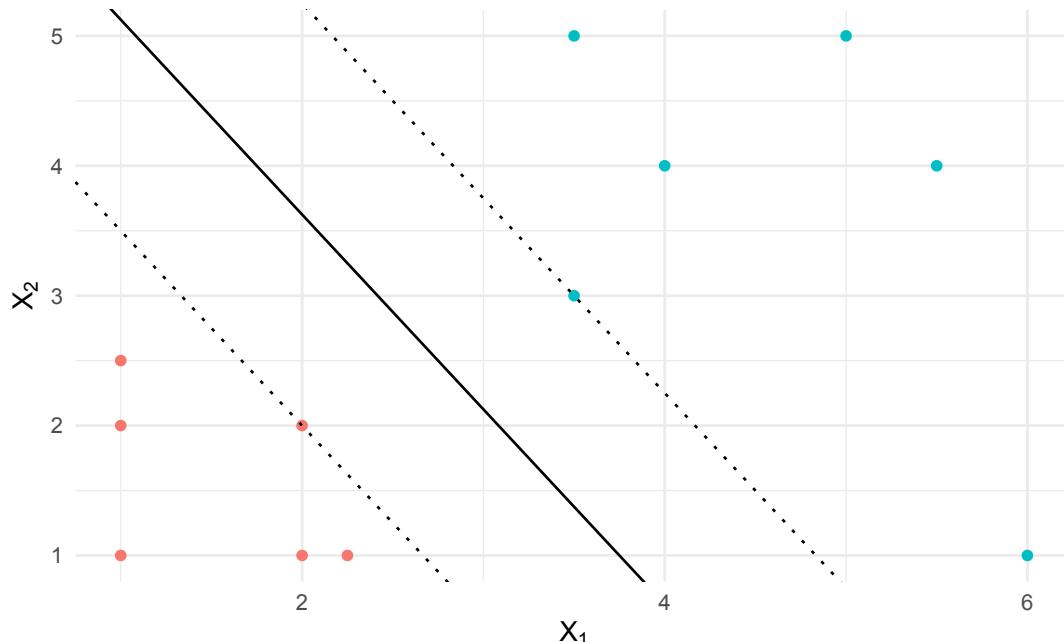


没有一个线性边界（超平面）linear boundary (hyperplane)能完美地分离出各个类别。这通常是指观测值没有一个完美的分离边界的情况。除了在 $n < p$ 的情况下（特征多于观测值之外）

- There is no linear boundary (hyperplane) that perfectly separates the classes.
- This is typically the case that observations don't have a perfect boundary of separation.
 - Except in the case when $n < p$ (more features than observations)

Effect of noisy data

Consider the impact of one extra observation

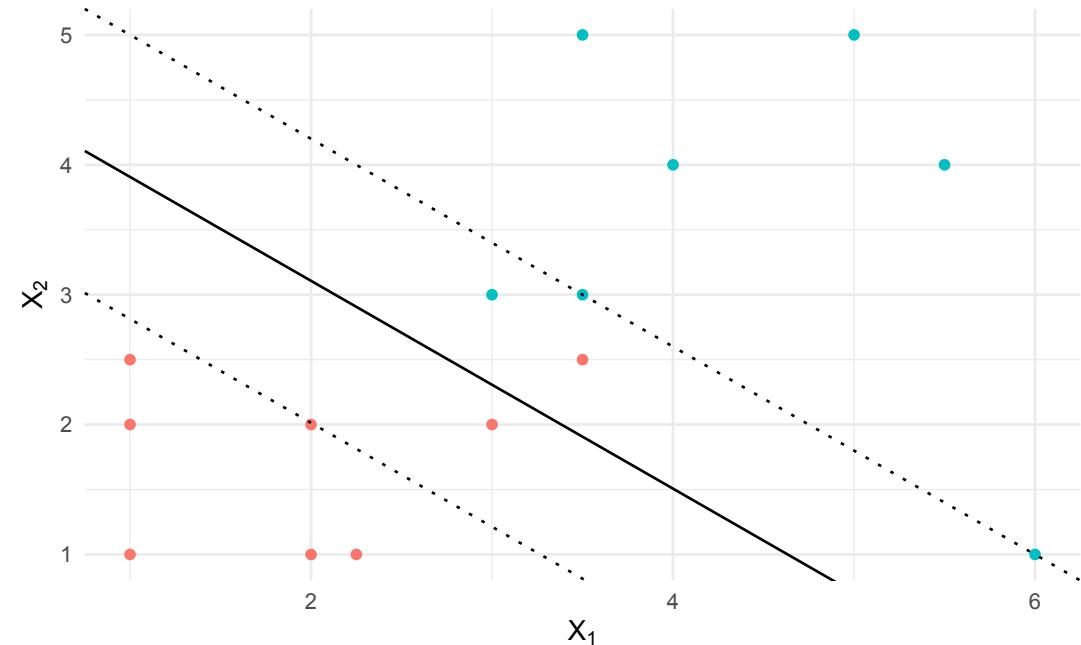
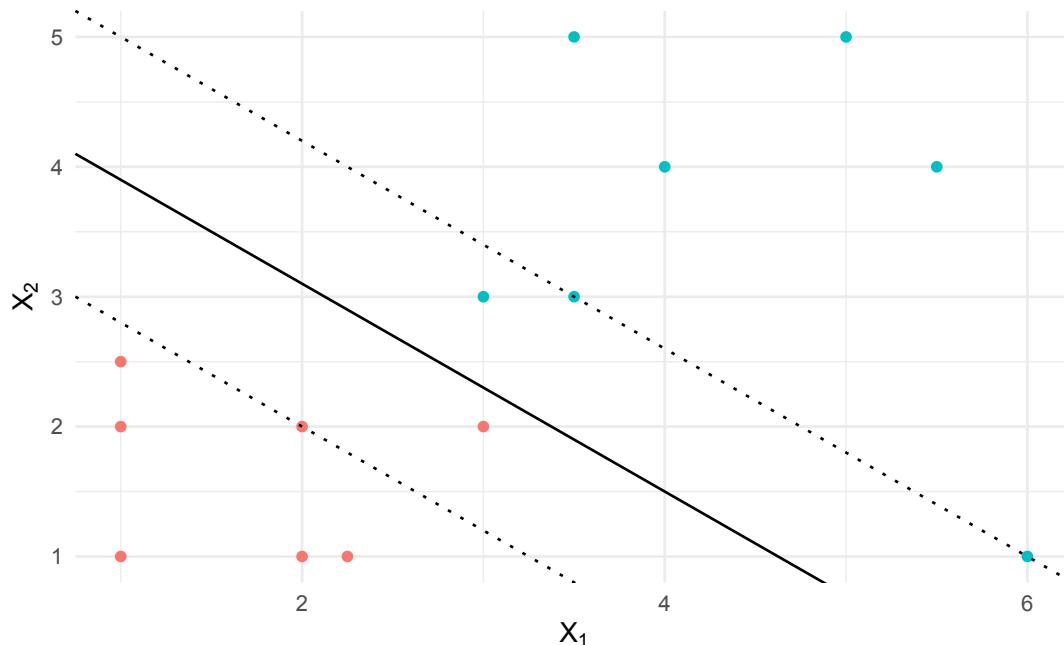


数据可能是可分离的，但
noisy的不稳定的解决方案的
maximal margin classifier。
支持向量分类器最大限度地提
高了一个软边际soft margin。
放宽了对所有观测值都在余量
的正确一侧的要求。

- Data could be separable, but noisy \rightsquigarrow unstable solution for the maximal margin classifier.
- The support vector classifier maximizes a soft margin.
 - relaxes requirement for all observations to be on the correct side of the margin

Soft margin examples

- Observations allowed in the margin



- Observations on the correct side of hyperplane

- Allow observations on incorrect side of hyperplane

Support Vector Classifier

Support Vector Classifier solves the following optimization problem:

$$\max_{\beta_0, \beta_1, \beta_2, \dots, \beta_p, \epsilon_1, \dots, \epsilon_n} M \quad \text{such that} \quad \sum_{j=1}^p \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

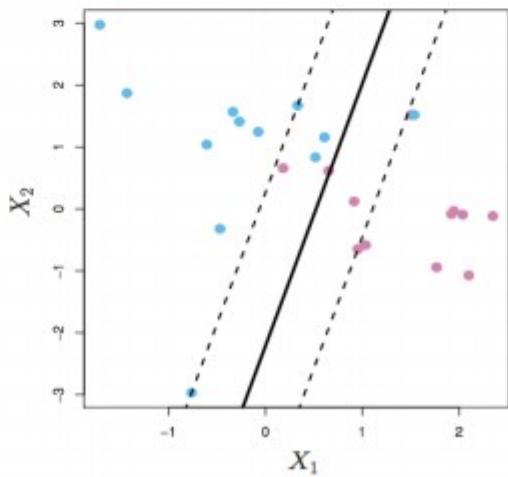
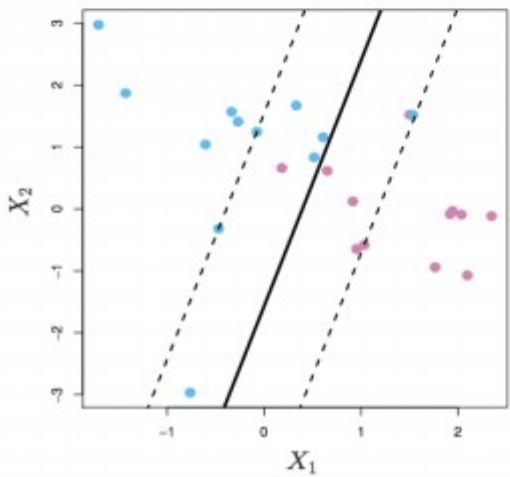
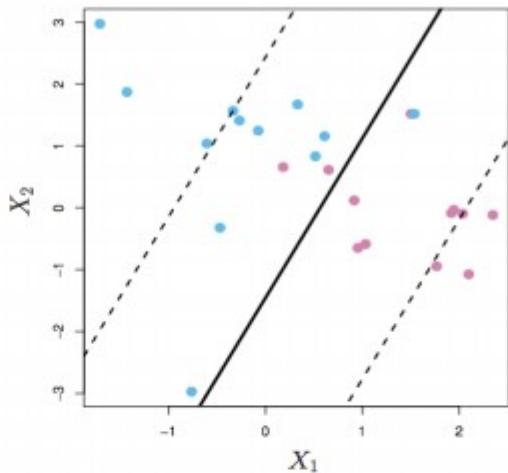
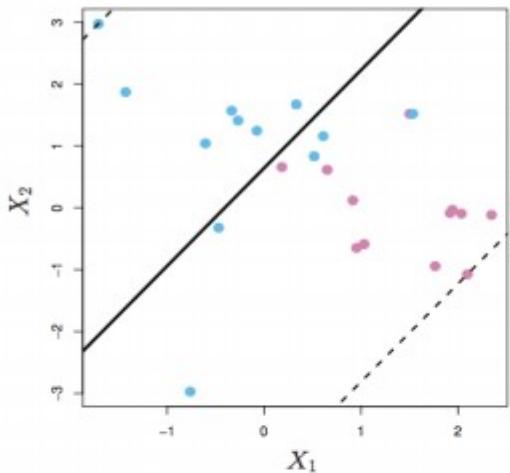
$$\epsilon_i \geq 0, \sum_{i=1}^n \epsilon_i \leq C$$

如果 $\epsilon_i > 1$ 则 i 在错边的超平面，如果 $0 < \epsilon_i \leq 1$ ，在正确的一侧，但在空白inside margin处，如果等于0，在正确的一侧，并越过边缘。

- C is a non-negative tuning parameter,
- M is the width of the margin,
- ϵ_i are slack variables that allow observations to be on the wrong side of the margin,
 - if $\epsilon_i > 1$, then observation i is on the wrong side of the hyperplane
 - if $0 < \epsilon_i \leq 1$, then observation i is on the correct side but inside margin
 - if $\epsilon_i = 0$, then observation i is on the correct side and past the margin.

Impact of cost parameter C

Largest C



Smallest C

Limitations of support vector classifier

- Single linear boundary can be insufficient

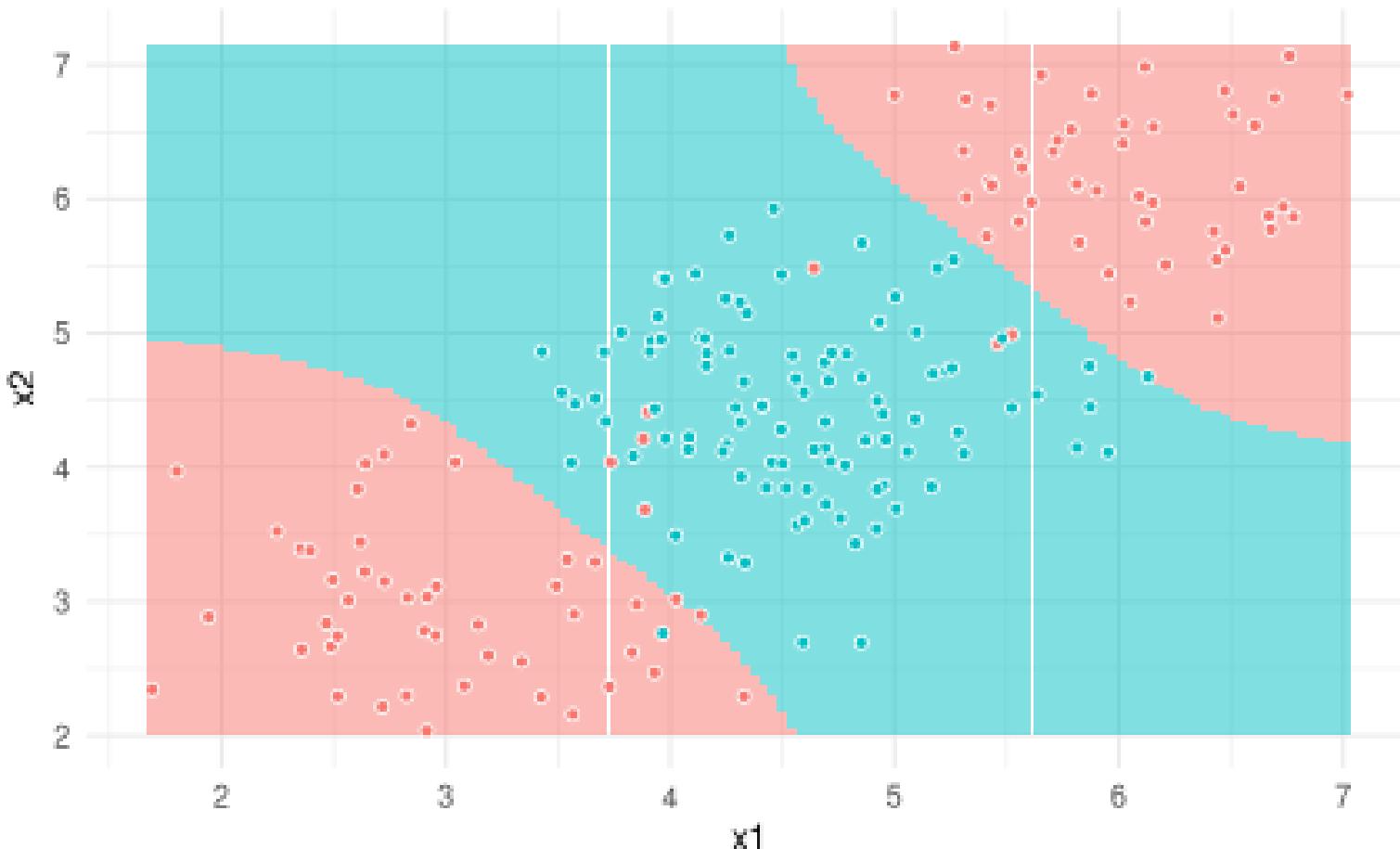
单一的线性边界可能是不够的

Feature space expansion

- Enlarge the space of features by including transformations:
 - e.g. new features that are powers and products X_1^2, X_1^3, X_1X_2
 - Hence go from p -dimensional space to $P > p$ dimensional
- Fit (linear) support vector classifier in the expanded feature space.
 - Impact is a **non-linear** decision boundary in original feature space.
- Example: Suppose we start off in 2-dimensional feature space (X_1, X_2) .
 - Make new feature space $(X_1, X_2, X_1^2, X_2^2, X_1X_2)$.
 - Then the decision boundary would be of the form:
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 = 0$$
- This leads to non-linear decision boundary in the original space (quadratic conic sections)

Sixth order polynomial

- Using degree six polynomial expansion



Non-linearity and kernels

- Polynomials get complicated and a burden very quickly as dimension increases.
- More elegant solution is to induce non-linear structure in Support vector classifier with **kernels**
- The elegance comes from the role of the **inner product** in the support vector classifier definition

随着维度的增加，多项式变得非常复杂，并成为一种负担。
更优雅的解决方案是在支持向量分类器中用核子诱导非线性结构
其优雅性来自于支持向量分类器中的内积的作用

Inner products and support vectors

- Recall $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$
- Inner product between vectors given by,

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \sum_{k=1}^p x_{ik} x_{jk}$$

- Recall the hyperplane equation

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- The linear support vector classifier can be represented as

$$f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^n \alpha_j \langle \boldsymbol{x}, \boldsymbol{x}_j \rangle$$

Support set

- Estimation of the parameters $\alpha_1, \alpha_2, \dots, \alpha_n$ and β_0 required
 - All that is required are the inner products between pairs of training observations.
- Usually $\hat{\alpha}_i = 0$ with the non-zero values occurring on the support vectors
 - I.e. ones that lie on the margin.

$$f(\mathbf{x}) = \beta_0 + \sum_{j \in S} \alpha_j \langle \mathbf{x}, \mathbf{x}_j \rangle$$

- Here, S is the support set of indices such that $\alpha_j > 0$

Kernel functions

Suppose now we replace the inner product with a generalized function of the form

$$K(\mathbf{x}_i, \mathbf{x}_j)$$

This function is called a **kernel**.

In this context it quantifies the similarity of two observations.

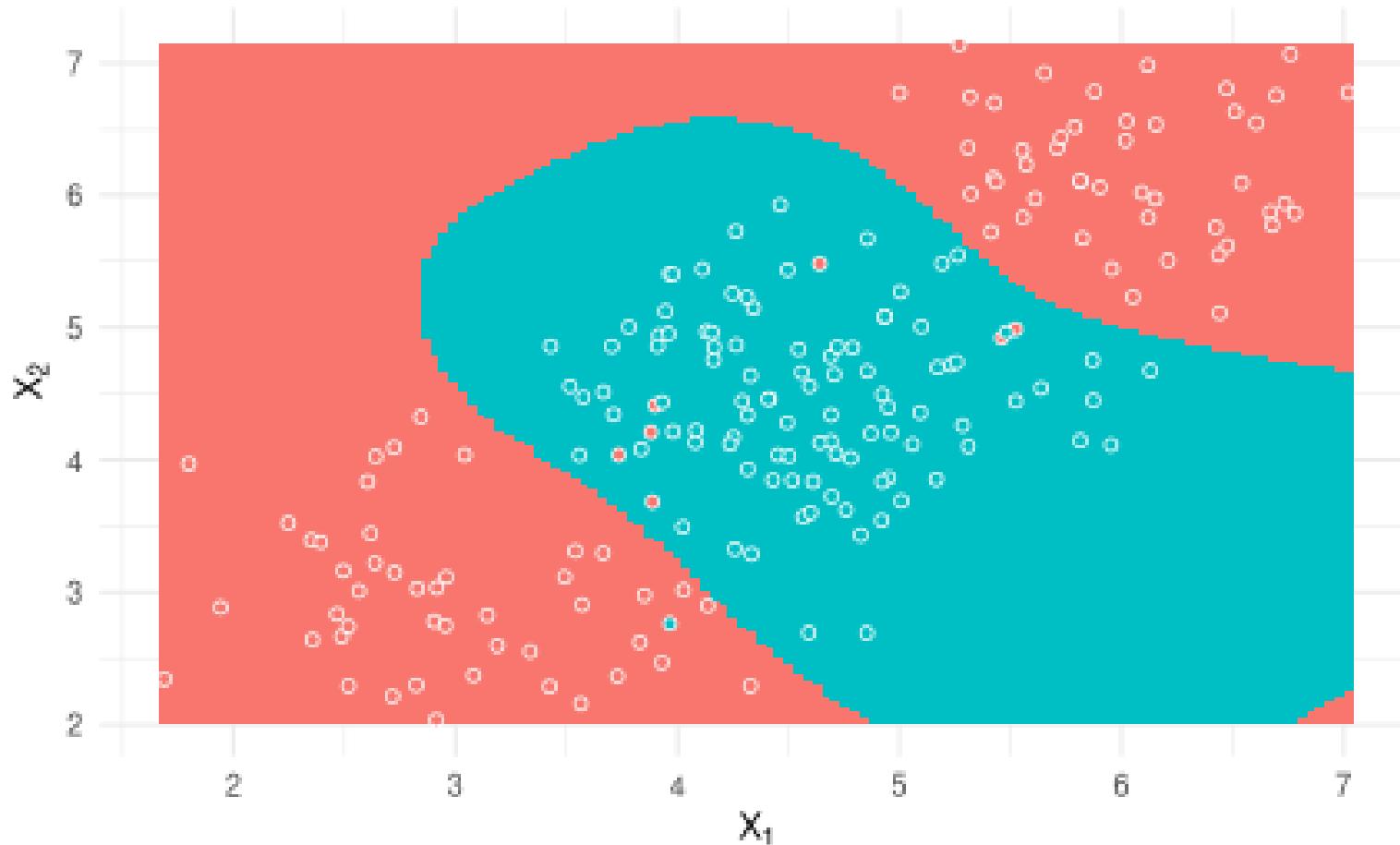
- Examples
 - Polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^d$$

- Gaussian radial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \sum_{k=1}^p (x_{ik} - x_{jk})^2\right)$$

Support vector machines with the radial kernel



SVM with more than two classes

- SVM covered previously are design for binary classification. To expand this to K classes there are two options

1. One vs all:

- Fit K different binary classifiers, $f_k(\mathbf{x})$ for $k = 1, 2, \dots, K$ where each boundary attempts to separate class k vs the rest.
- Then \mathbf{x}_i is classified to k^* where $f_{k^*}(\mathbf{x}_i) > f_j(\mathbf{x}_i)$ for all $j \neq k^*$. (i.e. the largest distance from the boundary).

2. One vs one:

- Fit all $\binom{K}{2}$ pairwise classifiers
 - Fit \mathbf{x}_i to the class that wins the most pairwise comparisons.
-
- Which to use?
- If K is small, do one vs one. Otherwise recommended One vs all.

一对all 把某个类别的样本归为一类, 其他剩余的样本归为另一类, 这样 k 个类别的样本就构造出了 k 个SVM, 最终的结果便是这四个值中最大的一个作为分类结果

优点: 训练 k 个分类器, 个数较少, 其分类速度相对较快。

缺点: ①每个分类器的训练都是将全部的样本作为训练样本, 这样在求解二次规划问题时, 训练速度会随着训练样本的数量的增加而急剧减慢;

一对一: 在任意两类样本之间设计一个SVM, 因此 k 个类别的样本就要设计 $k(k-1)/2$ 个SVM。

如果 K 是小的, 做一个对一个。否则建议一个对所有。

References

- Hastie, T, R. Tibshirani, and J. Friedman (2017). *The elements of statistical learning: data mining, inference, and prediction*. Second Edition, 12th printing. Springer Science & Business Media.
- James, G, D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

STAT5003

Week 6 : Cross validation and bootstrapping

Dr. Justin Wishart



Readings and R functions covered



Readings

- Cross validation and bootstrap covered in Chapter 5 in James, Witten, Hastie, and Tibshirani (2013)

R functions

- `caret::createDataPartition`
- `caret::train`
- `pROC::roc`
- `pROC::auc`

Training error vs test error



THE UNIVERSITY OF
SYDNEY

Training error vs test error

Training error is the performance metric applied to the observations used to train the model.

Test error is the average error when applying a model to predict the response on new (test) observations that were **not** used in the training of the model.

- Training error is usually very different in magnitude to the test error.
 - Training error can **underestimate** the test error.

训练误差是应用于用于训练模型的观测的性能指标。

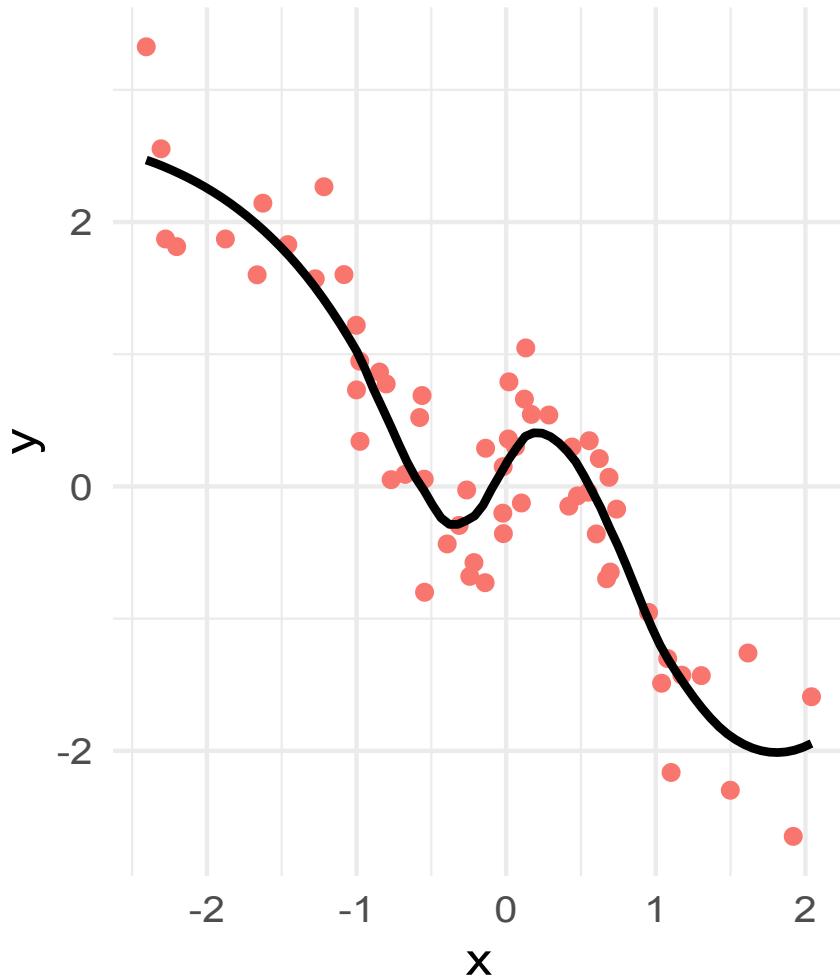
测试误差是应用模型预测新（测试）观测值的反应时的平均误差。

训练误差的大小通常与测试误差大不相同。

训练误差可以低估测试误差

Pick the better model

Low training error



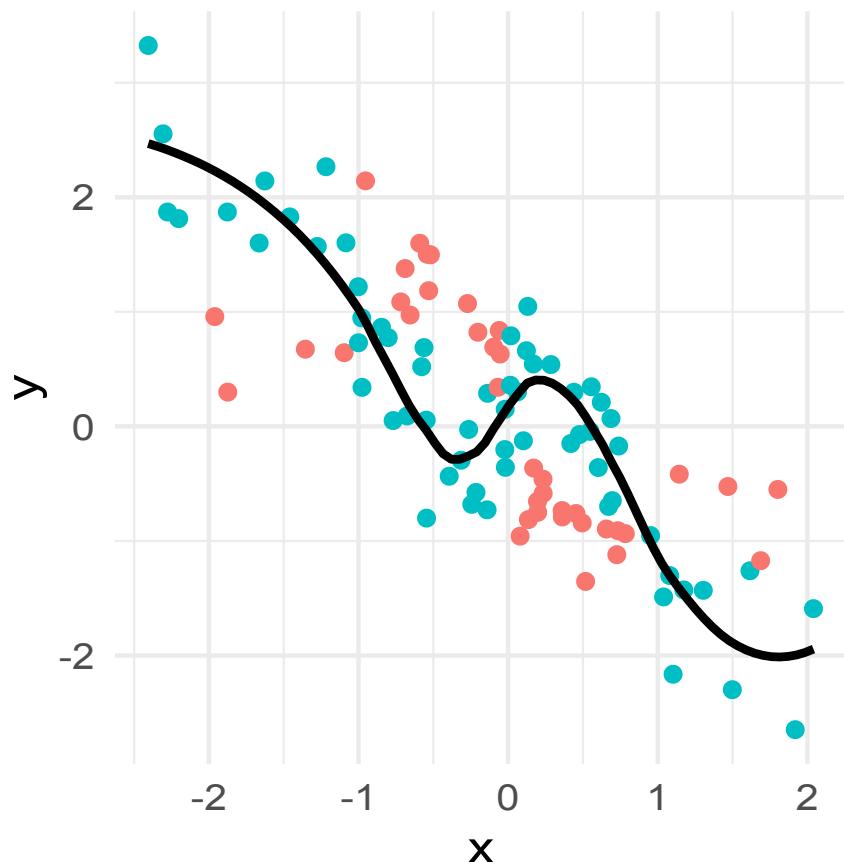
High training error



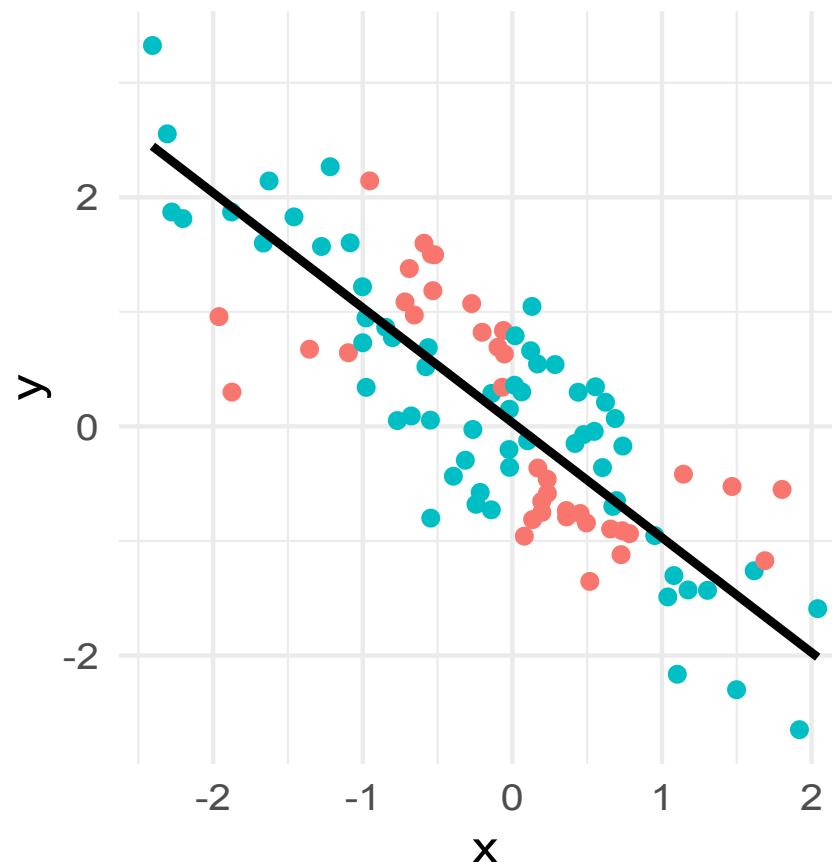
Pick the better model

Set • Test • Training

Low training, High test error

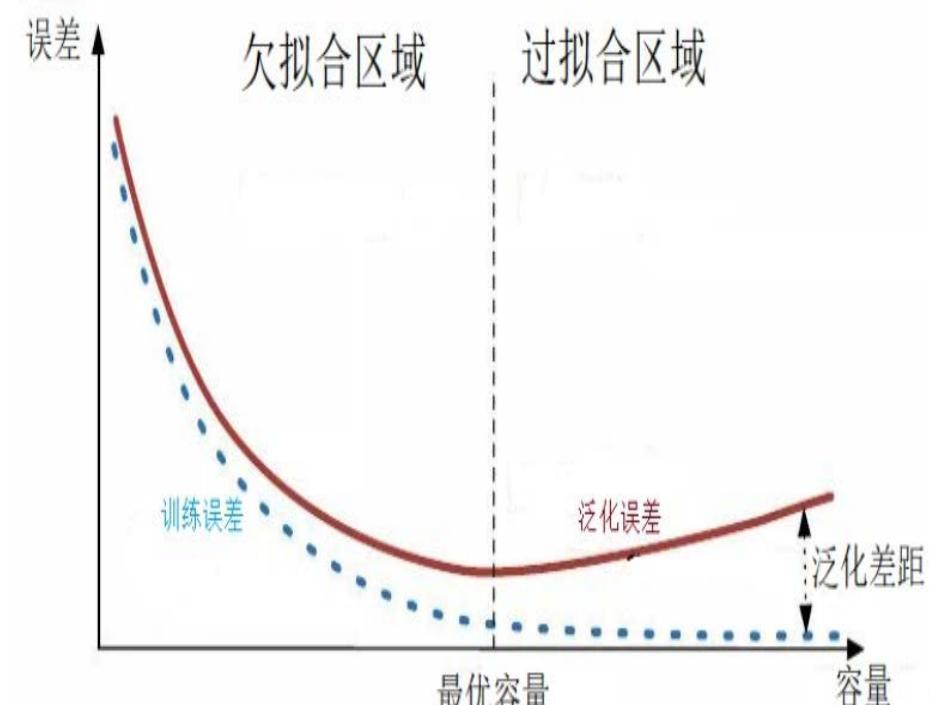
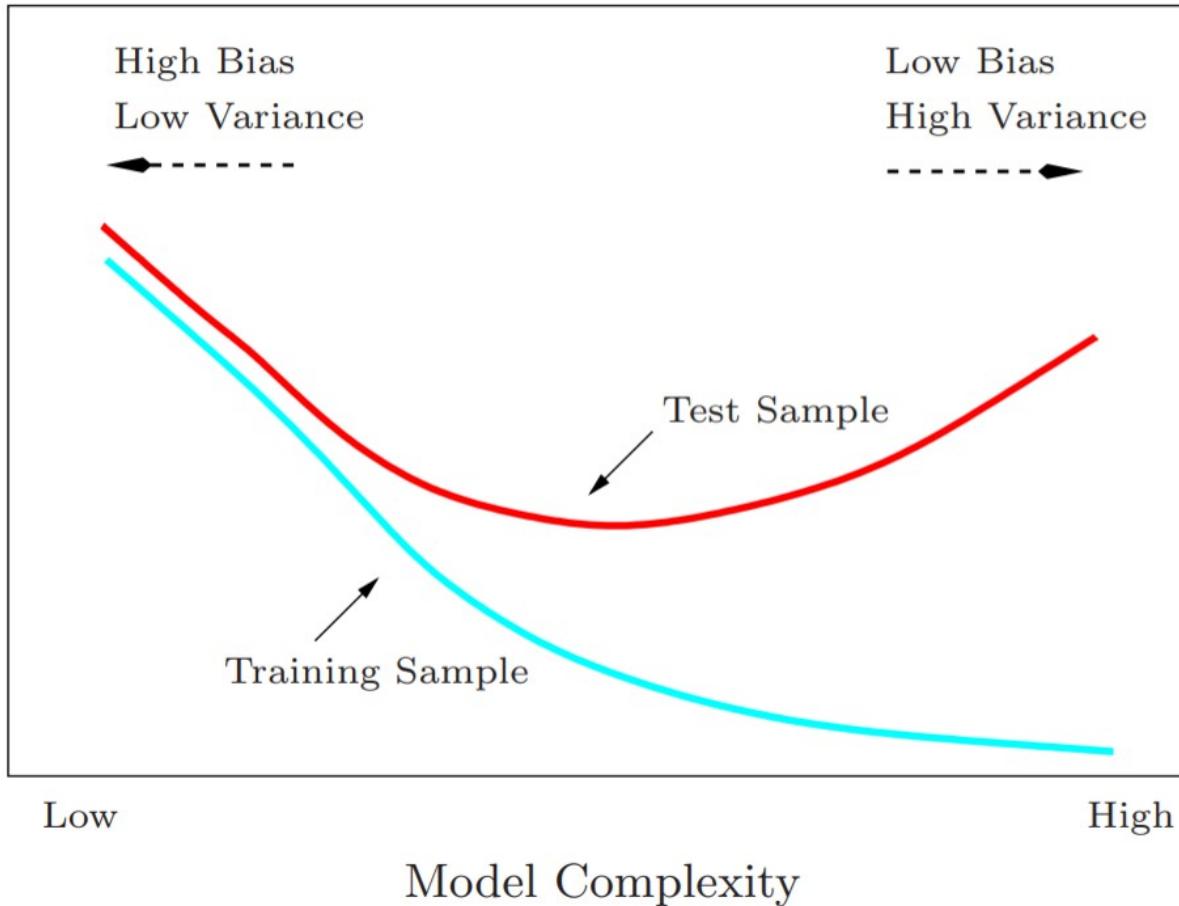


High training, low test error



Training set vs Test set error

Prediction Error



Estimate the test error

- Gold standard:
 - Use a large designated test set. Often not available
- Adjust the training error to estimate the test error
 - Common to add a penalty term to the model
 - BIC
 - Adjusted R^2
- Cross validation
 - Remove or hold out a subset of observations (test set) and use the rest to train the model.
 - Assess model performance on the test set.

调整训练误差以估计测试误差

C_p 选择最小

BIC 选择最小

Adjusted R² 选择最大

交叉验证

移除或保留一个观察的子集

(测试集)，用其余的观察来训练模型。

评估模型在测试集上的表现

Test Set approach

- Here we randomly divide the available set of samples into two
 - a training set
 - test set
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the test set.
- The resulting test-set error provides an estimate of the test error. Typically assessed using
 - MSE in the case of a quantitative response
 - Misclassification rate in the case of a qualitative (discrete) response.

由此产生的测试集误差提供了一个测试误差的估计。典型的评估方法是使用
在定量反应的情况下MSE
在定性（离散）反应的情况下，错误分类率。

Example of the training and test split



- Random split of the data into two halves
 - The left is the training indices
 - The right is the test indices

Drawbacks of test set approach

- The estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the test set.
- In the test set approach, only a subset of the observations are used to fit the model.
 - This suggests that the test set error may tend to overestimate the test error for the model fit on the entire data set.

测试误差的估计可能是高度可变的，这恰恰取决于哪些观测值被包括在训练集中，哪些观测值被包括在测试集中。

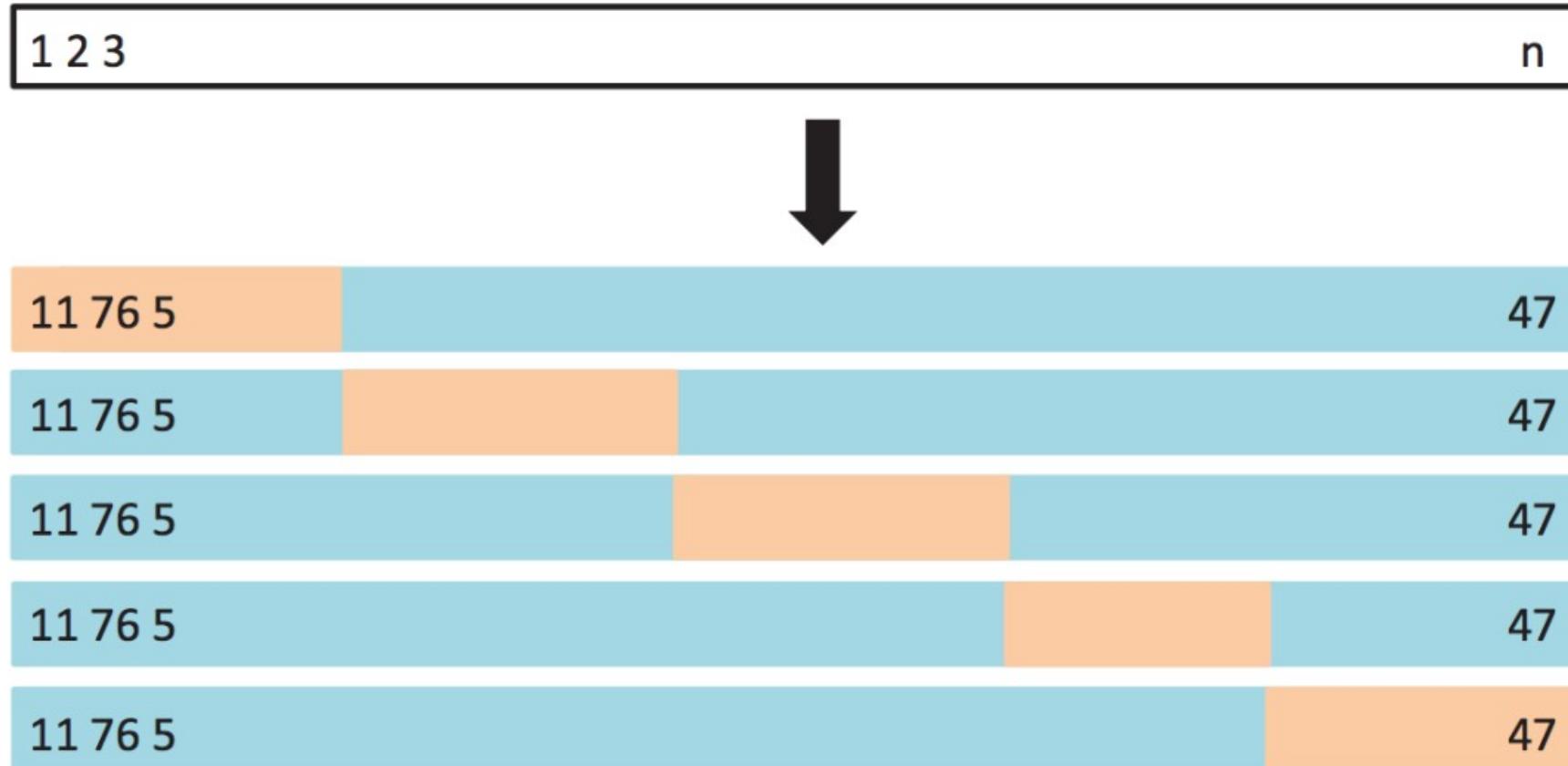
在测试集方法中，只有观测值的一个子集被用来拟合模型。测试集误差可能倾向于高估整个数据集上模型拟合的测试误差。

K -fold cross validation

- Widely used approach for estimating test error.
 - Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts.
 - We leave out part k , fit the model to the other $K - 1$ parts (combined), and then obtain predictions for the left-out k^{th} part.
- This is done in term for each part $k = 1, 2, \dots, K$ and then the results are combined.

把数据集分成多部份其中一个数据作为测试集，其他作为训练集，重复N次,最后平均MSE

Example: 5-fold



Cross-validation formula

- Let the K parts be C_1, C_2, \dots, C_K , where C_k denote the indices of the observations in part k .
 - There are n_k observations in part k :
 - if n is a multiple of K , then $n_k = \frac{n}{K}$
- Compute

$$CV_k = \sum_{k=1}^K \frac{n_k}{n} MSE_k$$

- where $MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$
- \hat{y}_i is the fit for observation i obtained from the data with part k removed.

Cross-validation for classification problems

- For classification problems, we can compute the accuracy for each fold by calculating:

$$CV_K = \sum_{k=1}^K \frac{n_k}{n} A_k$$

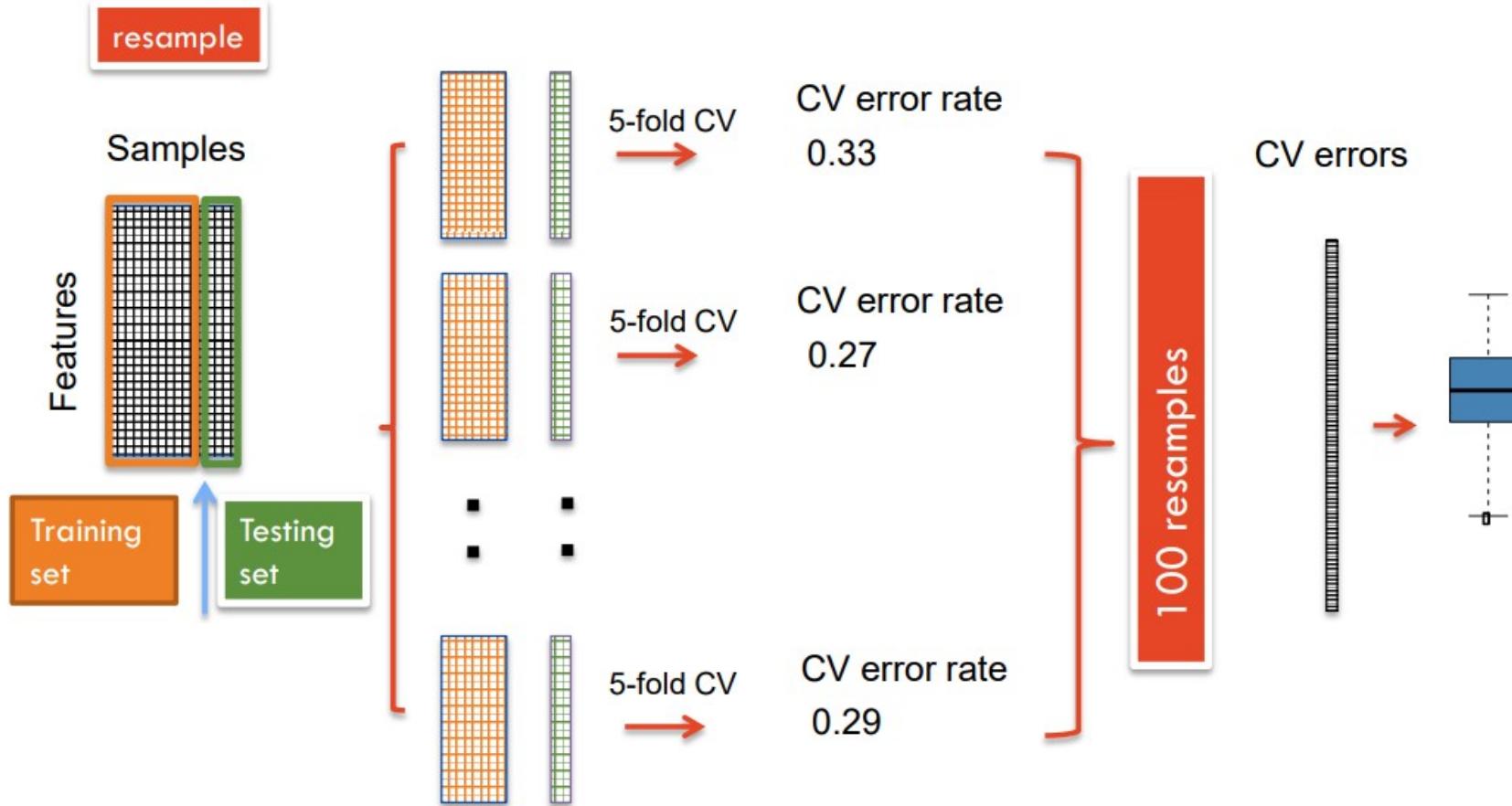
where the terms are

- n : The total number of observations in the dataset
- n_k : The number of observations in the belonging to class k
- A_k : The accuracy of the classifier in fold k

- e.g. $A_k = \frac{1}{n_k} \sum_{i \in C_k} 1_{\{\hat{y}_i = y_i\}}$

Repeated Cross validation

重复多次实施

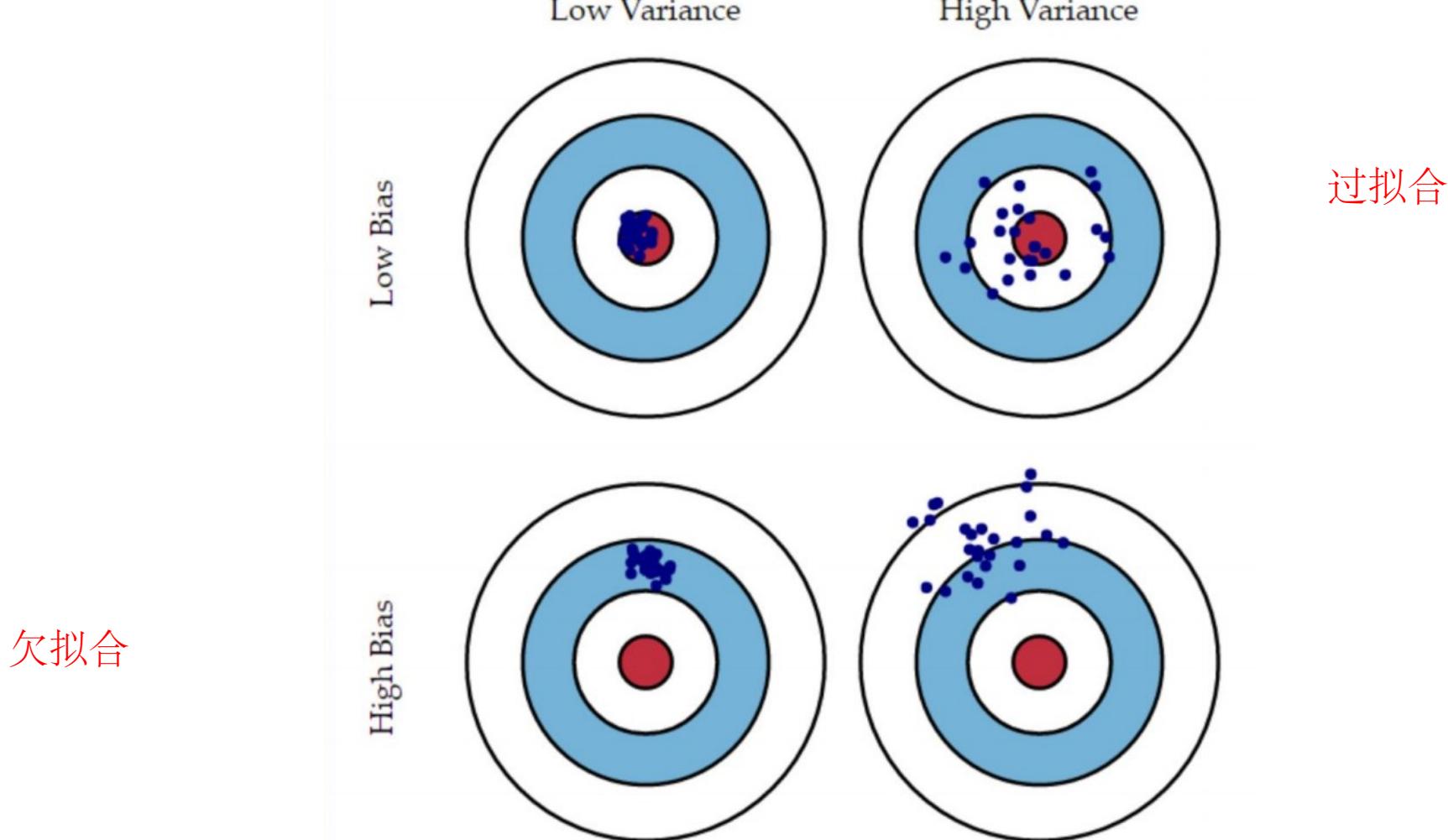


Repeated cross validation properties

In general, repeated CV provides a less biased CV error estimate

- Repeated CV also gives you the variance of the CV error
- However, it comes with a computational cost
- Implemented in the `caret` package in R

Dart board interpretation of bias & variance



Example of CV procedure

Consider a problem where you have a high dimensional data set, all entirely numeric, and need dimension reduction to proceed.

- You decide to reduce the dimensions of the data and use the following CV procedure:
 1. Compute correlation matrix, select the top 50 variables that have the highest correlation with the response.
 2. Use these 50 variables as features and perform K -fold cross validation

你有一个高维数据集，全部都是数字，并且需要降维来进行。

你决定减少数据的维度，并使用以下CV程序。

1. 1. 计算相关矩阵，选择与反应有最高相关性的前50个变量。
响应。

2. 2. 使用这50个变量作为特征，并进行K-fold交叉验证。

Issue with the previous slide

- Variable selection performed once using both the training and the test datasets
- Information can leak from the test to the training set
- Hence, the CV error estimate is likely to be biased.
- Ideally you shouldn't use the test data in any way in the training step.
 - If absolutely necessary some pre-processing on the features can be done so long as it doesn't involve the response variables.

使用训练和测试数据集进行一次变量选择

信息可能从测试集泄露到训练集上

因此，CV误差估计可能会有偏差。

理想情况下，你不应该在训练步骤中以任何方式使用测试数据。

如果绝对有必要，可以对特征进行一些预处理，只要不涉及响应变量。

Corrected CV procedure

- Split the dataset into K folds
- For each $k = 1, 2, \dots, K$
 - Determine the variables that correlate the best with the response using all the data except the data in fold k
 - Train your model using the selected variables above.
 - Run your classification algorithm and record accuracy against the test set.

将数据集分成k个折页

对于每个 $k = 1, 2, \dots, K$

使用除褶皱k中的数据外的所有数据，确定与响应关联度最高的变量

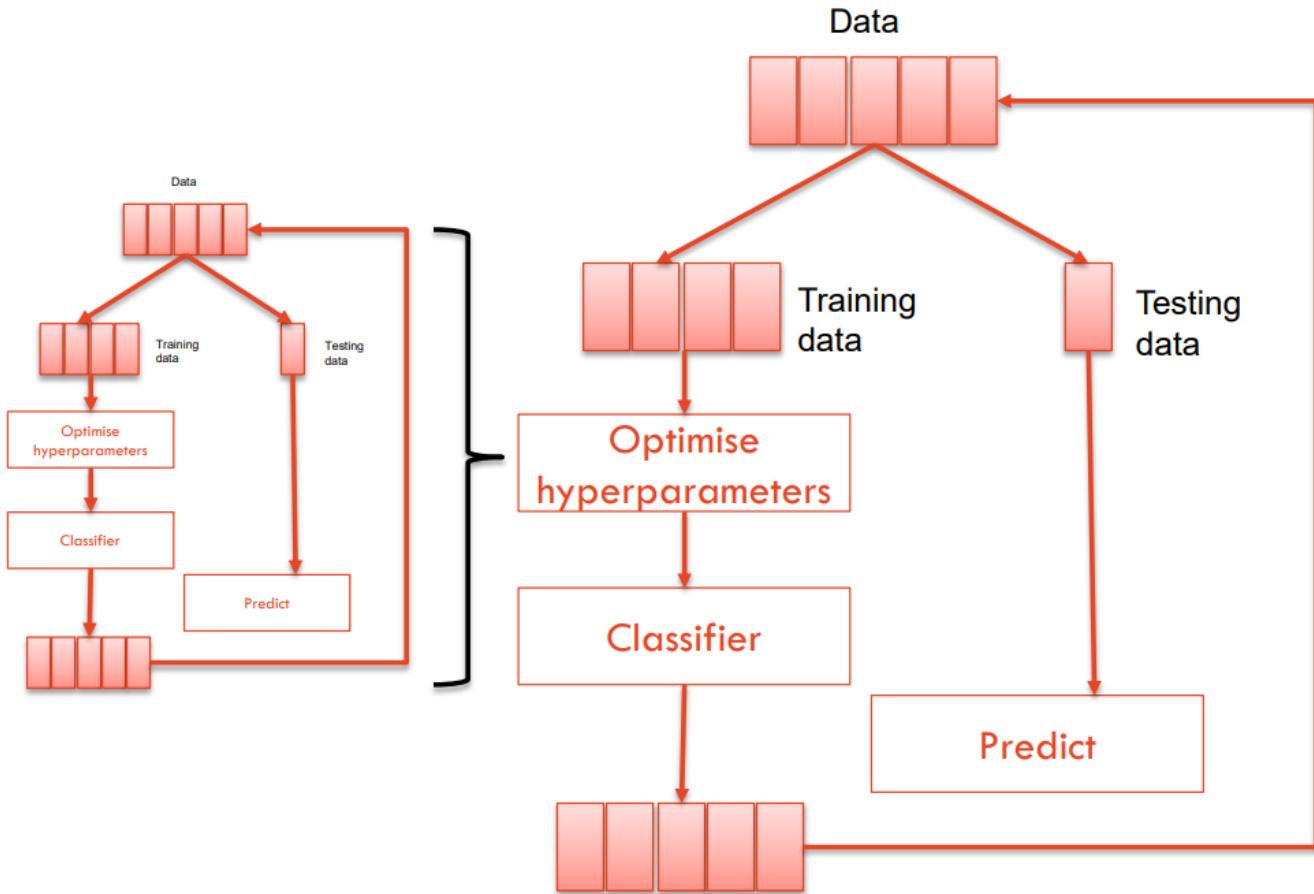
使用上述选定的变量训练你的模型。

运行你的分类算法，并针对测试集记录准确性。

Other information leakage to check

- Other things you should not do once but do it within with CV loop
 - Feature selection
 - Hyperparameter optimization
 - Missing data imputation
- Another method is nested cross validation

Nested cross validation



嵌套交叉验证是通过对基础模型泛化误差Generalization bias的估计来进行超参数Hyperparameters的搜索，以得到模型最佳参数

Final model building

- The reason for doing cross-validation is to evaluate the different models by estimating their performance on unseen data
- Example. If you need to choose between kNN, LDA and logistic regression and SVM, then you can run each of these classification algorithms with cross-validation, and pick the one with the highest CV accuracy
- But then, you can go back to use all the data to build a final model

进行交叉验证的原因是通过估计不同模型的。在未见过的数据上的表现来评估不同的模型例子。如果你需要在kNN、LDA和逻辑回归以及SVM之间做出选择，那么你可以用交叉验证法运行这些分类算法，然后挑选出具有最高准确率

Classification evaluation metrics



THE UNIVERSITY OF
SYDNEY

Classification accuracy

- Overall classification accuracy:
- Disadvantages:
 - Makes no distinction about the type of errors being made.
 - In spam filtering, the cost of erroneous deleting an important email is likely to be higher than incorrectly allowing a spam email past a filter.
 - Does not consider the natural frequencies of each class

缺点。
对所犯错误的类型不作区分。不考虑每个类别的自然频率

Confusion Matrix

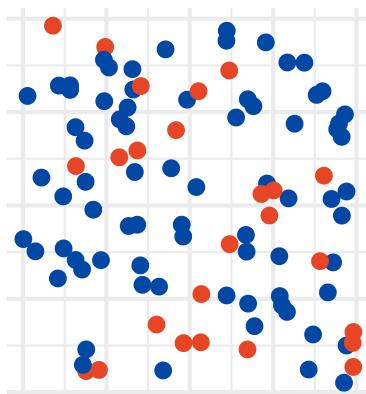
		Actual	
		True	False
Predicted	True	True Positive	False Positive
	False	False Negative	True Negative

- True positive: Are positive class and predicted to be positive class
- False positive: Are negative class but predicted to be positive class
- False negative: Are positive class but predicted to be negative class
- True negative: Are negative class and predicted to be negative class

Sensitivity and Specificity

100% Sensitivity

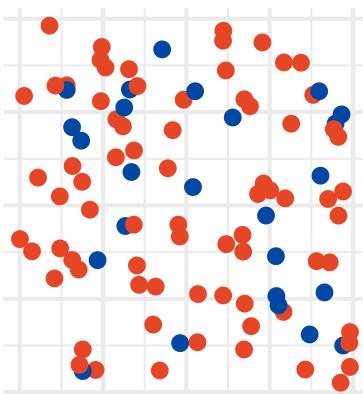
Class • Negative • Positive



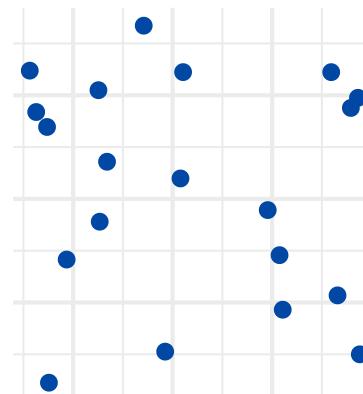
Test Positive

100% Specificity

Class • Negative • Positive



Test Negative



Test Positive

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} = \frac{TP}{P}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} = \frac{TN}{N}$$

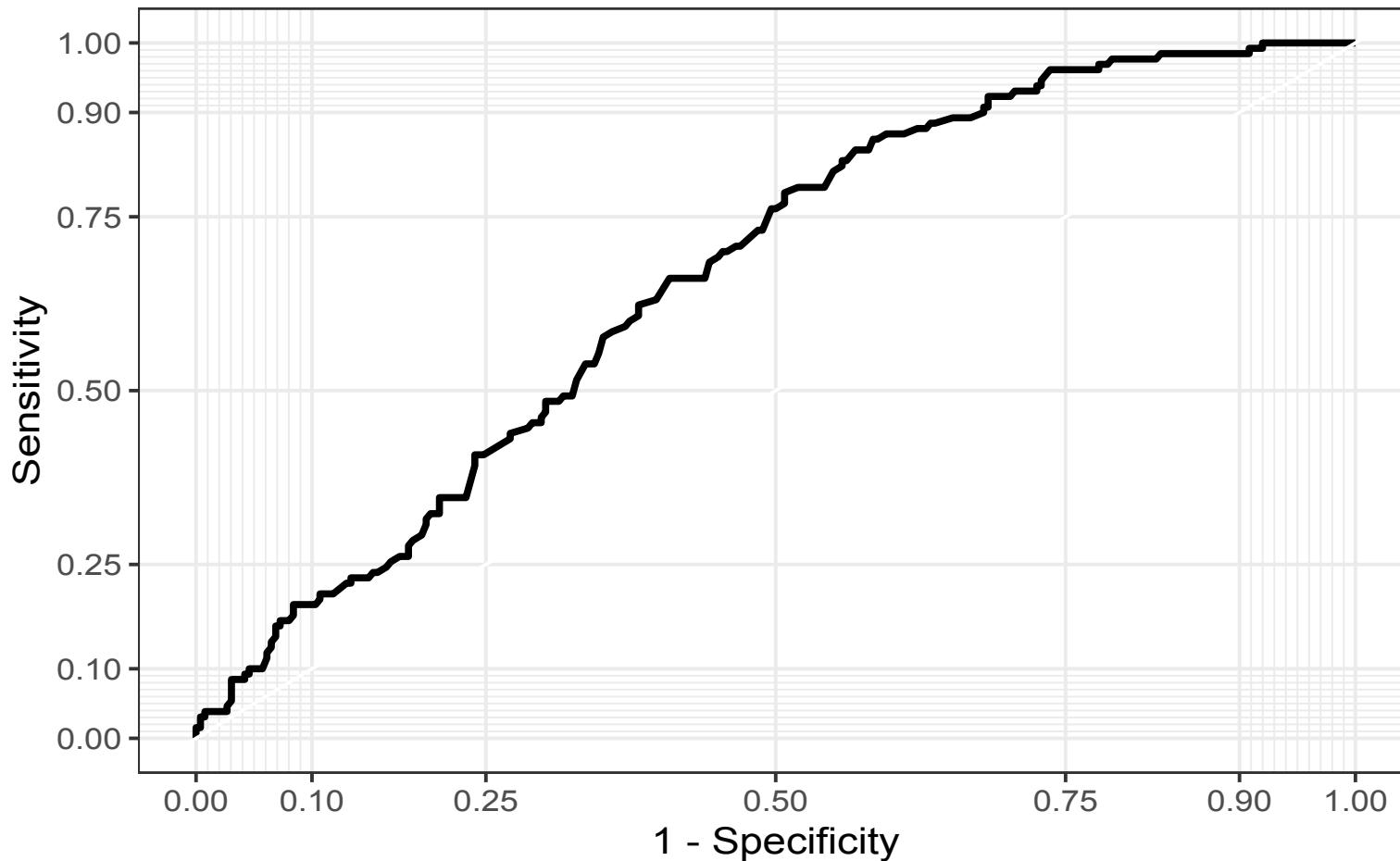
$$\text{Precision} = \frac{TP}{(TP+FP)}$$

$$\text{Recall} = \frac{TP}{(TP+FN)} = \frac{TP}{P}$$

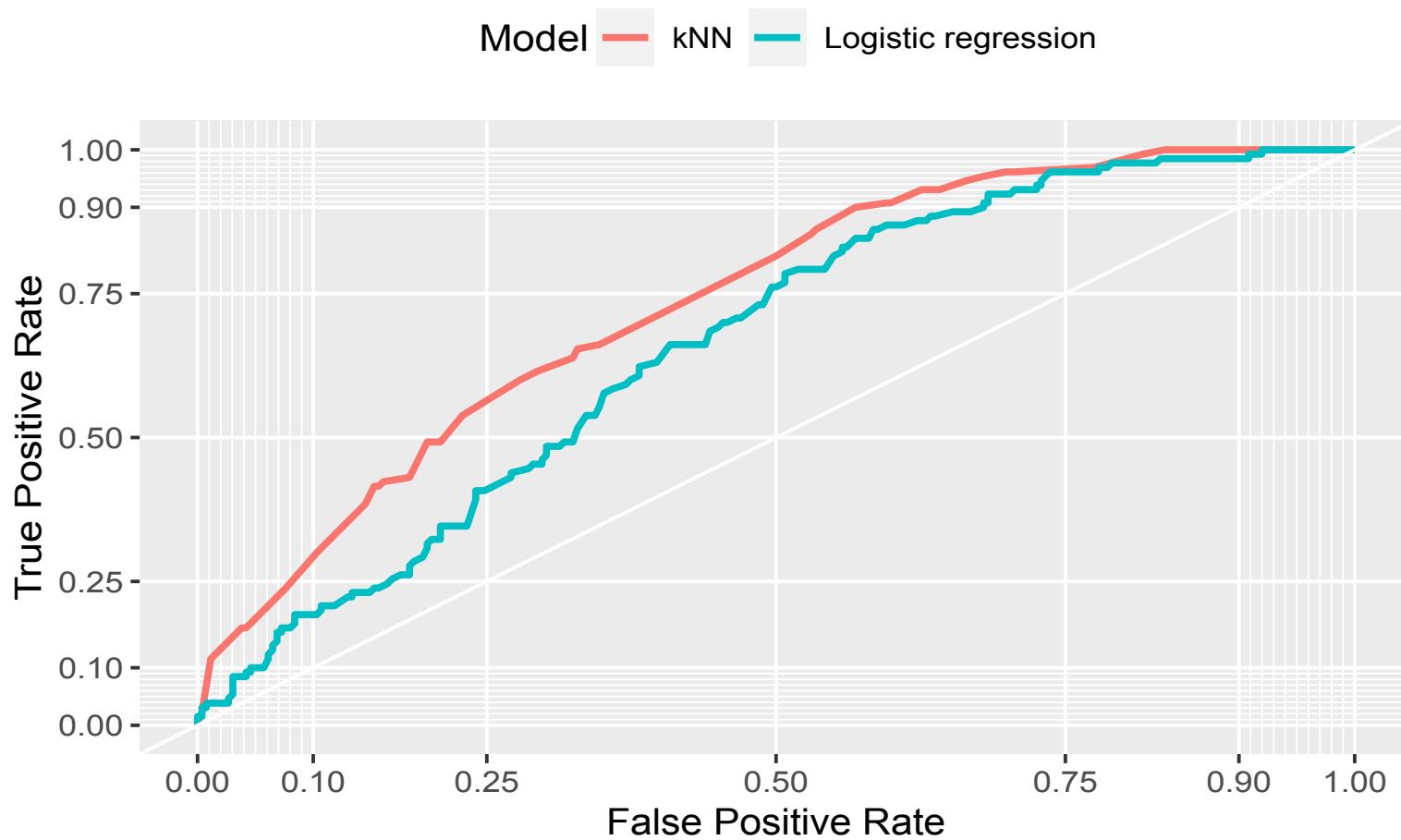
$$F_1 = \frac{2 \text{ Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \text{ (Harmonic mean)}$$

$$\text{GM} = \sqrt{\text{Precision} \times \text{Recall}} \text{ (Geometric mean)}$$

Receiver Operating Characteristics (ROC) curve



Comparing ROC curves



Precision是指预测样本正确，实际正确的概率，它显示了模型对阳性案例的准确度。而sensitivity和recall是指实际正确的样本被预测为正确的概率，它反映了模型的全面性，衡量分类器识别阳性案例的能力

Accuracy

binary classification 且正反例不平衡的情况下，模型过于简单导致无法输出错误，而是输出正确的例如真实的癌症患者比例为5%，也就是说100人中，有5个人患有癌症，95个人健康。如果我们建立一个模型，帮助医生建模去做癌症诊断。这个模型很简单，它只会输出‘健康’，而不会输出‘癌症’。将所有病人分类为‘健康’。 Accuracy = 95%

Receiver Operating Characteristics (ROC) curve

我们可以根据ROC曲线是否在 $y=x$ （随机猜测）上方或下方来判别它的性能。若在上面，则性能较好，若在下方，则性能较差。分类器的ROC曲线越靠近(0,1)，则证明该分类器性能越好F1分数是精确性和敏感性的结合和平衡，用来评价分类器的性能。

AUC

ROC面下面积

<https://www.zhihu.com/question/30643044>

Bootstrap



THE UNIVERSITY OF
SYDNEY

Bootstrap (非参数)

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

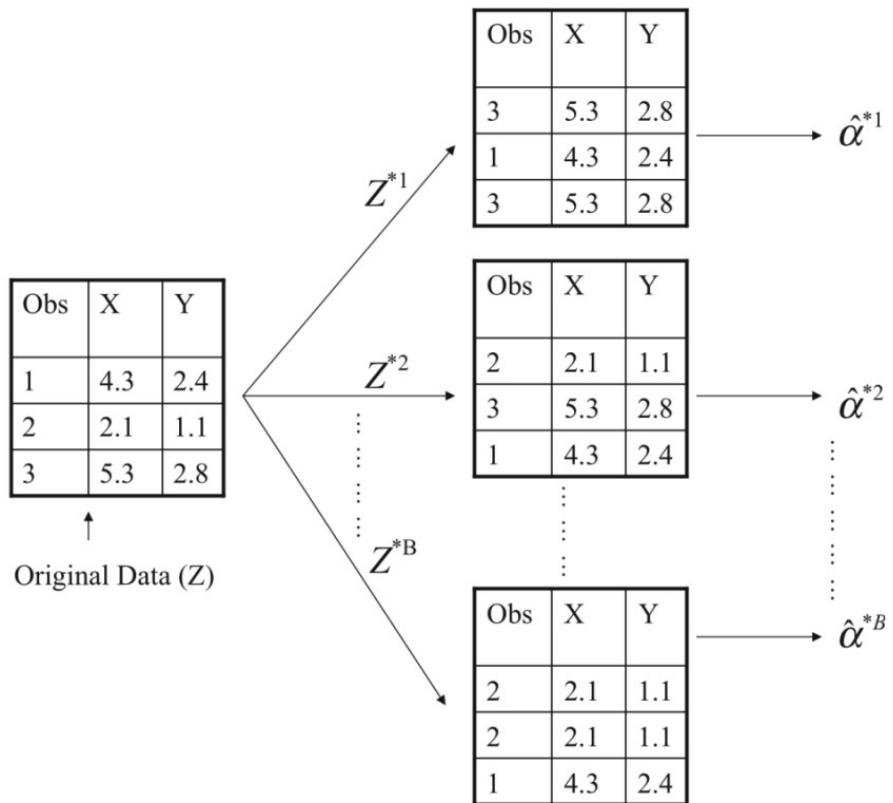
一种灵活而强大的统计工具，可用于量化与给定估计器或统计学习方法相关的不确定性。

例如，它可以提供一个系数的标准误差估计值，或该系数的置信区间

<https://murphypei.github.io/blog/2019/03/bootstrap-bagging-boost>

Bootstrap resampling algorithm

- Essentially sampling with replacement



1. 在原有的样本中通过重抽样抽取一定数量（比如100）的新样本，重抽样（Re-sample）的意思就是有放回的抽取，即一个数据有可以被重复抽取超过一次。

2. 基于产生的新样本，计算我们需要估计的统计量。

在这例子中，我们需要估计的统计量是 a ，那么我们就需要基于新样本的计算样本方差、协方差的值作为以及，然后通过上面公式算出一个 a_1

3. 重复上述步骤n次（一般是

$n > 1000$ 次）。

在这个例子中，通过n次（假设 $n=1000$ ），我们就可以得到1000个 a_i 。

4. 最后，我们可以计算被估计量的均值和方差

Simple example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y where X and Y are random quantities.
- The goal is to create a portfolio by investing fraction α of our wealth in X and $(1 - \alpha)$ in Y .
- Want to choose to minimise the total risk of the investment. Mathematically this involves minimising $Var(\alpha X + (1 - \alpha)Y)$
- The solution to this problem (calculus) is,

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \quad (1)$$

- where $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$ and $\sigma_{XY} = Cov(X, Y)$

Example

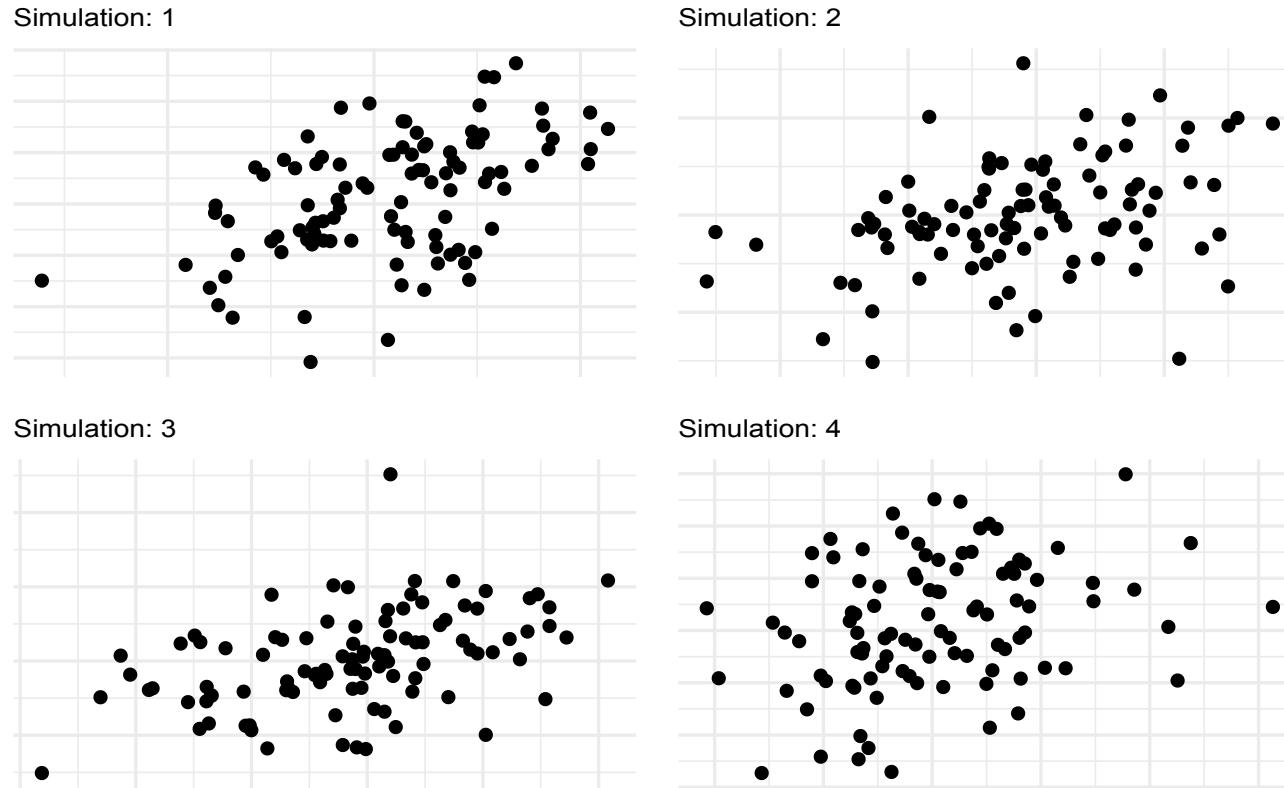
- The values of σ_X^2 , σ_Y^2 and σ_{XY} are unknown but estimates can be computed from the data.
- The estimate of α that minimises the variance of the investment can then be computed with

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}} \quad (2)$$

- Suppose that X and Y can be sampled from the population repeatedly
- To estimate the standard deviation of $\hat{\alpha}$, paired observations (X, Y) can be repeated simulated, say 100 pairs to get a single estimate of α . Repeat this process to get 1,000 estimates for α .
- Denote these estimates $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{1000}$

Bootstrap simulations

- Consider example with $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.5$ and $\sigma_{XY} = 0.5 \Rightarrow \alpha = 2/3$.



- Each panel shows 100 simulated returns. From left to right, top to bottom, the estimates for α are 0.659, 0.683, 0.726, 0.68.

Parameter estimates

- Consider the mean of all the parameter estimates

$$\bar{\hat{\alpha}} = \frac{1}{1,000} \sum_{k=1}^{1000} \hat{\alpha}_k = 0.6662595$$

- This is close to the true value of 0.6666667
- Estimate of the standard error using the standard deviation of all the estimates.

$$\sqrt{\frac{1}{1000 - 1} \sum_{k=1}^{1000} (\hat{\alpha}_k - \bar{\hat{\alpha}})^2} = 0.0760217$$

- This gives an intuitive description of the reliability of the estimator.
 - For a random sample the estimate would vary around the true value by 0.0760217

Application in reality

- Cannot apply this directly in reality
 - cannot generate new observations from the population model.
- Bootstrap attempts to mimic this process
- Instead of sampling new independent observations from the population
 - Re-sample observations from the data *with replacement*
- Some observations appear more than once and some not at all

不能在现实中直接应用

不能从人口模型中产生新的观察结果。

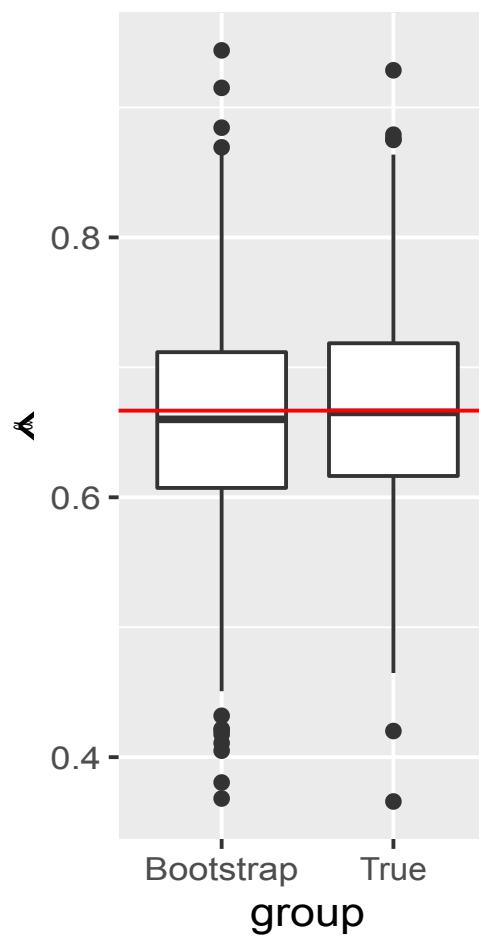
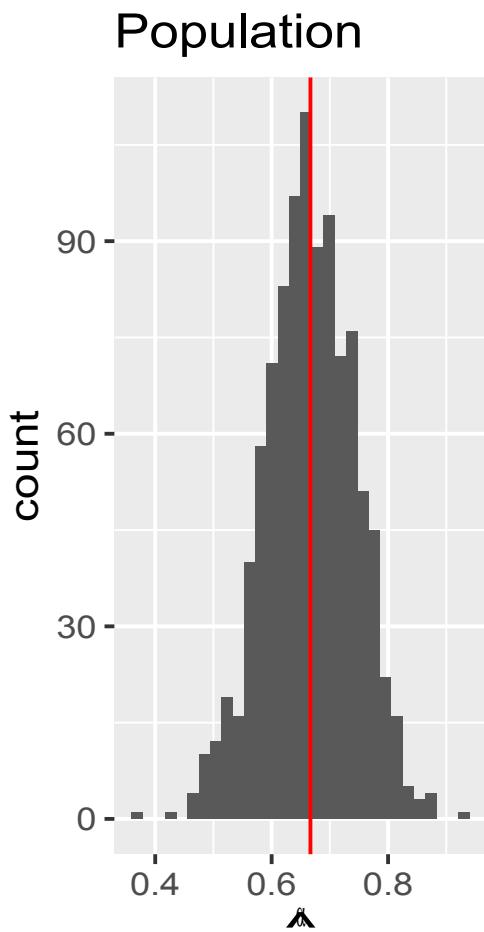
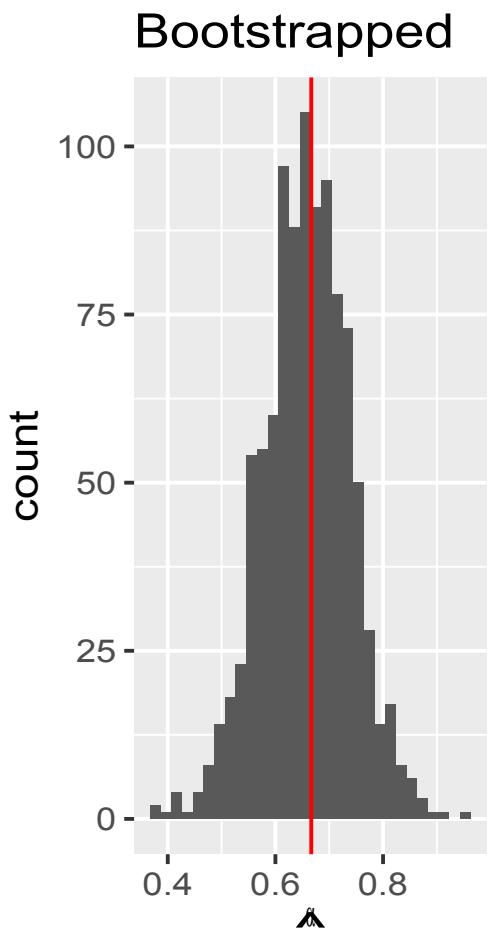
Bootstrap试图模仿这个过程

而不是从人口中抽取新的独立观测值

用替换法从数据中重新取样观测值

有些观测值出现了不止一次，有些根本没有出现

Results bootstrap vs population



References

James, G., D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

STAT5003

Week 7 : Missing data and class imbalance

Dr. Justin Wishart



Missing data



THE UNIVERSITY OF
SYDNEY

Mechanism for missing data

与人无关

- Missing Completely At Random (MCAR) 例如：假设我们做了一个调查，有些人不愿意在问卷中提供他们的年龄。但这与任何其他变量（包括他们的政党偏好）无关
 - E.g. Let's say we run a survey and some people don't want to give their age in the questionnaire, but this does not relate to any other variable (including their party preference)

- Missing At Random (MAR)

与预测数据无关与其他变量有关，可以从其他信息预测

- The missingness of the data in a variable is not related to the variable but related to some other variables.
一个变量中的数据缺失与该变量无关，而是与其他一些变量有关。例如：在一项调查中，如果来自社会经济地位较低的人可能不太愿意提供工资信息
- E.g. In a survey, if people from a lower socioeconomic status may be less willing to provide salary information (but we know their SES status).
- In other words. The salary value is missing not entirely due to the salary but conditional on the socioeconomic status.

- Missing Not At Random (MNAR)

确实内容与
变量相关

- When data is not MCAR or MAR.
- The missingness of the data is due to the value of the variable itself even after accounting for other variables

当数据不是MCAR或MAR时。数据的缺失是由于变量本身的价值，即使在考虑了其他变量

Identifying different types of missingness

Unfortunately, there is no statistical method to determine the mechanism of missingness

- You can guess the mechanism of missingness by knowing something about the data, and something about the data collection method
- To see if the data is MAR, can try to fit a classification model to predict missingness

Dealing with missing values

- For categorical data, "missing" can be a category.
 - For example, in a survey poll, if someone does not want to disclose who they want to vote for, can be in the category "undecided"
- Delete data with missing value. Two options.
 - Omit the variable with missing data.
 - Omit the observation with missing data.
 - Drawbacks are that you might be throwing away valuable information, or inadvertently introduce bias into the data
- Impute i.e. fill in the missingness.
 - Can replace missing values with the mean of the ones observed for that feature

对于分类数据，"缺失"可以是一个类别。
例如，在一个调查投票中，如果有人不愿意透露他们想投给谁。

删除有缺失值的数据。有两个选项。省略

有缺失数据的变量。省略有缺失数据的观察值。

缺点是你可能会丢掉有价值的信息，或者
不经意间把偏差到数据中

添加数据替代缺失值可以用观察到的该特征的平均值来代替缺失值

Single imputation

- Single imputation replace the missing value with a single value.
- Examples:
 - Replace the missing values of a feature with the mean/median value of that feature
 - Use a predictive method for filling in the missing values e.g. regression trees, kNN
 - Replace the missing value with the last observed value for that feature
 - With single imputation, once the missing data is added back, it is treated as valid observed data, hence the uncertainty in the missing value data is lost.

用单一数值代替缺失的数值。

用一个特征的平均值/中位数替换该特征的缺失值。使用预测方法来填补缺失值，例如回归树、kNN。用该特征的最后观察值替换缺失值在单一归因法中，一旦缺失的数据被添加回来，它就被视为有效的观察数据。因此，缺失值数据的不确定性就消失了。

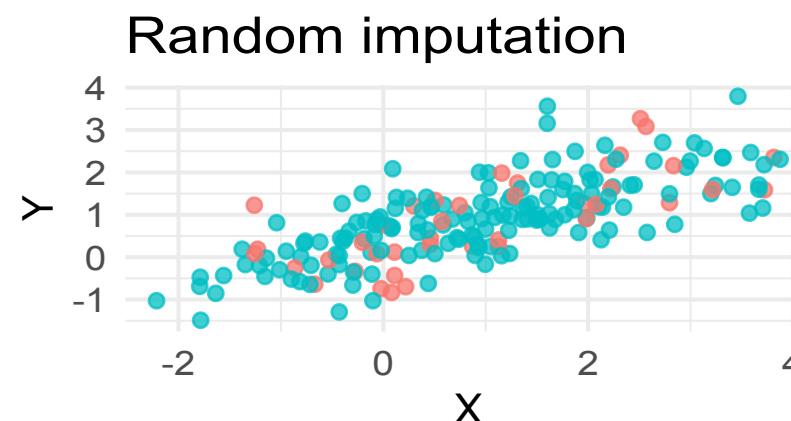
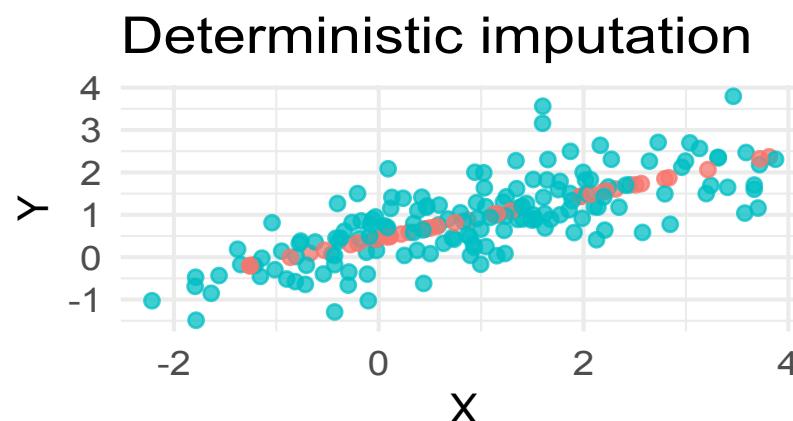
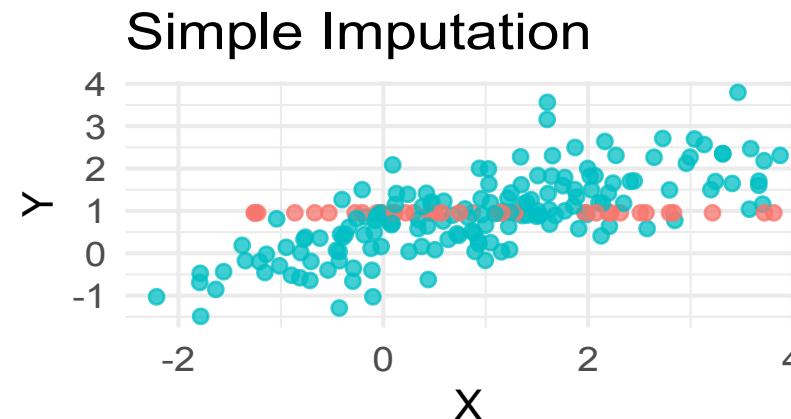
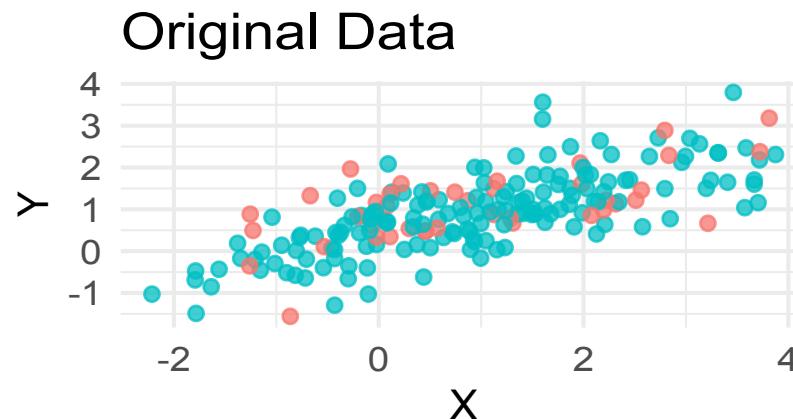
Multiple imputation

- Multiple imputation accounts for uncertainty in the imputation process.
- Generally follows three steps:
 - Impute the data k times (this can be done using a single imputation method)
 - Perform analysis (e.g. regression) on each of the k imputed data sets
 - Pool the k results together
- Multiple Imputation by Chained Equation (MICE) is a popular method
 - See van Buuren and Groothuis-Oudshoorn (2011)

考虑了归因过程中的不确定性。一般来说遵循三个步骤。
对数据进行 K 次归因（可以使用单一的归因方法进行）。对每一个
经过归因的数据集进行分析（如回归）。将 K 个结果汇集在一起。
链式方程多重归因法（MICE）是一种流行的方法。

Basic impute, deterministic imputation, random imputatation

Missing • Missing • Not missing



Other practical suggestions

- It is highly recommended that you visualize your data to look for patterns of missingness
- Be wary of variables with high proportion of missing data. However, this might not be a problem if imputation is applicable and performs well.
- Some algorithms can cope with missingness (e.g. decision trees) and so you may not need to do imputation
- If you believe the pattern of missingness is informative, you can include it as a dummy variable

R packages for dealing with missing values

- Impute (Bioconductor package)
 - KNN imputation, written for microarray data
- MICE
 - Multiple imputation
- missForest
 - Uses Random Forest (in upcoming tree based module) to predict the missing values
 - Can be used for continuous and categorical data
- Amelia II
 - Multiple imputation

Class imbalance

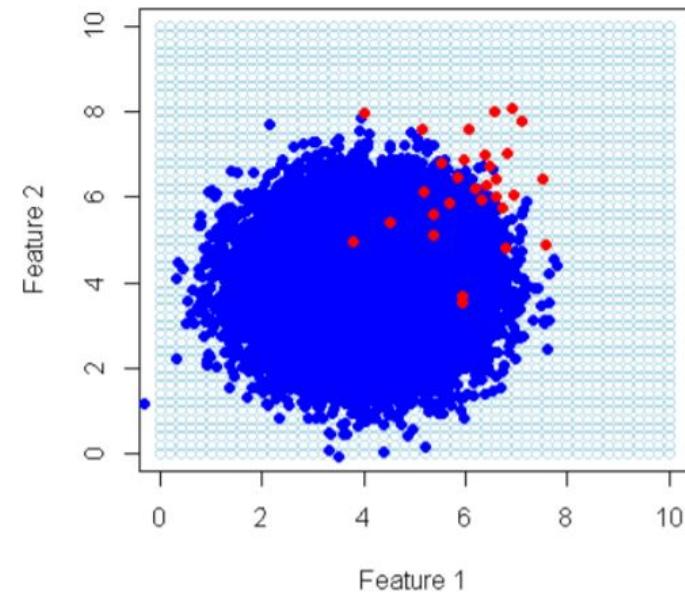
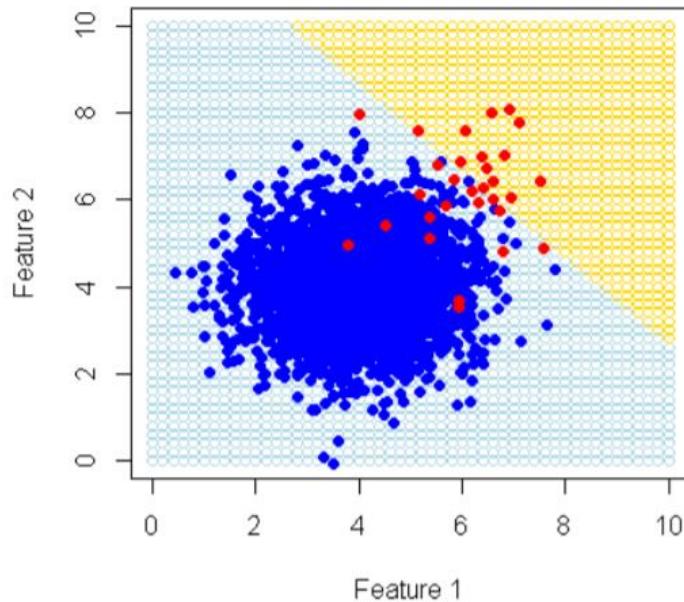
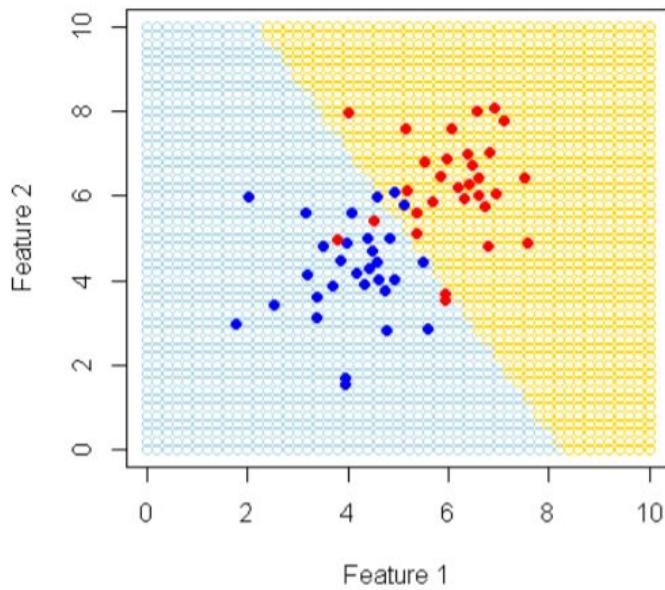


THE UNIVERSITY OF
SYDNEY

Class imbalance

- Let's say we have a classification problem to detect credit card fraud, but only 1% of transactions are fraud.
- If you use accuracy as the metric to optimize, then just by classifying every transaction as not-fraud will get you to 99% accuracy!

Inspect a SVM



- Notice the negative (blue) class swamps the classifier
- Boundary moves and disappears favouring only the negative class.

Use different metric

- F_1 score

$$F_1 = \frac{2TP}{2TP + FP + FN} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Area Under Curve (AUC) for the ROC
- Cohen's Kappa
 - $\kappa = \frac{p_0 - p_e}{1 - p_e}$
 - Compares expected to observed accuracy

What is Cohen's Kappa metric?

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

- p_o = observed agreement of classes = Overall Accuracy
- p_e = hypothetical expected agreement

Consider the simple case of binary classification (two classes) with n cases. Call the two judges, judges A and B .

- Judge A could denote the observed class assignment labels.
- Judge B could denote the predicted class assignment labels

p^o = Proportion of Agreement between judge A and B = This is the accuracy.

Pe: 所有类别分别对应的“实际与预测数量的乘积”，之总和，除以“样本总数的平方”

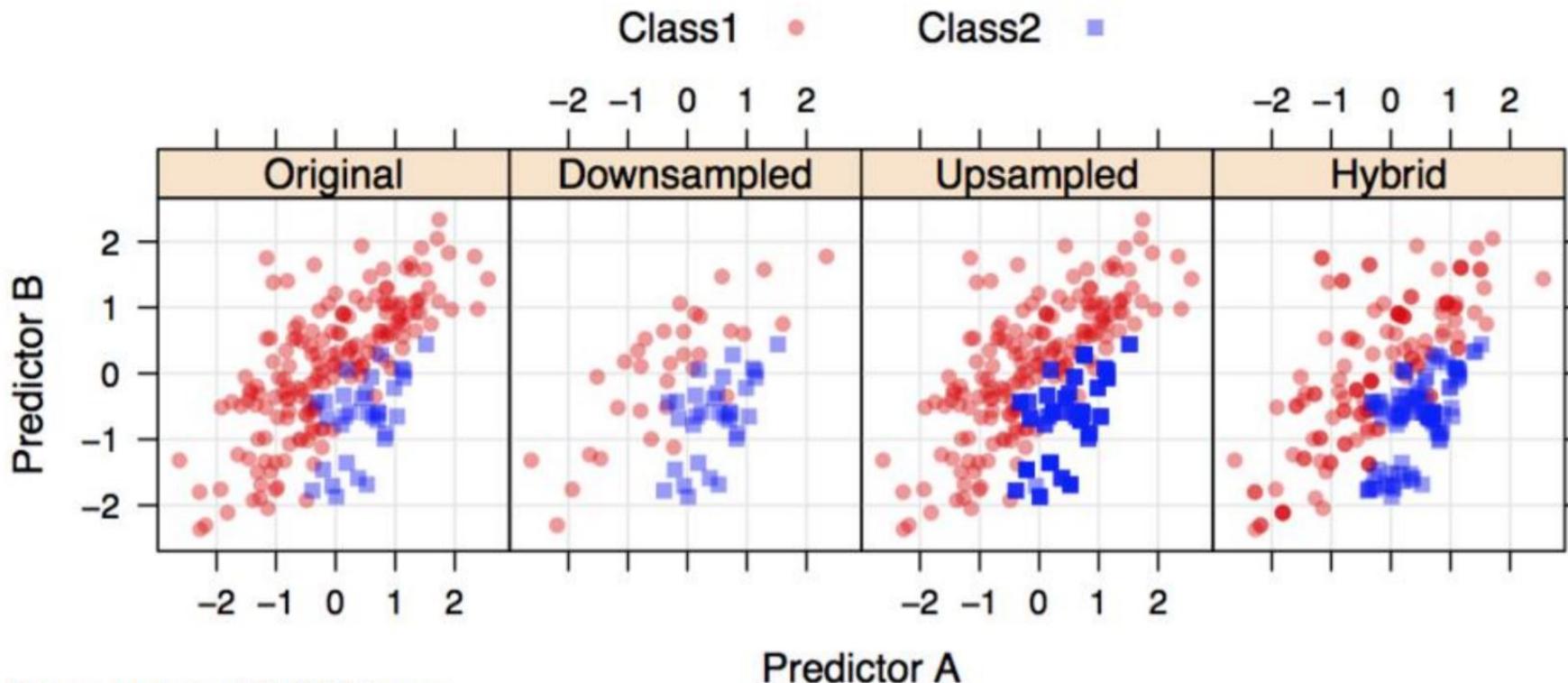
Cohen's Kappa: Empirical chance measure

The other metric p_e is the empirical chance these judges give the classification by chance.

- i.e. Computes the chance that Judge A and Judge B agree on a randomly picked element.
 - Chance that either both judges classify positive class or both judges classify negative class.
 - Assuming they pick at random.

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction      0      1
##           0 980000      0
##           1  20000      0
##
##                 Accuracy : 0.98
##                 95% CI : (0.9797, 0.9803)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1
##
##                 Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
```

Alternatively, modify the input data

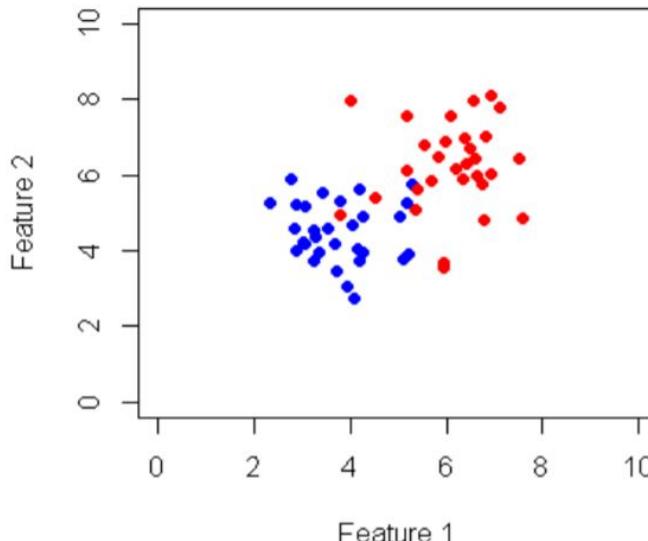
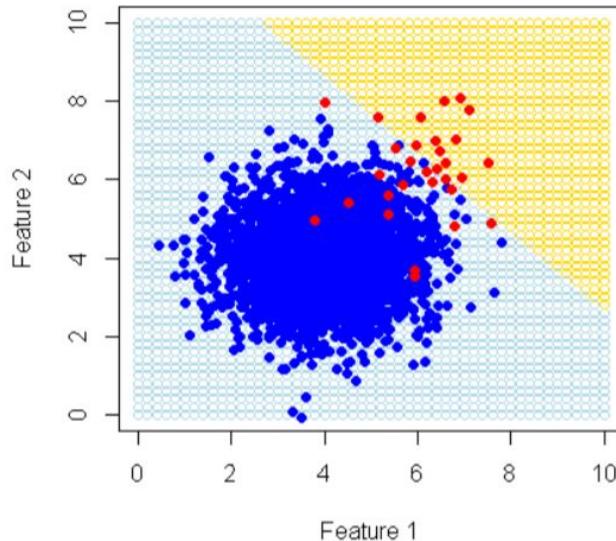


- Random up-sampling (增加分类少的observation)
- Disadvantages:
 - creates duplicated and/or artificial instances
 - can introduce bias and/or noise to the original data

产生重复的和/或艺术性的实例，给原始数据引入偏见和/或噪音

Down-sampling

(删减分类多的observation)



- Advantage is it does not introduce duplicates and/or artificial instances
- Disadvantages:
 - Not all data points are used.
 - Potentially removing useful information.
- Better choice for data with very high class imbalance.

优点是它不会引入重复和/或艺术性的实例劣势。不是所有的数据点都被使用。有可能删除有用的信息。对于类不平衡度非常高的数据来说是更好的选择。

Create synthetic samples of the minority class

- Synthetic Minority Over-sampling Technique (SMOTE) is a popular algorithm
- It creates synthetic samples from the minority class by:
 - Finding the k-nearest-neighbours for minority class observations
 - Randomly choosing one of the k-nearest-neighbours, then using it to create a similar but random new observation
- Be careful you split your data into training/validation before doing any oversample/SMOTE. Otherwise, you will leak information from training to validation data set.
- The R package DMwR implements SMOTE
 - See Torgo (2010)

为少数群体的观察值找到knn。随机选择其中一个最近的knn，然后用它来创建一个类似但随机的新的观察值。在做任过度采样/SMOTE之前，将你的数据分成训练/验证。否则，你会从训练数据集泄露信息到验证数据集

References

- Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC. URL: <http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR>.
- van Buuren, S. and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1-67. URL: <https://www.jstatsoft.org/v45/i03/>.

STAT5003

Week 8 : Feature Selection

Dr. Justin Wishart



Readings



- Feature selection covered in Chapter 6.1-6.2 in James, Witten, Hastie, and Tibshirani (2013)

Feature Selection



THE UNIVERSITY OF
SYDNEY

Goals of feature selection

- **Prediction accuracy:** especially when $p > n$
 - where p is the number of features and n denotes number of observations
- **Model interpretability:**
 - Removing irrelevant or poor features (that is, by setting the corresponding coefficient estimates to zero) \rightsquigarrow we can obtain a model that is more easily interpreted
- Some approaches for feature selection are presented.

Approaches for feature selection

1. Subset selection:

确定一个我们认为与反应或类别有关的预测因子子集。在减少的变量集上拟合一个分类或回归模型。

- Identify a subset of the p predictors that we believe to be related to the response or class (y).
- Fit a classification or regression model on the reduced set of variables.

2. Shrinkage:

- Primarily used for regression models
 - Fit a model involving all p predictors.
 - Some estimation coefficients are shrunk towards zero relative
 - This shrinkage (also known as regularisation) has the effect of reducing variance and can also be used for feature selection.
- 主要用于回归模型。拟合一个涉及所有预测因子的模型。一些估计系数被缩减为零。
这种收缩（也被称为正则化）具有减少方差的作用，也可用于特征选择。

3. Dimension reduction:

- We project the p predictors into M -dimensional subspace, $M < p$

将预测器 p 投射到 M 维的子空间, $M < p$

Best subset selection

1. Denote \mathcal{M}_0 to be the null model
 - Contains no predictors.
2. For $k = 1, 2, \dots, p$
 - Fit all $\binom{p}{k}$ models that contain exactly k predictors
 - Denote \mathcal{M}_k the best among the $\binom{p}{k}$ models.
 - Measured as best against some metric (smallest residual sum of squares or highest accuracy etc.)
3. Select the single best model among the $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$
 - Using cross-validated prediction error or Residual sum of squares etc.
 - Consider as an example Linear regression
 - $\mathcal{M}_0 : Y = \beta_0 + \varepsilon$
 - $\mathcal{M}_p : Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$

Best subset selection methods

- It can be too computationally expensive to apply best subset selection when p is large.
 - Too many possible feature subsets. 如果 p 较大，可以使用best subset selection
- Statistical problems with large p
 - Larger search space \rightsquigarrow increased chance of finding models that overfit.
 - Perform well on training data

Forward Stepwise selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.

正向逐步选择从一个不包含任何预测因子的模型开始，然后将预测因子逐一添加到模型中。到模型中，直到所有的预测因子都在模型中。特别是，在每一步中，对拟合有最大额外改善的变量被添加到模型中。

Forward stepwise selection

1. Denote \mathcal{M}_0 to be the null model (e.g. $Y = \beta_0 + \varepsilon$ in linear reg)
 - Contains no predictors.
2. For $k = 0, 1, 2, \dots, p - 1$
 - Consider all $p - k$ models that augment the predictors in \mathcal{M}_k with one additional predictor
 - Choose the *best* among these $p - k$ models and assign it as \mathcal{M}_{k+1} .
 - Best measured against some metric (RSS or classification error)
3. Select the single best model among the $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$
 - Using cross-validated prediction error or Residual sum of squares etc.
 - Computational advantage over best subset selection is clear.
 - However,
 - not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
 - Why not?

Backward stepwise selection

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.
- However, unlike forward stepwise selection,
 - begins with the **full model** containing all p predictors,
 - iteratively **removes** the **least useful** predictor, one-at-a-time.

与前向逐步选择一样，后向逐步选择也提供了一种有效的替代方法。最佳子集选择。然而，与前向逐步选择不同的是，后向逐步选择从包含所有预测因子的完整模型开始，逐次迭代地删除最不有用的预测因子。

Backward stepwise selection

1. Denote \mathcal{M}_p to be the **full** model (e.g. $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ in linear reg)
 - Contains all predictors.
2. For $k = p, p - 1, \dots, 1$
 - Consider all k models that contain all but one of the predictors in \mathcal{M}_k , for a total of $k - 1$ predictors
 - Choose the *best* among these k models and assign it as \mathcal{M}_{k-1} .
 - Best measured against some metric (RSS or classification error)
3. Select the single best model among the $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$
 - Using cross-validated prediction error or Residual sum of squares etc.

More on backward stepwise selection

- Similarities to forward selection
 - it searches through only $1 + p(p + 1)/2$ models
 - can be applied in settings where p is too large to apply best subset selection
 - backward selection is not guaranteed to yield the best model containing a subset of the p predictors.
- **Note:** for some models such as linear regression, backward selection requires that the number of cases n is larger than the number of features p (so that the full model can be fit).
 - In contrast, forward stepwise can be used even when $n < p$.

添加或减少

Linear model (feature) selection

- Recall the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

How to choose the optimal model?

- The model containing all of the predictors will always have **the smallest RSS**, since these quantities
- We wish to choose a model with **low test error**, not a model with low training error.
 - Training error is usually a poor estimate of test error.
- Therefore, RSS are not suitable for selecting the best model among a collection of models with different number of predictors.

希望选择一个具有低测试误差的模型

Estimating test error - two approaches

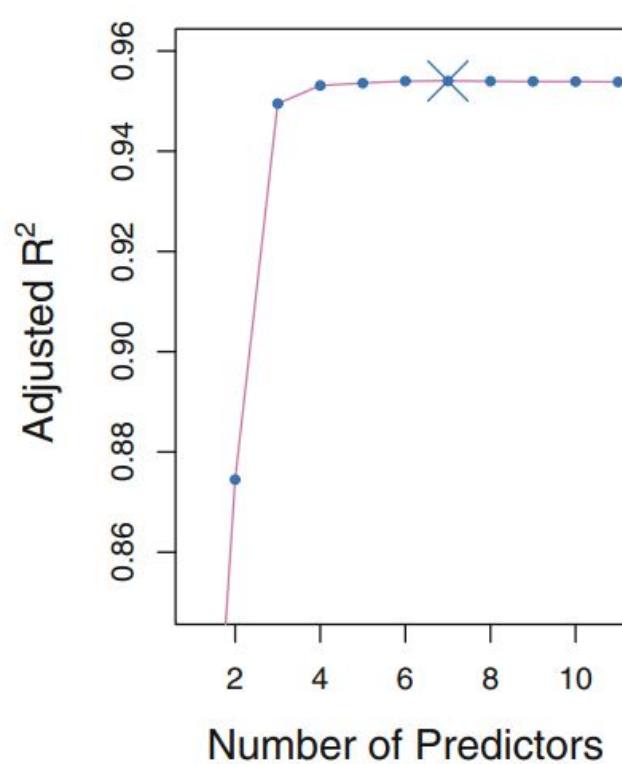
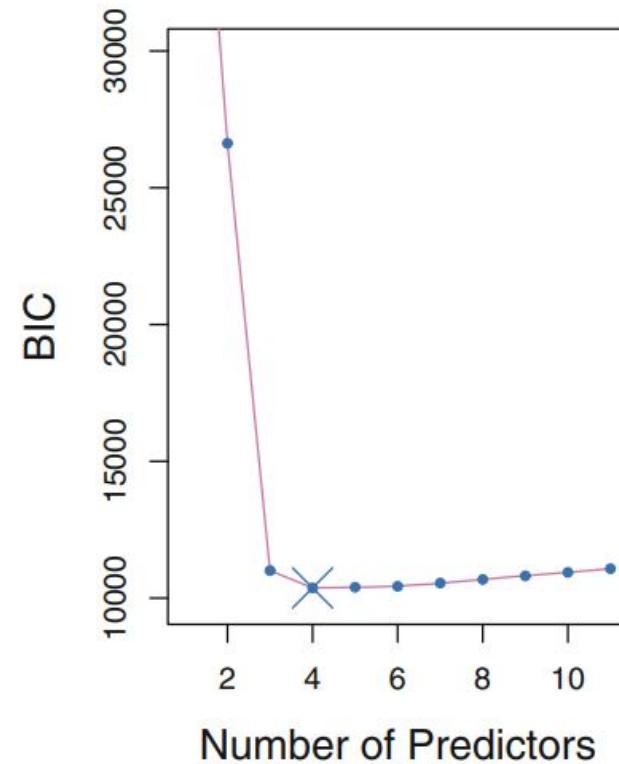
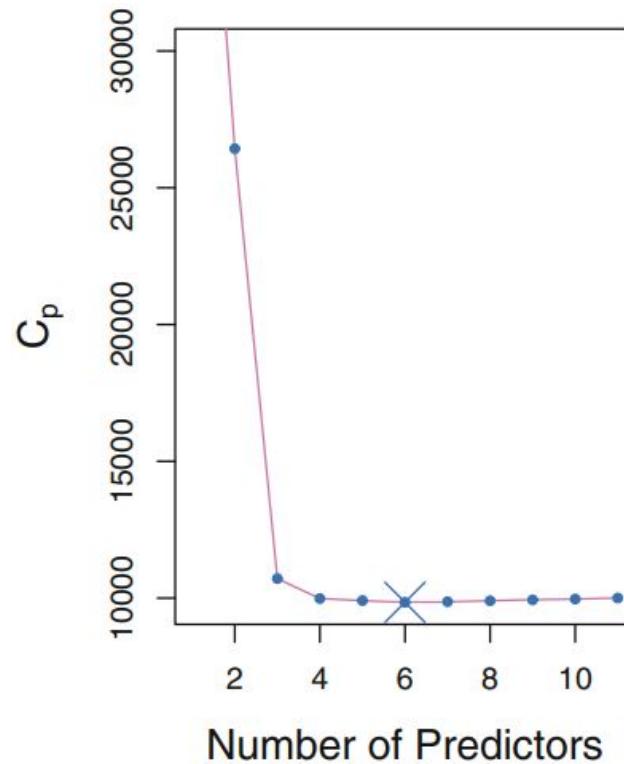
- **Indirectly** estimate test error by making an adjustment to the training error.
 - Account for the bias due to overfitting.
- **Directly** estimate the test error, using either a test set or cross-validation set approach (covered in previous lectures)
- Will illustrate the **indirect** approach and also review cross-validation.

通过对训练误差进行调整来间接估计测试误差。考虑到由于过度拟合而产生的偏差。

直接估计测试误差，使用测试集或交叉验证集的方法

Indirect approaches (e.g. C_p and BIC)

- Adjust the training error for the model size (model complexity)
 - Can be used to select among a set of models with a different number of features
- Figure below displays Mallow's C_p and the Bayesian Information Criterion (BIC) and adjusted R^2 for the best model produced by best subset selection on the credit data set.



Details of C_p and BIC

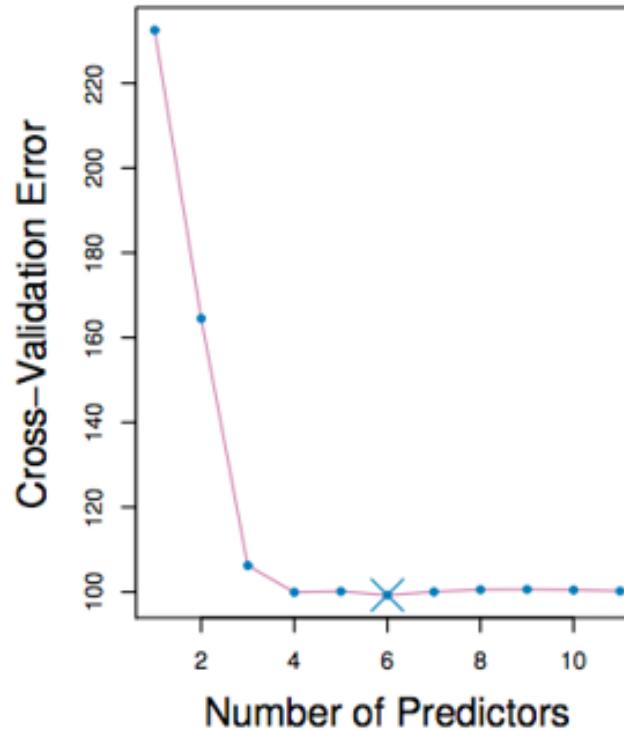
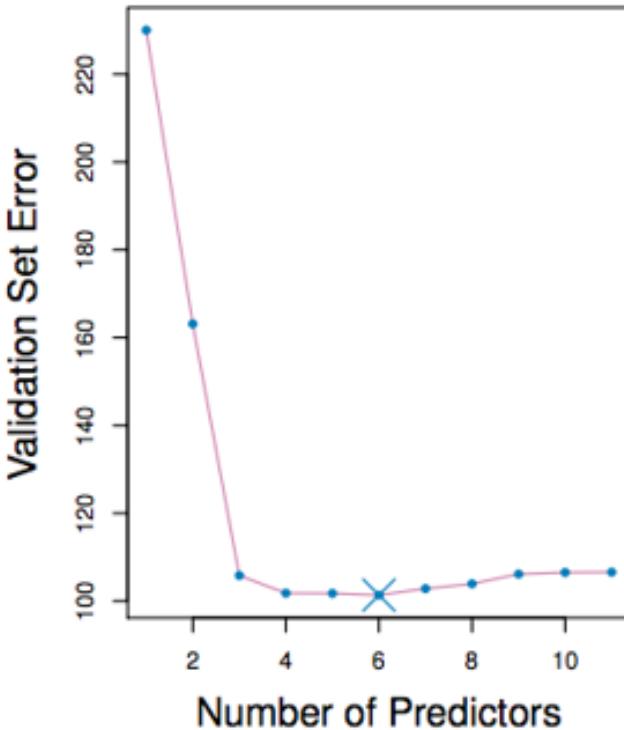
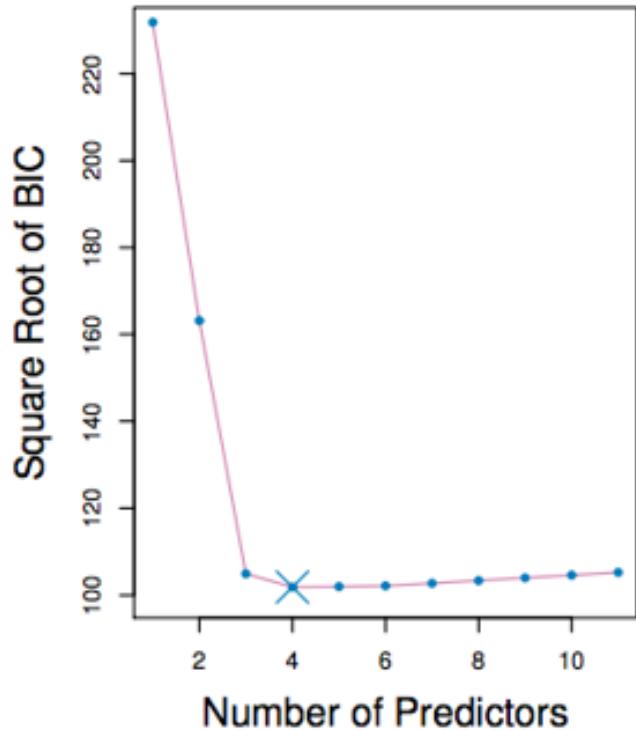
- Mallow's $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$ 梅洛 Cp 统计量
 - d is the total number of parameters
 - $\hat{\sigma}^2$ is an estimate of the variance of ε
- Bayesian Information Criterion $BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$
- Like C_p , the BIC will tend to take on small value for model with a low test error, and so generally we select model that has the lowest BIC value.
- Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term, where n is the number of samples.
- Since $\log n > 2$ when $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p .

Test set and cross-validation

- Each of the procedures returns a sequence of models \mathcal{M}_k indexed by model size $k = 0, 1, 2, \dots$. Our job here is to select k . Once selected, we will return model \mathcal{M}_k .
- We compute the validation set error or the cross-validation error for each model \mathcal{M}_k
 - select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to C_p and BIC, in that it provides direct estimate of the test error, and doesn't require an estimate of the error variance σ^2 .
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2 .

将sample重复

Credit card example



Shrinkage methods

<https://zhuanlan.zhihu.com/p/30535220>

- We will introduce two methods specifically designed for linear regression

Ridge-regression and Lasso

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Ridge regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \dots, \beta_p$ using the values that minimize

$$RSS = \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 = \|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- The ridge regression coefficient estimates $\hat{\boldsymbol{\beta}}_R$ are the values that minimize

$$RSS + \lambda \sum_{j=1}^p \beta_j^2$$

- where $\lambda \geq 0$ is a **tuning** parameter, to be determined separately.

岭回归将在最终模型中包括所有预测因子

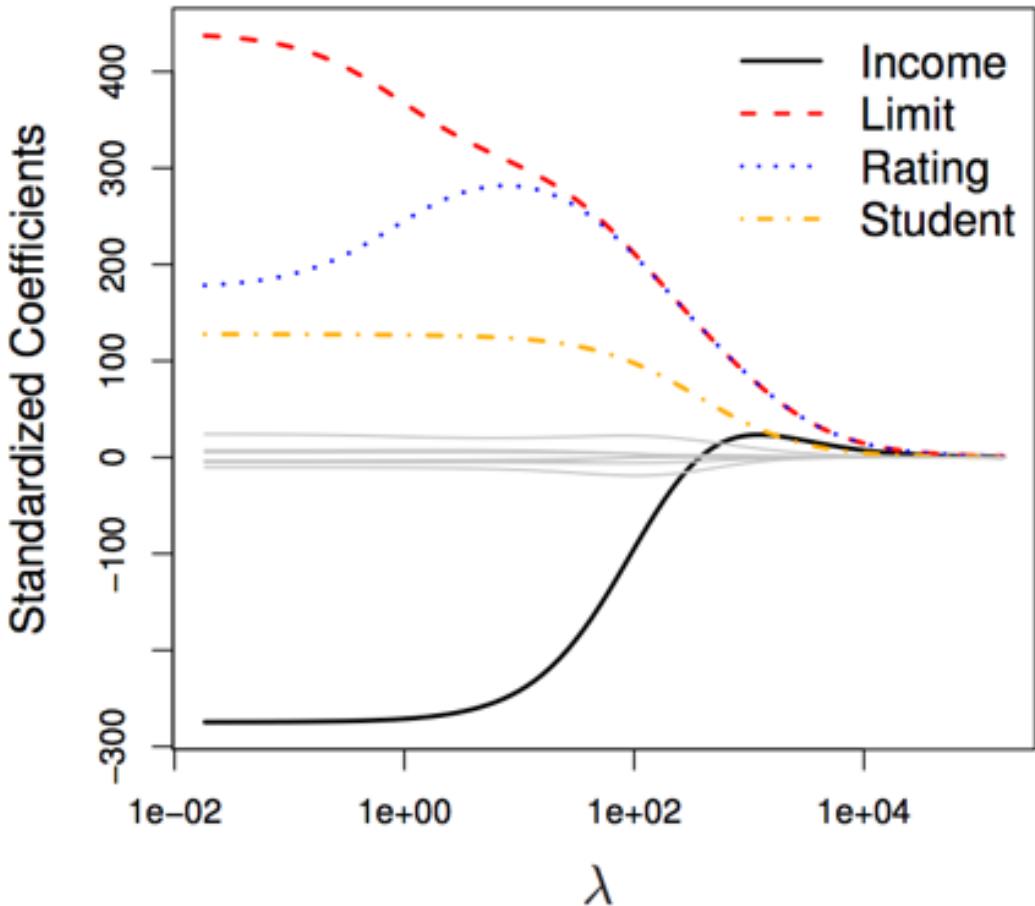
当岭参数 lamda时，得到的解是最小二乘解

当岭参数 lamda 趋向更大时，岭回归系数 wi趋向于0，约束项 t 很小

Ridge regression

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term, $\lambda \sum_{j=1}^p \beta_j^2$, is called a shrinkage penalty
 - is small when β_1, \dots, β_p are close to zero, and so it has the effect of shrinking the estimates of β_j towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for λ is critical; cross-validation is used for this.

Credit card example



在最左边 [lambda系数最小时, 可以得到所有系数的原始值(与标准线性回归相同);而在右边, 系数全部缩减为0, 从不稳定趋于稳定;

Ridge regression: scaling of predictors

- The standard least squares coefficient estimates are scale **invariant**:
 - multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$.
 - In other words, regardless of how the j^{th} predictor is scaled, $X_j, \hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficients estimates can change **substantially** when multiplying a given predictor by a constant,
 - due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after **standardising the predictors**, using a formula such as below:

$$\widetilde{X_{ij}} = \frac{X_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}}$$

The Lasso

- Ridge regression does have one obvious disadvantage:
 - Ridge regression will include all p predictors in the final model
 - Subset selection will generally select models that involve a subset of the predictors
- The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage.
 - The lasso coefficients, $\hat{\beta}_L$, minimise the quantity

$$RSS + \lambda \sum_{j=1}^p |\beta_j|$$

lasso 比reige 区别在它的有的变量的lambda很快趋于0了，适合于做变量选择， reige的需要lambda很大才趋向于0

- The lasso uses an ℓ_1 penalty instead of the ℓ_2 penalty.
 - The ℓ_1 norm of a coefficient vector β is $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$

LASSO(The Least Absolute Shrinkage and Selection Operator)是另一种缩减方法，将回归系数收缩在一定的区域内。LASSO的主要思想是构造一个一阶惩罚函数获得一个精炼的模型, 通过最终确定一些变量的系数为0进行特征筛选。

The Lasso

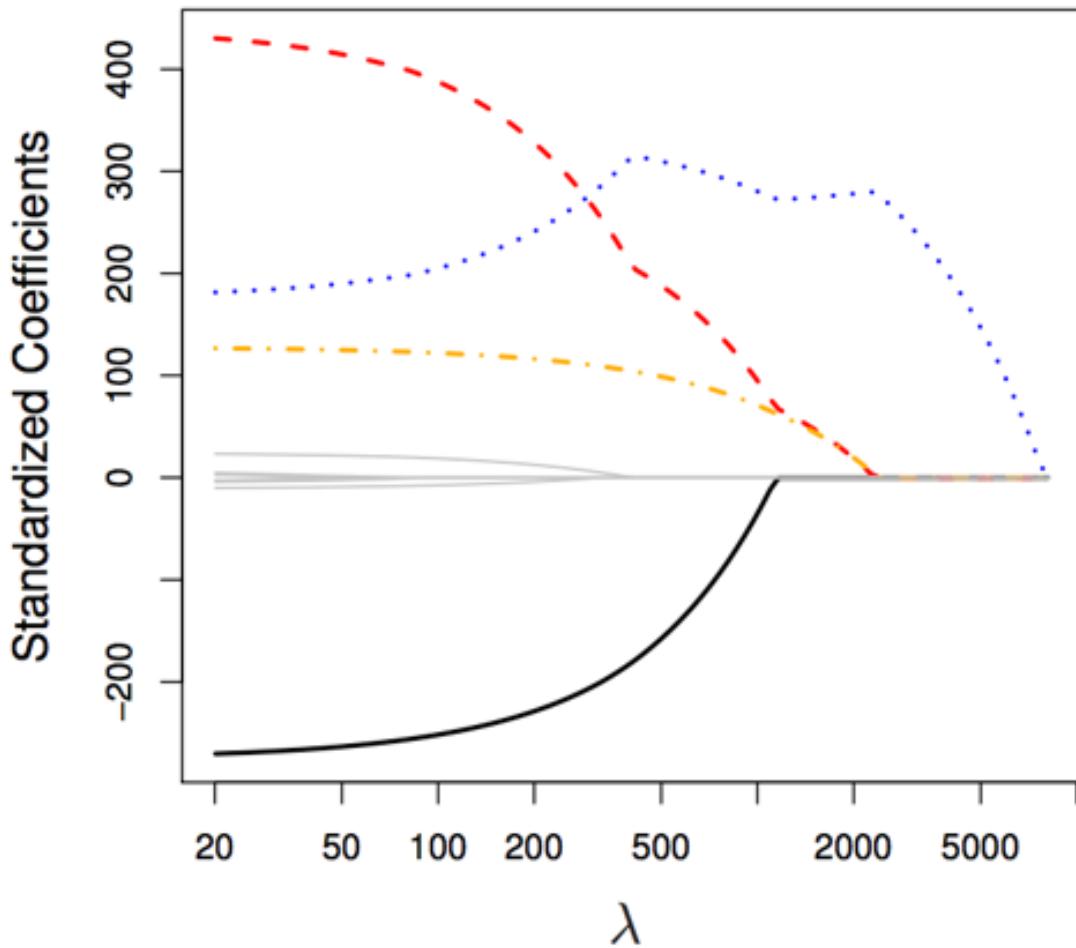
- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, the lasso performs **feature selection** (in an embedded manner).
- We say that the lasso yields **sparse** models – that is, models that involve only a subset of variables.
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice.

至于子集选择，对于岭回归和套索，我们需要一种方法来确定哪一个的模型是最好的。也就是说，我们需要一种方法来选择调整参数的值，或者说，选择约束条件的值。交叉验证法提供了一个解决这个问题的简单方法。我们选择一个值的网格，然后计算每个值的交叉验证错误率。

然后我们选择交叉验证误差最小的调整参数值。

最后，使用所有可用的观测数据和选定的调谐参数值对模型进行重新测试

Example: Credit dataset



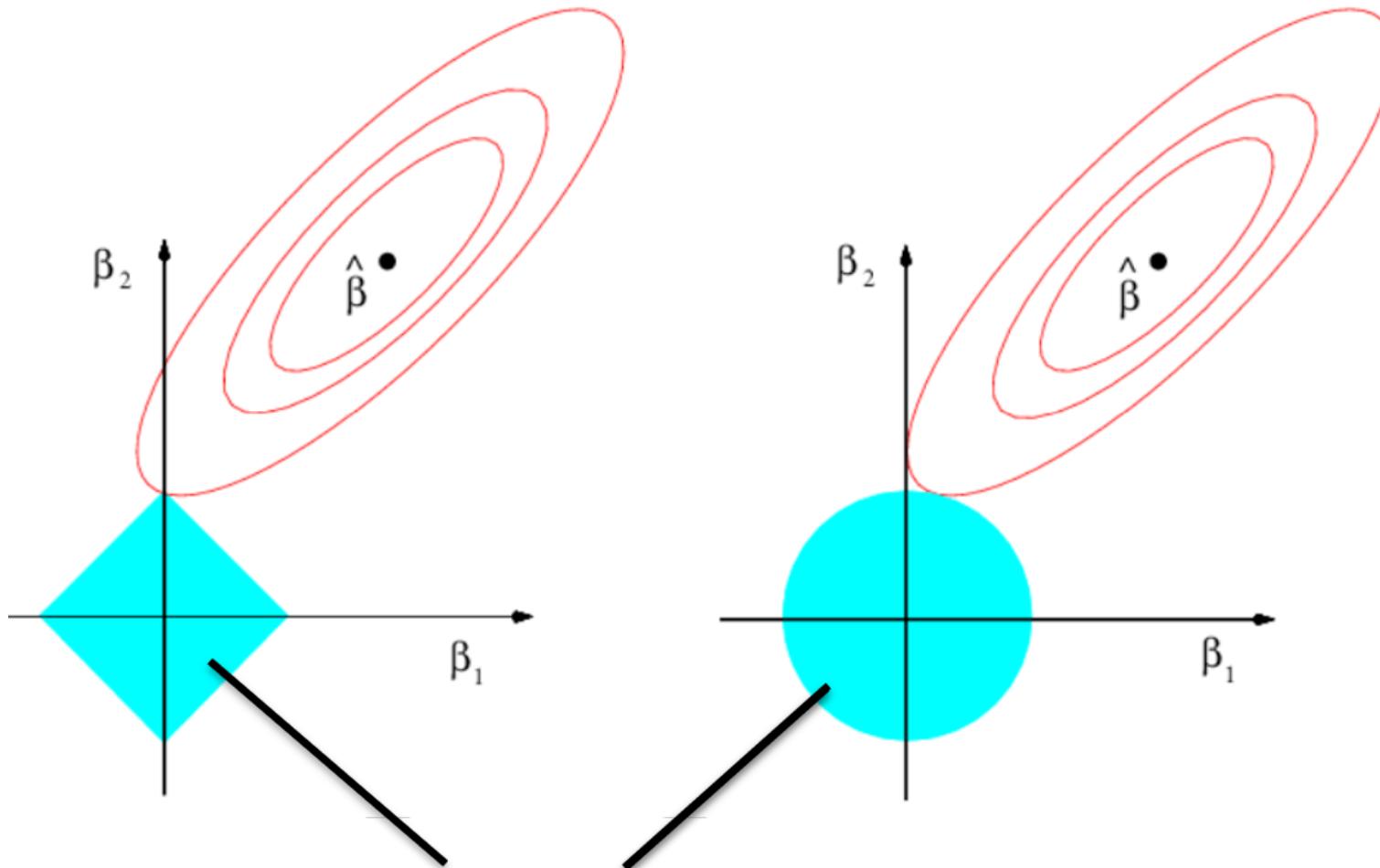
Variable selection property of the Lasso

- Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?
- One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\min_{\beta} \sum_{i=1}^n (Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2 \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s.$$

$$\min_{\beta} \sum_{i=1}^n (Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2 \quad \text{subject to} \quad \sum_{j=1}^p \beta_j^2 \leq s.$$

Comparison of ℓ_1 and ℓ_2 constraints



这个时候等值线与圆相切的点便是在约束条件下的最优点，

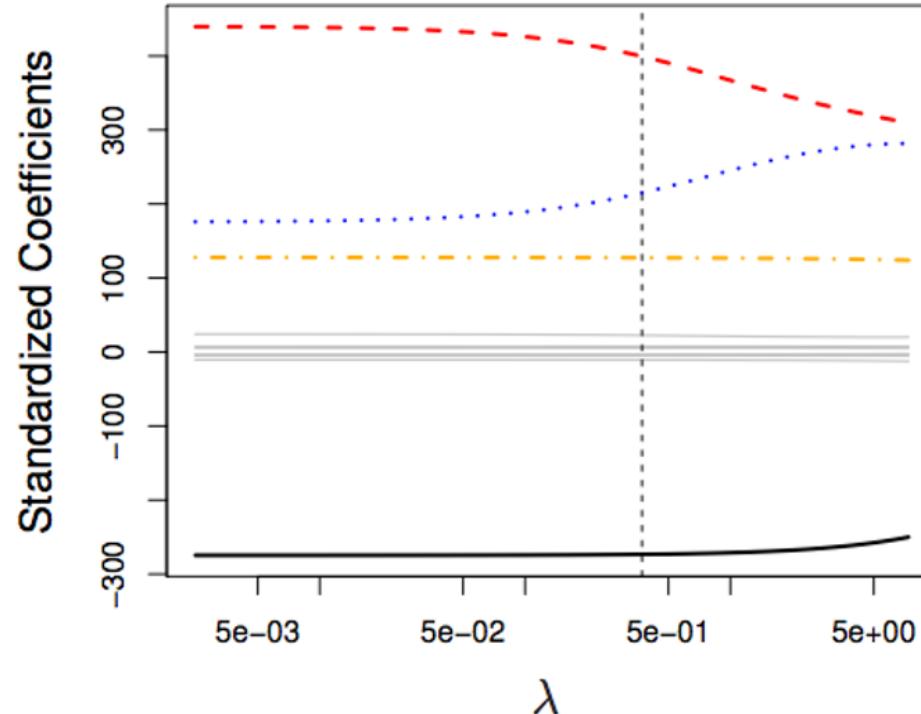
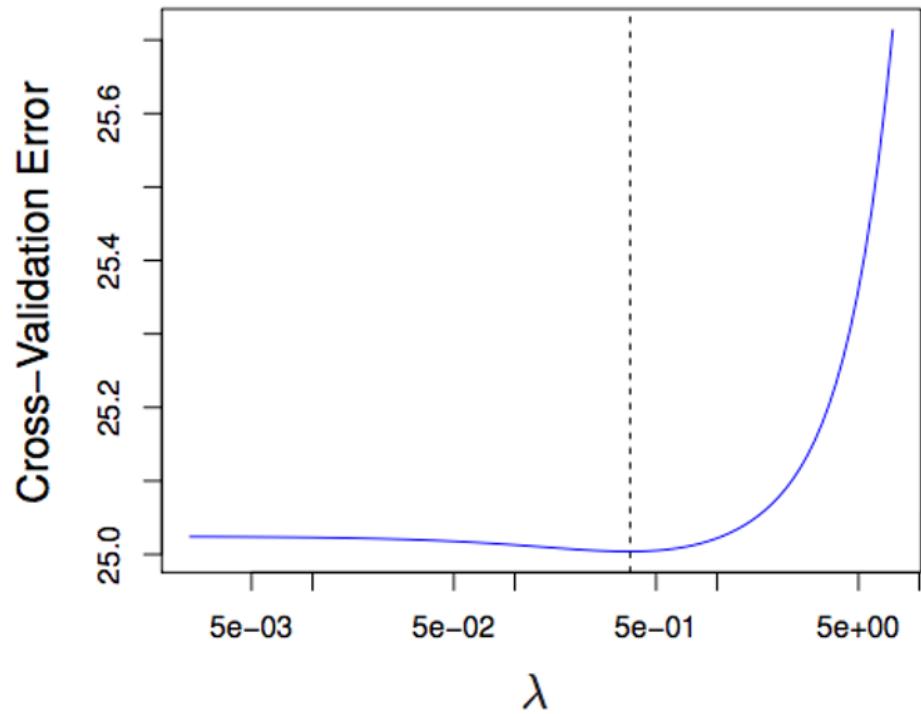
相比圆，方形的顶点更容易与抛物面相交，顶点就意味着对应的很多系数为0，而岭回归中的圆上的任意一点都很容易与抛物面相交很难得到正好等于0的系数。这也就意味着，lasso起到了很好的筛选变量的作用。

- Solution is feasible if it is within these blue regions for the Lasso (left) and Ridge (right) respectively.

Selecting the tuning parameters for Ridge regression and Lasso

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is the best.
- That is, we require a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint s .
- **Cross-validation** provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ .
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter

Credit data example



纵向虚线表示交叉验证选择的最佳值。

- Left illustrates cross-validation errors that result from applying ridge regression to the Credit data set with a range of λ values.
- Right illustrate the coefficient estimates as a function of λ . The vertical dashed lines indicate the best value of λ selected by cross-validation.

Elastic net in `glmnet`

- In `glmnet` (see Friedman, Hastie, and Tibshirani (2010)) the implementation is actually for a more general model called the Elastic net.
- It solves the following penalised minimization problem

$$\arg \min_{\beta} \|Y - \mathbf{X}\beta\|_2^2 + \lambda((1 - \alpha)/2\|\beta\|_2^2 + \alpha\|\beta\|_1)$$

- Can consider it a weighted combination (mixture) of ℓ_1 and ℓ_2 penalties.
- Objective function actually more general than this as the RSS term can be further weighted

References

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1-22. URL:
<http://www.jstatsoft.org/v33/i01/>.

James, G., D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

STAT5003

Week 9 : Tree and Ensemble methods

Dr. Justin Wishart



THE UNIVERSITY OF
SYDNEY



Readings



- Tree methods covered in Chapter 8 in James, Witten, Hastie, and Tibshirani (2013)

Regression/Decision Trees



THE UNIVERSITY OF
SYDNEY

Decision tree for regression

- Baseball player salary data: how would you stratify it?
 - **Salary** is colour-coded (NA values coded grey).

2.3 剪枝策略

为什么要剪枝：过拟合的树在泛化能力的表现非常差。

2.3.1 预剪枝

在节点划分前 来确定是否继续增长，及早停止增长的主要方法有：

节点内数据样本低于某一阈值；

所有节点特征都已分裂；

节点划分前准确率比划分后准确率高。

预剪枝不仅可以降低过拟合的风险而且还可以减少训练时间，但另一方面它是基于“贪心”策略，会带来欠拟合风险。

2.3.2 后剪枝

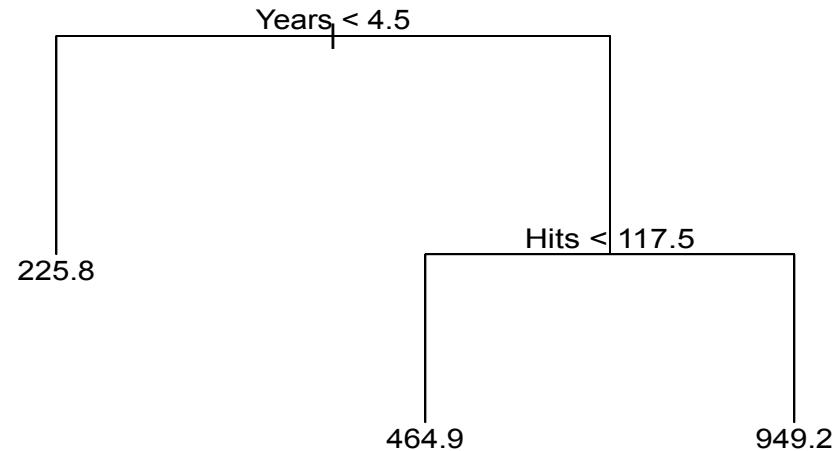
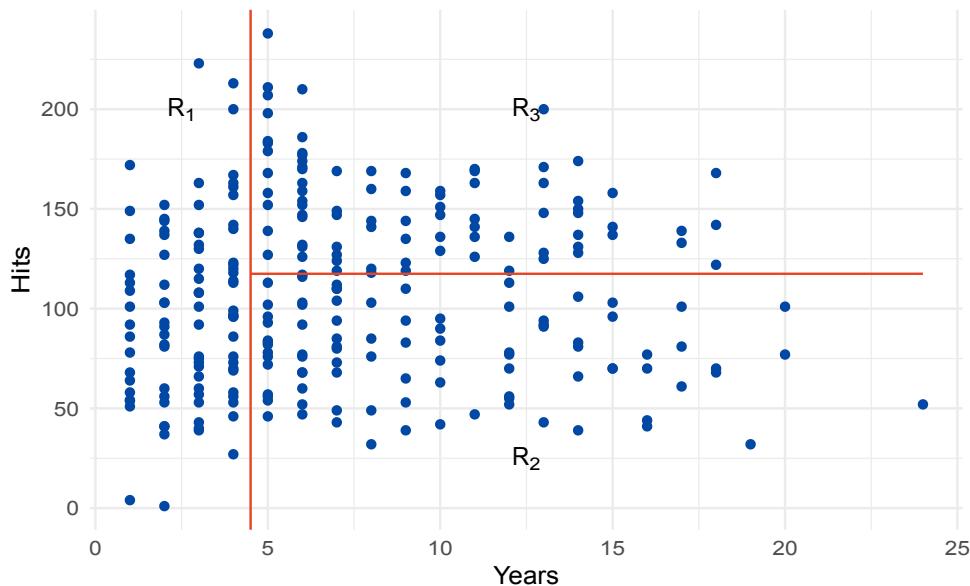
在已经生成的决策树上进行剪枝，从而得到简化版的剪枝决策树。

C4.5 采用的悲观剪枝方法，用递归的方式从低往上针对每一个非叶子节点，评估用一个最佳叶子节点去代替这课子树是否有益。如果剪枝后与剪枝前相比其错误率是保持或者下降，则这棵子树就可以被替换掉。C4.5 通过训练数据集上的错误分类数量来估算未知样本上的错误率。

后剪枝决策树的欠拟合风险很小，泛化性能往往优于预剪枝决策树。但同时其训练时间会大的多。

Decision Tree

- Overall, the tree segments the players into three regions of predictor space:
 - $R_1 = \{X | \text{Years} < 4.5\}$
 - $R_2 = \{X | \text{Years} \geq 4.5, \text{Hits} < 117.5\}$
 - $R_3 = \{X | \text{Years} \geq 4.5, \text{Hits} \geq 117.5\}$



Terminology for trees

- In keeping with the tree analogy,
 - the regions R_1 , R_2 and R_3 are known as terminal nodes.
 - Decision trees are typically drawn upside down, in the sense that the leaves are at the bottom of the tree.
 - The points along the tree where the predictor space is split are referred to as internal nodes.
- In the Baseball player salary tree, the two internal nodes are indicated by the text Years < 4.5 and Hits > 117.5.

Interpretation of Results

- **Years** is the most important factor in determining **Salary**,
 - Players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced,
 - the number of **Hits** that he made in the previous year seems to play little role in his **Salary**.
- Among players who have been in the major leagues for five or more years,
 - the number of Hits made in the previous year does affect Salary,
 - players who made more Hits last year tend to have higher salaries.
- Obviously an over-simplification,
 - compared to some other classification models (such as a regression model), it is easy to display, interpret and explain.

Details on tree building process

- In theory, the decision regions could have any shape. However, we choose to divide the predictor space into high-dimensional rectangles, or boxes, for simplicity and for ease of interpretation of the resulting predictive model
- The goal is to find boxes R_1, \dots, R_J that minimizes the residual sum of squares (RSS), given by:

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

- where \hat{y}_{R_j} is the mean response for the training observations within the j^{th} box.

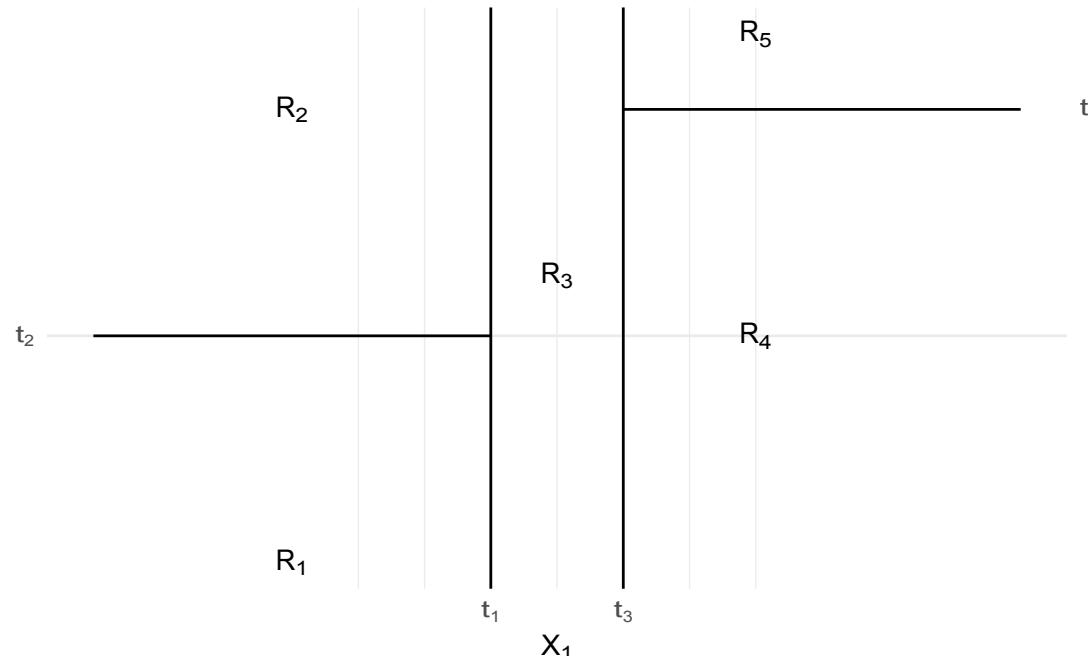
Tree building with recursive binary splitting

- It is computationally **infeasible** to consider every possible partition of the feature space into J boxes
- Take a top-down, **greedy** approach that is known as **recursive binary splitting**
- The approach is **top-down** because it begins at the top of the tree and then successively splits the predictor space; each split is indicated via two new branches further down on the tree.
- It is **greedy** because at each step of the tree-building process, the best split is made at that particular step,
 - rather than looking ahead and picking a split that will lead to a better tree in some future step.

采取一种自上而下的、贪婪的方法，即所谓的递归二元分割法**recursive binary splitting**。这种方法是自上而下的，因为它从树的顶端开始，然后依次分割预测空间。每次分割都通过树上的两个新分支来表示。它是贪婪的，因为在建树过程的每一步，都会在该特定的步骤中进行最佳分割。而不是向前看，挑选一个在未来某个步骤中会导致更好的树的分裂。

Using Decision Trees for prediction

- We divide the predictor space
 - the set of possible values for X_1, X_2, \dots, X_p into J distinct and non-overlapping regions, R_1, R_2, \dots, R_J .
- For every observation that falls into the region R_j , we make the same prediction, which is simply the mean of the response values for the training observations in R_j .



Decision trees for classification

- Very similar to a regression tree,
 - exception being the prediction of a qualitative response rather than a quantitative one.
- For a classification tree,
 - Inspect the region that the observation belongs and predict the most commonly occurring class in that region.

用于分类的决策树与回归树非常相似。

例外的是对定性反应的预测，而不是定量反应。

对于一个分类树来说。检查观察结果所属的区域，并预测该区域中最常出现的类别。

Gini index

在分类设置中，RSS不能作为进行二元拆分的标准。如基尼指数被用来代替。

- Just as in the regression setting, we use recursive binary splitting to grow a classification tree.
- In the classification setting, RSS cannot be used as a criterion for making the binary splits
- Alternative measure such as Gini index is used instead.
- The Gini index is defined by

$$G = \sum_j \sum_{k=1}^K \hat{p}_{jk} (1 - \hat{p}_{jk})$$

- where \hat{p}_{jk} represents the proportion of training observations in the j^{th} region that are from the k^{th} class.

Gini index

- The Gini index is a measure of total variance across the K classes.
 - It takes on a small value if all of the \hat{p}_{jk} values are close to zero or one.
 - This occurs when there is a clear majority class!
- For this reason the Gini index is referred to as a measure of node **purity**
 - a small value indicates that a node contains predominantly observations from a single class.

基尼指数是对K类中总差异的衡量。

如果所有的 p_{jk} 值都接近于零或一。

当有一个明显的多数阶级时就会出现这种情况!

由于这个原因，基尼指数被称为衡量节点纯度的一个指标

一个小的值表示一个节点主要包含来自一个类的观察值

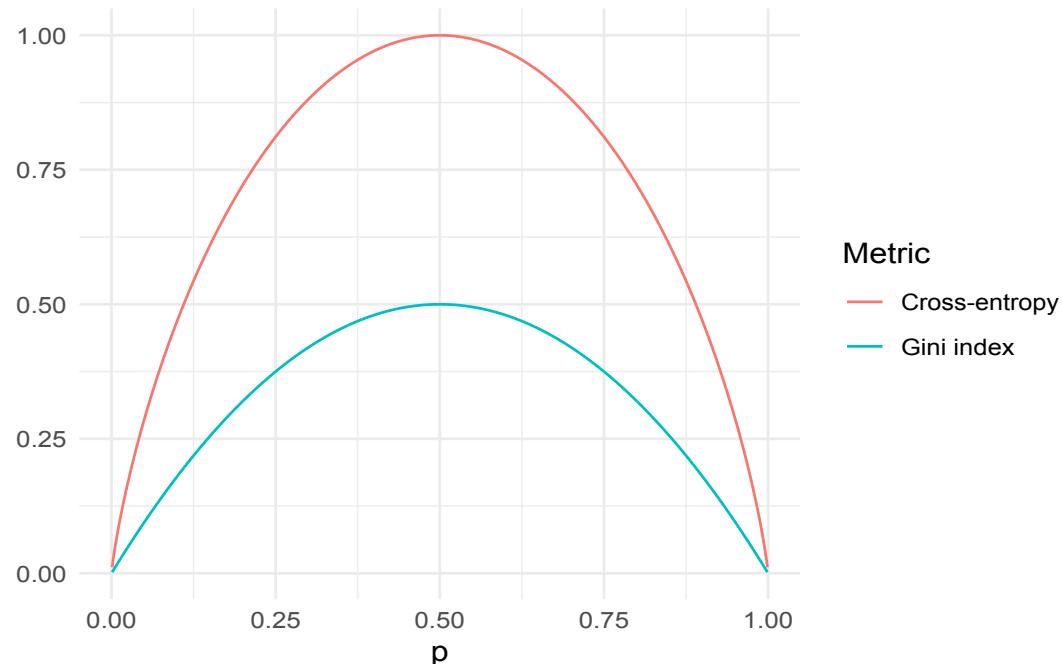
Cross-entropy

吉尼指数的一个替代方法是交叉熵

- An alternative to the Gini index is cross-entropy, given by

$$D = - \sum_m \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$$

- It turns out that the Gini index and the cross-entropy are very similar numerically.



Tree vs linear model

针对上述离散型数据，按照体温为恒温和非恒温进行划分。其中恒温时包括哺乳类 5 个、鸟类 2 个，非恒温时包括爬行类 3 个、鱼类 3 个、两栖类 2 个，如下所示我们计算 D1, D2 的基尼指数。

$$\text{Gini}(D1)=1 \quad [(5/7)^2+(2/7)^2]=20/49$$

$$\text{Gini}(D2)=1 \quad [(3/8)^2+(3/8)^2+(2/8)^2]=42/64$$

然后计算得到特征体温下数据集的 Gini 指数。

$$\text{Gain_Gini}(D, \text{体温})=7/15*20/49+8/15*42/64$$

cart tree 实例：<https://murphypei.github.io/blog/2019/04/cart-tree.html>

Advantages and disadvantages of trees

Advantages:

- Trees are very easy to explain to people
- Some people believe that decision trees closely relate to human decision-making
- Trees can be displayed graphically
- Can handle different data types and doesn't require scaling

树很容易向人们解释。有些人认为，决策树与人类的决策密切相关

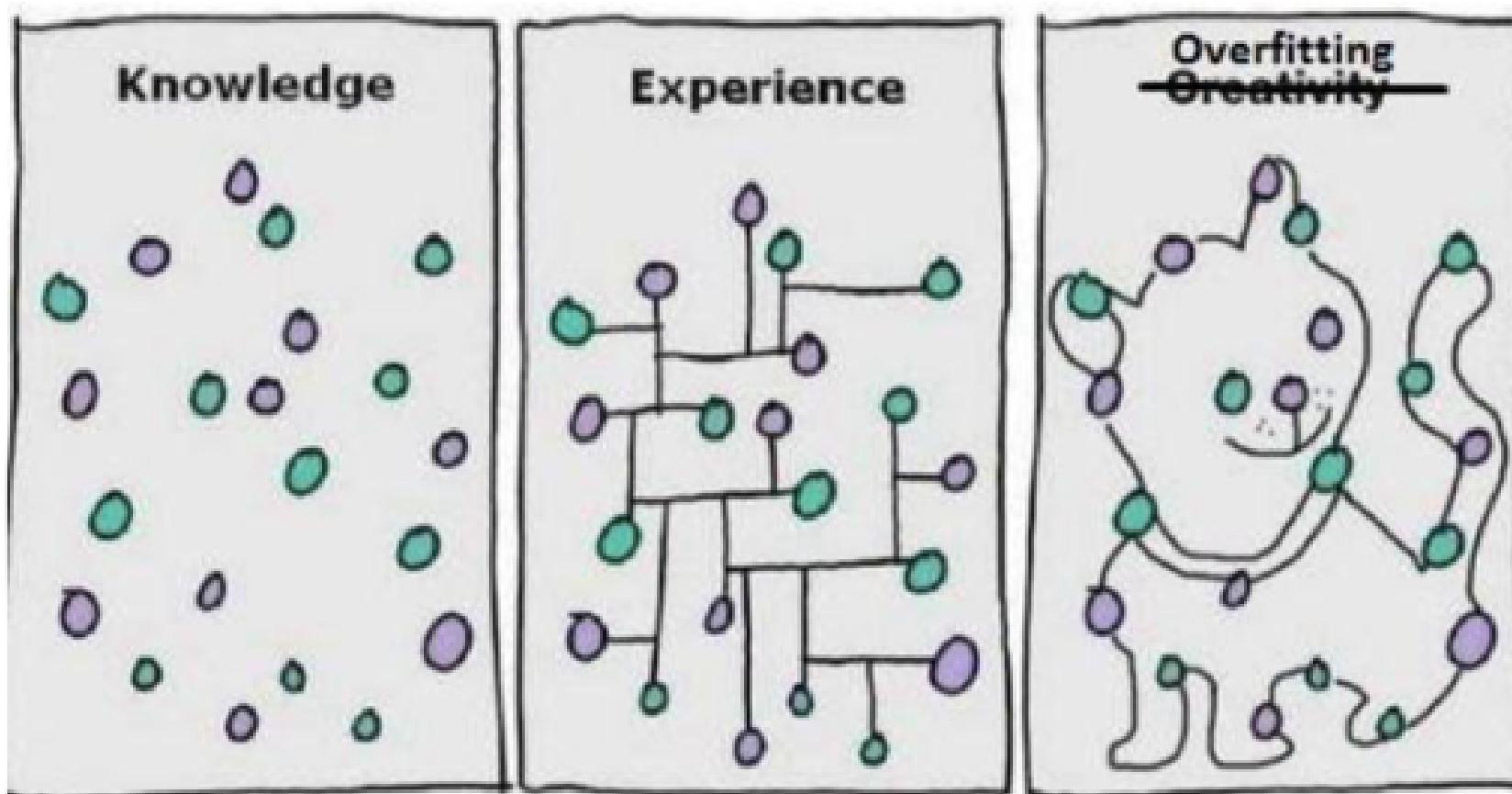
树可以用图形显示。可以处理不同的数据类型，不需要缩放

Disadvantages:

- Trees do not have the same level of predictive accuracy as some of the other regression and classification approaches we have discussed.

树的预测准确性不如我们讨论过的其他一些回归和分类方法

A single decision tree is prone to over-fitting



Ensemble methods



THE UNIVERSITY OF
SYDNEY

Ensemble of trees

- An alternative is available than just relying on one tree and hope we make the right decision at each split.

除了依赖一棵树并希望我们在每一次分裂时做出正确的决定外，还有一个替代方案。

Ensemble Methods

- Take a sample of Decision Trees into account
- Calculate which features to use at each split
- Make final prediction model using the **aggregated** result from an **ensemble** of trees.

将决策树的样本考虑在内

计算在每次分割时使用哪些特征

使用树群的汇总结果建立最终的预测模型

Bagging (Bootstrap Aggregation)

- Bootstrap aggregation, or bagging, is a general purpose procedure for reducing the variance of a statistical learning method. It is particularly useful and frequently used in the context of decision trees.
- Recall that given a set of n observations Z_1, \dots, Z_n each with a variance of σ^2 ,
 - the variance of the mean \bar{Z} is given by σ^2/n
- In other words, averaging a set of observations reduces variance.
- Since it is typically not possible to have access to multiple training sets,
 - we can use **bootstrapping** to create multiple training sets.

从所有样本中随机抽取一个样本
将1中样本放回到样本集中，充分混合。
重复步骤1和2，直到满足采样数目要求

对一组观察值进行平均化可以减少方差。由于通常不可能获得多个训练集。我们可以使用引导法来创建多个训练集。使用bootstrapping从一个（单一的）训练数据集中重复取样。在这种方法中，我们产生不同的引导训练数据集。

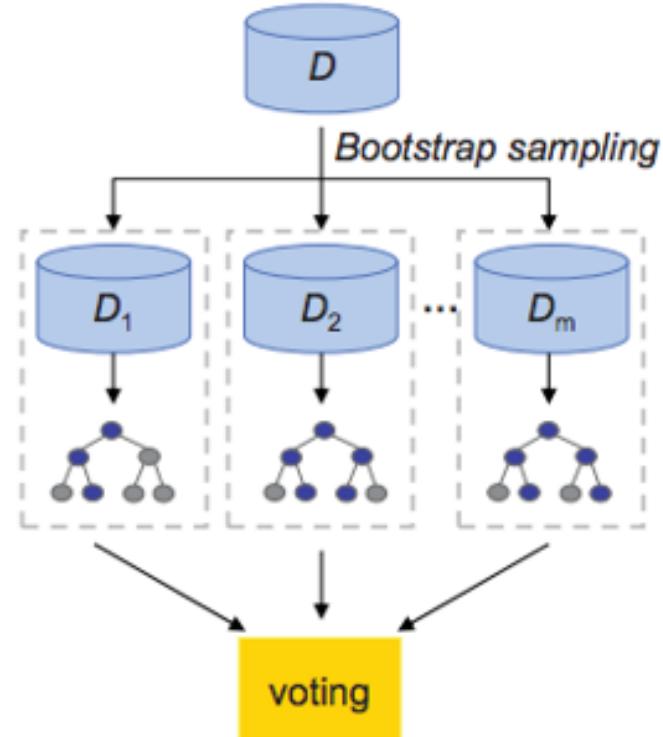
Bagging continued

- Use bootstrapping to take repeated samples from a (single) training data set.
- In this approach, we generate B different bootstrapped training data sets.
 - Train our method of the b^{th} bootstrapping set in order to obtain $\hat{f}_b^*(x)$, the prediction at a point x .
- Average all the observations to obtain:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}_b^*(x)$$

- This is called bagging

对于一个样本，它在一个包含 n 个样本的训练集中随机采样，则每次被采到的概率是 $1/n$ ，则不被抽取的概率是 $1 - 1/n$ 。若抽取次数和数据集相同，则抽取完毕，一个样本不被抽取的概率是 $(1 - 1/n)^n$ ，当 $m \rightarrow \infty$ ， $(1 - 1/n)^m \rightarrow e^{-1} \approx 0.368$ 。



Out of bag error estimation

事实证明，有一个非常直接的方法来估计一个袋装模型的测试误差。
回顾一下，装袋的关键是树被重复到引导的子集的
观察结果。我们可以看到，平均而言，每棵袋装树都使用了大约三分之二的

- It turns out that there is a very straightforward way to estimate the test error of a bagged model.
- Recall that the key to bagging is that trees are repeatedly fit to bootstrapped subsets of the observations. One can show that on average, each bagged tree makes use of around two-thirds of the observations.
- The remaining one-third of the observations not used to fit a given bagged tree are referred to as the out-of-bag (OOB) observations.
- We can predict the response for the i^{th} observation using each of the trees in which that observation was OOB. This will yield around $B/3$ predictions for the i^{th} observation, which we average. This estimate is essentially the LOO cross-validation error for bagging, if B is large.

训练集中大约有36.8%的数据没有被采样集采集中。对于这部分大约36.8%的没有被采样到的数据，我们常常称之为袋外数据(Out Of Bag, 简称OOB)。这些数据没有参与训练集模型的拟合，因此可以用来检测模型的泛化能力。

我们可以使用每个观察点的树来预测第1个观察点的响应，该观察点在其中是OOB的。这将产生对第1个观察点的大约 $B/3$ 的预测，我们对其进行平均。这个如果是 B 大的话，这个估计值本质上是袋法的LOO交叉验证误差。

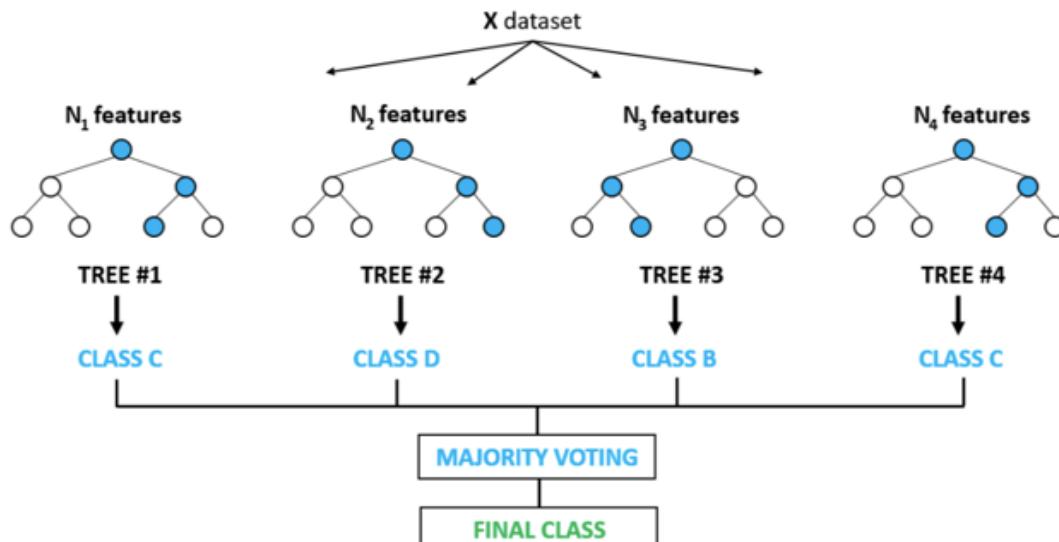
From bagging to Random Forest

- Random forests (some times) provide an improvement over bagged trees by way of a small tweak that decorrelates the trees. This reduces the variance when we average the trees.
- As in bagging, we build a number of decision trees on bootstrapped training samples.
- But when building these decision trees, each time a split in a tree is considered, a random selection of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.
- A fresh selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$ - that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors.

随机森林（有些时候）通过一个小的调整，提供了一个比Bagging更好的方法对树木进行装饰。这减少了我们对树进行平均化时的方差。如同袋法，我们在Bootstrap的训练样本上建立一些决策树。但在构建这些决策树时，每次考虑树的分裂时，都会随机选择一个从全部 p 个预测器中选择 m 个预测器作为分裂候选者。分割时，只允许使用这 m 个预测器中的一个。
每次拆分都要重新选择预测器，通常我们会选择--即每次分割时考虑的预测器数量大约等于预测器总数的平方根。

Random Forest algorithm pseudocode

```
Set number of models to build, B
for i = 1 to B
    Generate a bootstrap sample of the original data
    Train a tree model on this sample where
        for each split
            Randomly select m (< p) of the original predictors
            select the best predictor among the k predictors and partition the data
        endfor
    endfor
```



RF的主要优点有：

各个弱分类器之间没有关联，训练可以高度并行化，对于大数据时代的大样本训练速度有优势。

由于可以随机选择决策树节点划分特征，这样在样本特征维度很高的时候，仍然能高效的训练模型。

在训练后，可以给出各个特征对于输出的重要性。

由于采用了随机采样，训练出的模型的方差小，泛化能力强。

相对于Boosting系列的Adaboost和GBDT， RF实现比较简单。

对部分特征缺失不敏感。

RF的主要缺点有：

在某些噪音比较大的样本集上， RF模型容易陷入过拟合。

取值划分比较多的特征容易对RF的决策产生更大的影响，从而影响拟合的模型的效果。

Boosting

- Like bagging, boosting is a general approach that can be applied to many statistical learning methods for regression or classification. We only discuss boosting for decision trees.
- Recall that bagging involves creating multiple copies of the original training data set using the bootstrap, fitting a separate decision tree to each copy, and then combining all of the trees in order to create a single predictive model.
- Notably, each tree is built on a bootstrap data set, independent of the other trees.
- Boosting works in a similar way, except that the trees are grown *sequentially*: each tree is grown using information from previously grown trees.

为每个副本建立一个单独的决策树，然后将所有的树结合起来，以创建一个单一的预测模型。Boosting的工作方式与此类似，只是树是按顺序生长的：每棵树的生长都是使用来自以前生长的树的信息

Boosting是一种迭代算法，针对同一个训练集训练不同的分类器(弱分类器)，然后进行分类，对于分类正确的样本权值低，分类错误的样本权值高
(通常是边界附近的样本)，最后的强分类器是很多弱分类器的线性叠加(加权组合，权重和分类器精度有关)

Idea behind boosting

- Unlike fitting a single large decision tree to the data, which amounts to *fitting the data hard* and potentially overfitting, the boosting approach instead *learns slowly*.
- Given the current model, we fit a decision tree to the residuals from the model. We then add this new decision tree into the fitted function in order to update the residuals.
- Each of these trees can be rather small with just a few terminal nodes.
- By fitting small trees to the residuals, we slowly improve $f(x)$ in areas where it does not perform well. The shrinkage parameter λ slows the process down even further, allowing more and different shaped trees to attack the residuals.

与对数据进行单一拟合的大型决策树不同，对数据进行严格的拟合，并有可能过度拟合。
而提升方法则是慢慢学习。

考虑到当前的模型，我们将一棵决策树与该模型的残差拟合。然后我们将这个新的决策树添加到拟合函数中，以更新残差residuals.。

这些树中的每一个都可以相当小，只有几个终端节点terminal nodes。

通过将小树拟合到残差中，我们可以慢慢改善它表现不佳的地方。缩减参数shrinkage parameter使这一过程进一步放慢，允许更多不同形状的树攻击残差。

Boosting for regression trees

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
 - Here r_i denotes the i^{th} residual and y_i the outcome.
2. For $b = 1, 2, \dots, B$
 - Fit a tree \hat{f}_b with d splits ($d + 1$ terminal nodes) to the new training data (X, r) .
 - That is r is the new response value.
 - Update \hat{f} by adding in a shrunken version of the new tree
 - $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}_b(x)$
 - Update the residuals
 - $r_i \leftarrow r_i - \lambda \hat{f}_b(x)$
3. Compute the final model

$$\hat{f}(x) = \sum_{i=1}^B \lambda \hat{f}_b(x)$$

Parameters to tune in boosting

- Number of trees B
- The shrinkage parameter λ a small positive number.
 - Typical values are between 0.01 and 0.001.
- Number of split d in each tree.
 - Controls the complexity of the boosted ensemble.
 - If $d = 1$, then the tree is just a stump. This actually usually works quite well.

Other boosting algorithms

- AdaBoost
- Stochastic gradient boosting
- XGBoost

AdaBoost

- Short for Adaptive Boosting, one of the first boosting algorithms
- Convert a set of weak classifiers into a strong one
- Basic idea
 - at each iteration, reweight the data to place more weight on data points that the classifier got wrong
- At the end, combine all the weak classifiers by taking a weighted combination. Put more weight on the weak classifiers with the higher accuracies.

将一组弱分类器转换成一个强分类器。在每次迭代中iteration，对数据进行重新加权，将更多的权重放在分类器得到出错的数据点上最后，通过加权组合，将所有的弱分类器结合起来。把更多的权重放在准确率较高的弱分类器

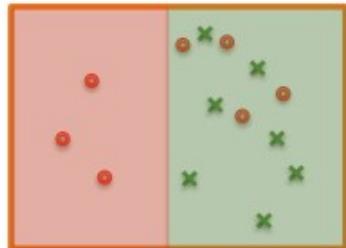
初始化样本权重，使用有权重的样本集训练弱学习器（包括分类和回归）

计算弱学习器的误差，更新弱学习器的系数

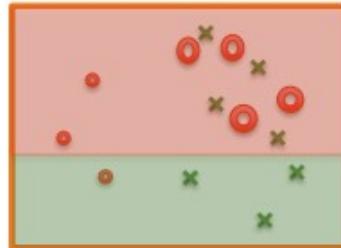
更新样本权重分布，重复2~5来构建多个弱学习器

根据结合策略，利用多个弱学习器构建最终的强学习器

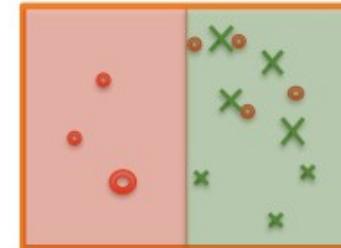
Adaboost plot



Iteration 1



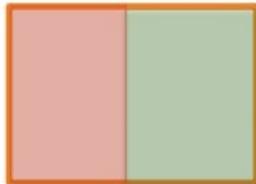
Iteration 2



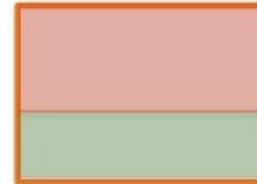
Iteration 3

Final Model

$$f(x) = 0.46$$



$$+ 0.29$$



$$+ 0.46$$



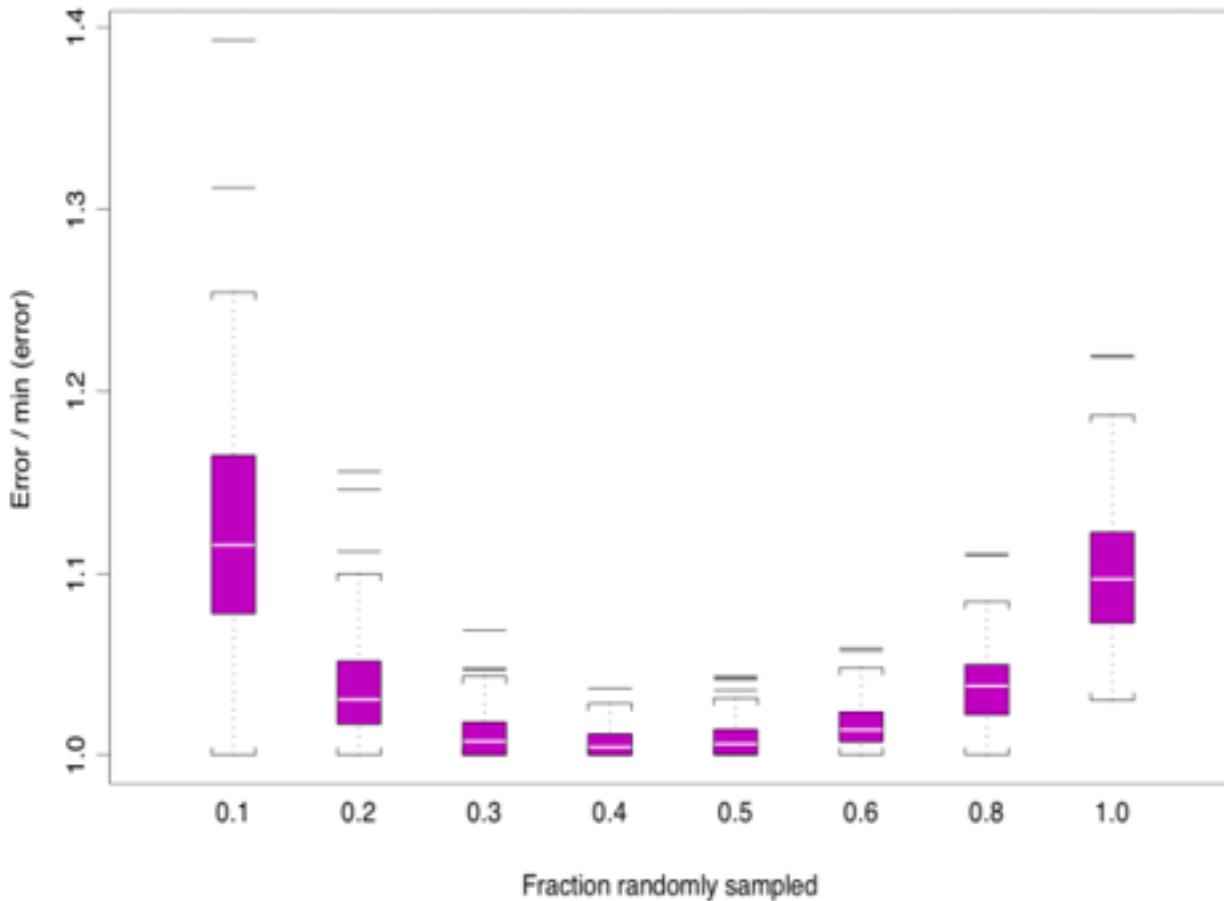
Stochastic gradient boosting

- Uses the idea behind bagging
- In each iteration of stochastic boosting, a sample of the training set instead of the full training set.
- Instead of a bootstrap sample (with replacement), the algorithms samples a fraction of the training set.
- This introduced randomness can improve performance
- Paper by Friedman (2002) covers this technique.

在随机提升的每一次迭代中，都要对训练集进行抽样，而不是全部训练集。而不是自举样本（带替换），该算法对训练集的一部分进行抽样。这种引入的随机性可以提高性能

Stochastic gradient boosting plot

$N = 500$



- Taken from Friedman (2002)

XGBoost

- XGBoost is short for eXtreme Gradient Boosting
- It is a library for high performance gradient boosting models written in C++ with implementations in Python, R and Julia
- Designed for speed – up to 10 times faster than the gbm package
- Has been very popular in recent years and has won a number of machine learning competitions (especially on tabular data)
- Supports regularization
- Can handle missing data

支持正则化
可以处理缺失数据

Bagging vs Boosting

- Both ensemble methods get N learners from 1 learner
 - built independently for bagging
 - built sequentially for boosting
- Trees built in boosting are weak learners (sometimes just a stump) while trees in Random Forest have higher complexity
- Few parameters to tune in Random Forest, many more in Boosting (depending on which variations)
- Both combine outputs from N trees

两种集合方法都是从1个独立建立的学习者中获得N个学习者

Boosting中建立的树是弱的学习者（有时只是一个树桩），而随机森林中的树有更高的复杂性，在随机森林中需要调整的参数很少，在Boosting中则更多两者都结合了来自N个树的输出。

Bagging 和 boosting 区别

样本选择上：

Bagging：训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的。

Boosting：每一轮的训练集不变，只是训练集中每个样例在分类器中的权重发生变化。而权值是根据上一轮的分类结果进行调整。

样例权重：

Bagging：使用均匀取样，每个样例的权重相等

Boosting：根据错误率不断调整样例的权值，错误率越大则权重越大。

预测函数：

Bagging：所有预测函数的权重相等。

Boosting：每个弱分类器都有相应的权重，对于分类误差小的分类器会有更大的权重。

并行计算：

Bagging：各个预测函数可以并行生成

Boosting：各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。

Summary

- Decision trees are simple and interpretable models for regression and classification.
- However they are often not competitive with other methods in terms of prediction accuracy.
- Bagging, random forests and boosting are good methods for improving the prediction accuracy of trees. They work by growing many trees on the training data and then combining the predictions of the resulting ensemble of trees.
- The latter two methods – random forests and boosting – are among the state-of-the-art methods for supervised learning. However their results can be difficult to interpret

决策树是用于回归和分类的简单和可解释的模型。然而，在预测精度方面，它们往往无法与其他方法竞争。Bagging、random forest和Boosting是提高决策树预测精度的好方法。树。它们的工作原理是在训练数据上生长出许多树，然后结合所产生的树的预测结果。形成的树的集合。random forest和Boosting--是监督学习的最先进的方法之一。然而，它们的结果可能很难解释

References

Friedman, J. H. (2002). "Stochastic gradient boosting". In: *Computational statistics & data analysis* 38.4, pp. 367-378.

James, G., D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

STAT5003

Week 10 : Monte Carlo

Dr. Justin Wishart



THE UNIVERSITY OF
SYDNEY

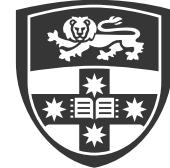


Readings



- Easily accessible text in Shonkwiler and Mendivil (2009)

Monte Carlo Methods



THE UNIVERSITY OF
SYDNEY

What are Monte Carlo methods?

- Monte Carlo are a class of computational methods that can be applied to a wide range of problems
- Key aspect of Monte Carlo methods are that they rely on **random sampling**
- Generally provide approximate solutions
 - not exact solutions
- Used in cases where analytic solutions don't exist or are too difficult to implement
- Monte Carlo methods are sometimes referred to as stochastic simulation

采样越多，近似结果是真实结果概率越大

蒙特卡洛是一类计算方法，可以应用于广泛的问题。蒙特卡洛方法的主要方面是它们依赖于随机抽样，通常提供近似的解决方案而不是精确的解决方案，用于不存在分析解或太难实现的情况下，蒙特卡洛方法有时被称为随机模拟
工作原理是通过大量随机样本，去了解一个系统，进而得到所要计算的值。

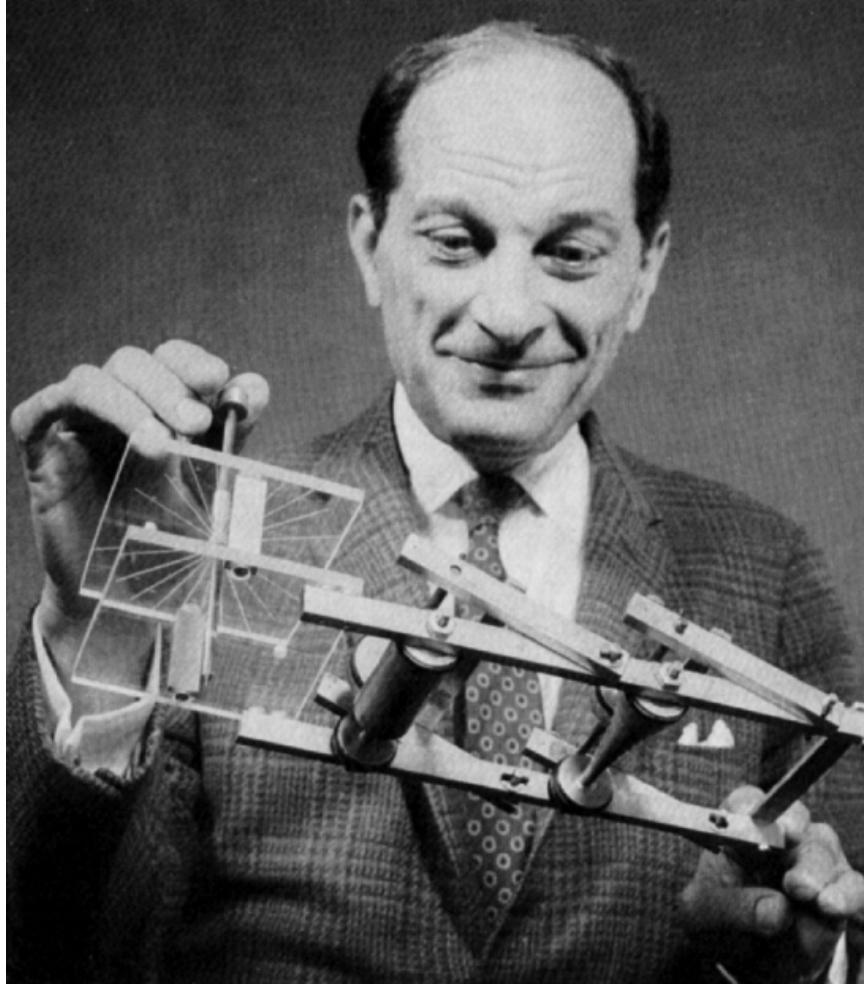
简言之就是一件事算N次，这N次的平均数就应该是这个事件的期望 $E(X)$ 。而 $E(X)$ 恰恰就是我们要求，但是又没法解出来的东西。比如一些奇异期权，我们能够列出微分方程，却解不出来奇异期权的解，就用monte carlo模拟微分方程所代表的路径，然后求平均数，就估计出这个期望值了，我们的定价目的也就达到了。Monte Carlo求的是数值解，BS公式之类得到的是解析解。

History of the Monte Carlo method

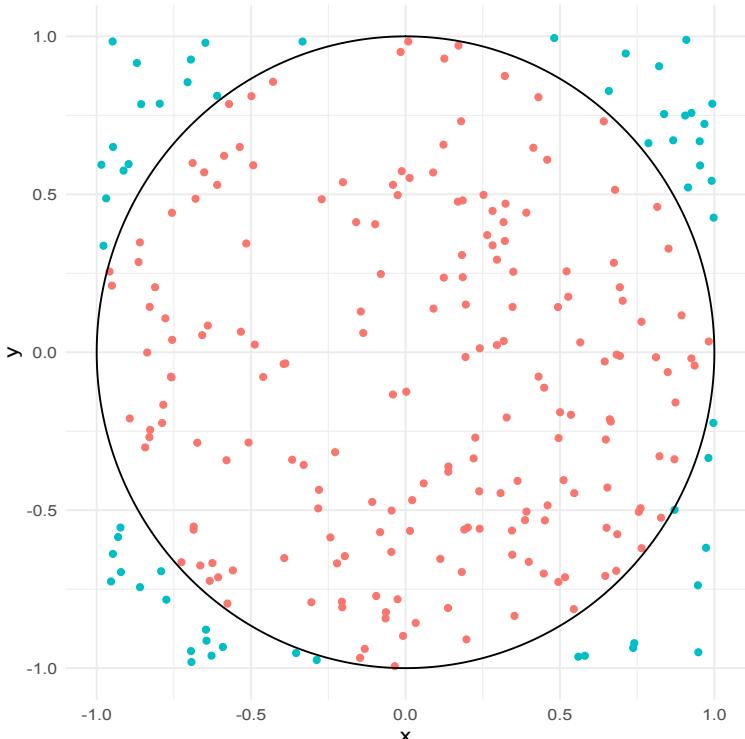
- Invented by a Polish mathematician called Stanislaw Ulam in the late 1940s
- Inspiration during recovery from illness with a thought experiment:

The question was what are the chances that a Canfield Solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays.

- Monte Carlo was the code name given for the project.
 - Inspired by the Monte Carlo casino in Monaco where Ulam's uncle would borrow family money to gamble.



Monte Carlo estimation of π



- Area of circle = πr^2 and area of square = $4r^2$
 - (r is the radius)
 - A Monte Carlo method to estimate π :
 1. Randomly draw N points within unit square at random
 2. Count the R points which are inside the unit circle
 3. Compute the ratio R/N and estimate π as $4R/N$
1. 在单元格内随机抽出N个点。
2. 计算单位圆内的R个点
3. 计算R/N的比率并估计为4R/N

<https://www.zhihu.com/question/263316961>

Using Monte Carlo to estimate π

In this example, the mathematical statistics of the procedure is:

- Write the parameter we want to estimate (π here) as an expectation
- Represent the parameter as a sample approximation

$$EX = \int t f(t) dt \approx \frac{1}{N} \sum_{i=1}^N X_i$$

- The law of large numbers **guarantees** that as the number of samples we draw increase, the parameter converges to the true parameter value.

Expectation of a random variable

From last slide, we can write the expected value of a random variable X as a sum:

$$EX = \int_A t \cdot f(t) dt \approx \frac{1}{N} \sum_{i=1}^N X_i$$

We can replace X with a function $g(X)$. Then the previous equation becomes:

$$Eg(X) = \int_A g(t) \cdot f(t) dt \approx \frac{1}{N} \sum_{i=1}^N g(X_i)$$

Where we assume to draw $x_i \sim f$

模拟赌局和股票

Monte Carlo Integration example

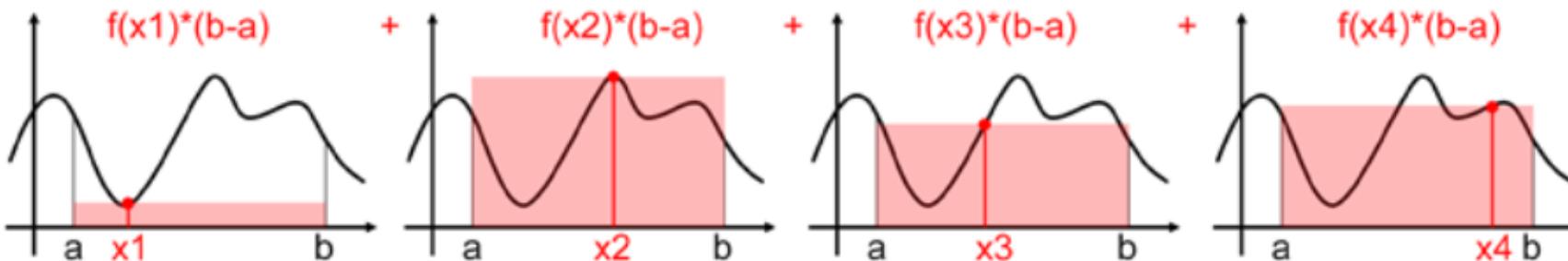
Goal is to evaluate the following definite integral

$$\int_0^1 e^{-x^2/2} dx \approx \frac{1}{N} \sum_{i=1}^N g(X_i).$$

Find a random variable such that the above holds. Need to be careful such that,

- Domain of the random variable and g agree with the above definition.
- One example is,
 - $g(x) = e^{-x^2/2}$
 - X_i is a uniform random variable on $(0,1)$
 - Then $Eg(X) = \int_0^1 e^{-x^2/2} dx \approx \frac{1}{N} \sum_{i=1}^n g(X_i)$

Monte Carlo visually



$$1/4 * (\text{red bar}_1 + \text{red bar}_2 + \text{red bar}_3 + \text{red bar}_4) \approx \text{black wavy line}$$

© www.scratchapixel.com

Simulating random variable

- Let's say we know how to simulate random variables that has a uniform distribution on some interval $U \sim [0, 1]$
 - Can do that in R with `runif`

How do we simulate random variables that can have any probability distribution?

In R, we can draw random Gaussian variables using the function `rnorm` but how do these functions really work?

Two methods (others exist):

1. Inverse-transform method
2. Acceptance-rejection method

Inverse transform method

- Let's say X is a random variable that has a cumulative distribution function (cdf) F .
- Remember that cdf is the integral of the pdf i.e.

$$F(x) = \int_{-\infty}^x f(t) dt$$

- If the inverse of $F(x)$ exists, then we can generate X as:

$$X = F^{-1}(U)$$

简言之就是一件事算N次，这N次的平均数就应该是这个事件的期望 $E(X)$ 。而 $E(X)$ 恰恰就是我们要求，但是又没法解出来的东西。比如一些奇异期权，我们能够列出微分方程，却解不出来奇异期权的解，就用monte carlo模拟微分方程所代表的路径，然后求平均数，就估计出这个期望值了，我们的定价目的也就达到了。Monte Carlo求的是数值解，BS公式之类得到的是解析解。

Example - Exponential distribution

The exponential distribution has a pdf of the form:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

and a cdf of the form:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

The inverse cdf is:

$$F^{-1}(x) = -\frac{\log(1-x)}{\lambda}$$

To sample from the exponential distribution, we can do the following:

1. Sample $U \sim \text{Uniform}(0, 1)$

2. Set $X = -\frac{\log(1-U)}{\lambda}$

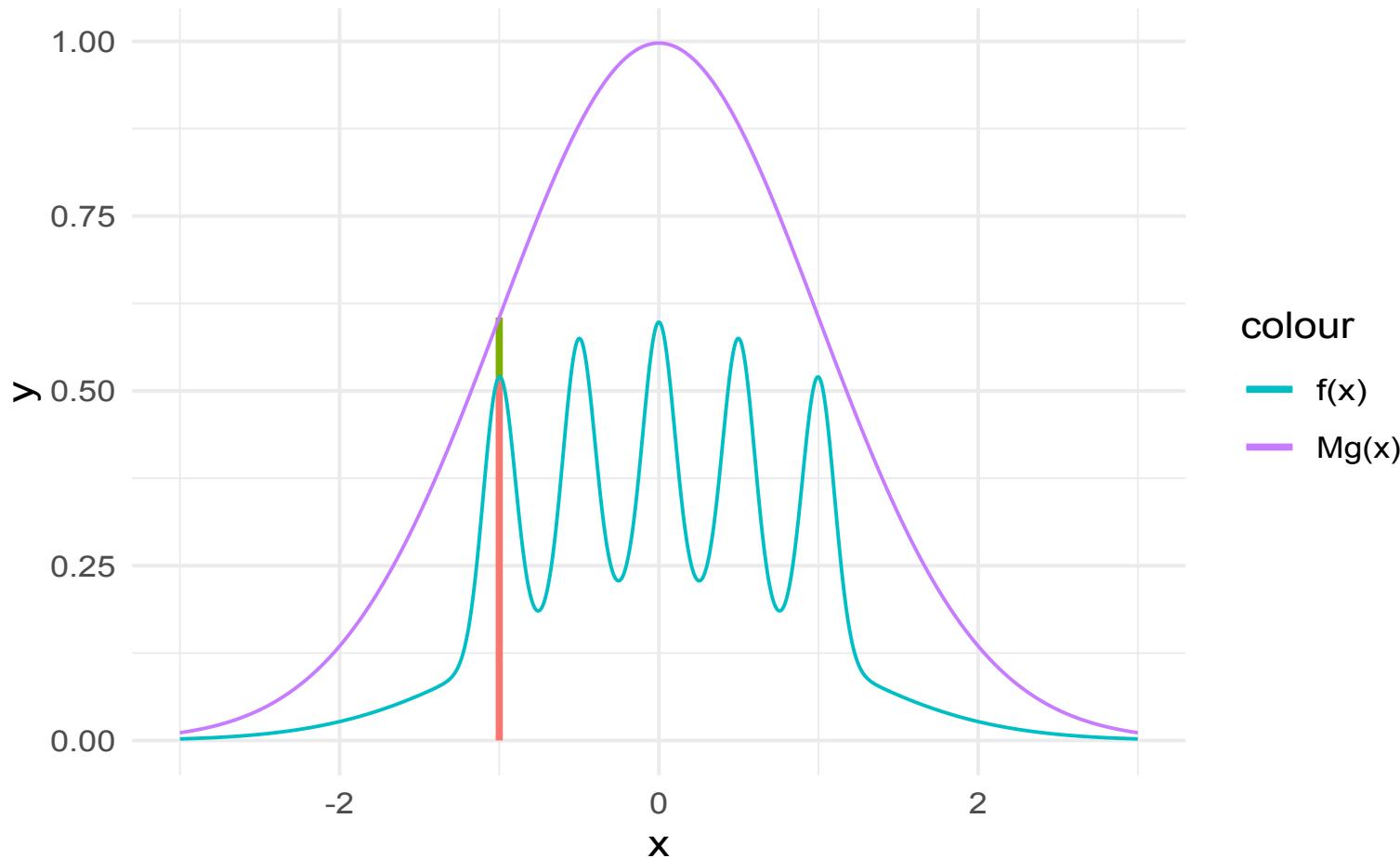
Acceptance-Rejection method

- The problem with the inverse transform method is that you need to know the functional form of the inverse of the cdf
- Acceptance-rejection method is a more general method can simulate 'difficult' distributions

Acceptance rejection method

- Given two probability densities $f(x)$ and $g(x)$, with $f(x) < Mg(x)$ for all x
- If $g(x)$ is a distribution that can be easily sampled, then we can sample $f(x)$ using the following procedure:
 1. Draw a random variable $X \sim g(x)$
 2. Accept X with probability $\frac{f(x)}{Mg(x)}$
 3. Repeat steps 1 and 2 until you have the desired number of random samples

Acceptance-rejection plot



Problems with the Acceptance-rejection method

- The envelop distribution $g(x)$ needs to closely resemble $f(x)$ for this method to work well
- If $g(x)$ does not match $f(x)$ well,
 - the method will draw a lot of unwanted samples hence is not very efficient
 - becomes very problematic when the data is high-dimensional.
 - Even at dimensions of ~ 10 , this method become very inefficient – i.e. you need to draw lots of samples before one sample is accepted

已知分布的概率密度函数，求服从此分布的样本

会抽取大量不需要的样本，因此不是很有效。
当数据是高维的时候，就会出现很大的问题。

<https://zhuanlan.zhihu.com/p/108258020>

Monte Carlo simulation examples



THE UNIVERSITY OF
SYDNEY

Collecting Coles or Woolworths kids toys

- Collect one little shop when you spend \$30
- How many \$30 shops do I need to collect all 30 items?



模拟股票

数学表示就是 $S_t = S_{t-1} + e$, e 就代表每天股价的波动，它是个随机数，所谓随机数就是取值不确定的数。

通过输入随机数符合正态分布，得到100天后的结果，模拟多条路径来进行计算，然后除以数量得到平均数

Theoretical aspect (before we see the Monte Carlo)

Expected number of shops you need to do:

$$ET = n \left(\frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{n} \right) \approx n \log n + 0.5772n + 0.5$$

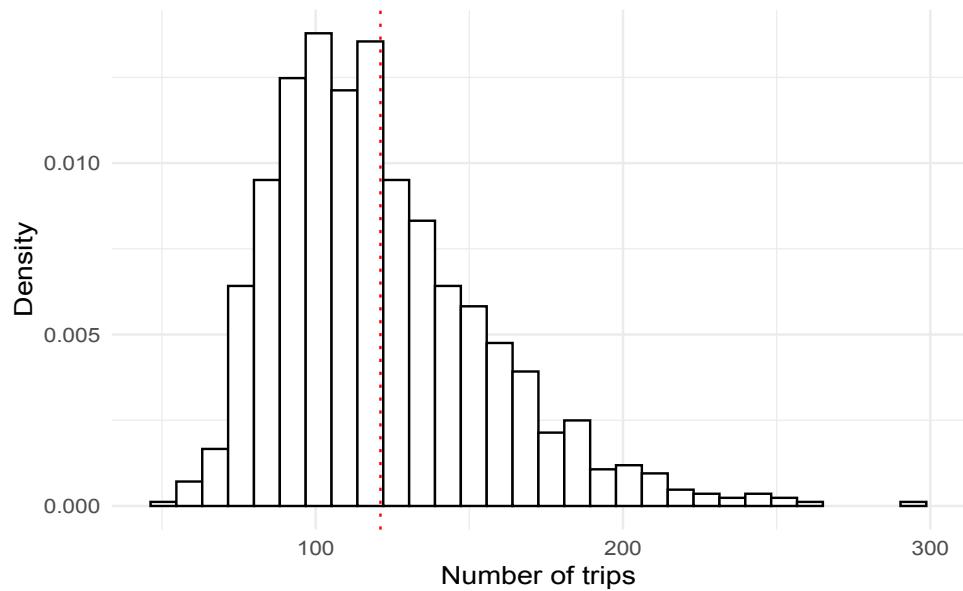
- where, T is the time to collect all items, n is the number of items to collect

So, for a Coles little shop would expect, $E(T) \approx 120$

Monte Carlo simulation of Cole's little shop

```
n.shops.sim <- c()
for(i in 1:1000) {
  collected <- c()
  n.shops <- 0
  while(length(collected) < 30) {
    newitem <- sample(30,1)
    collected <- union(collected, newitem)
    n.shops <- n.shops + 1
  }
  n.shops.sim <- c(n.shops.sim, n.shops)
}
mean.sim <- mean(n.shops.sim)
ggplot(data.frame(x = n.shops.sim)) +
  theme_minimal() +
  geom_histogram(aes(x = x, y = ..density..),
                 colour = "black",
                 fill = "white") +
  labs(x = "Number of trips", y = "Density") +
```

```
geom_vline(xintercept = mean.sim,
            linetype = "dotted",
            colour = "red")
```



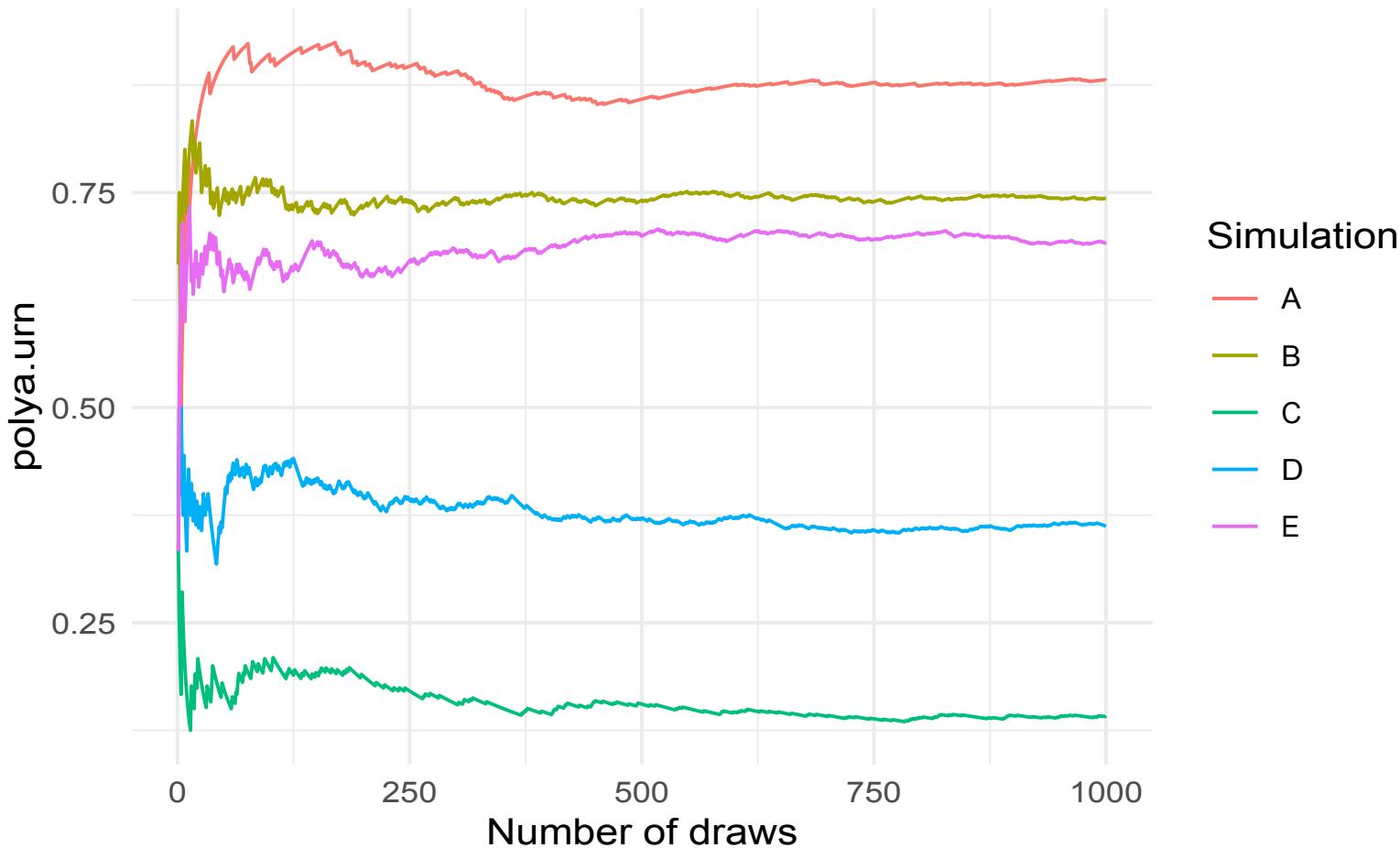
Polya urn problem

- Let an urn start with 1 black ball and 1 white ball.
- Repeatedly draw a single ball from the urn
- Each time a ball is drawn
 - you put back that ball plus one extra ball of the same colour back into the urn.

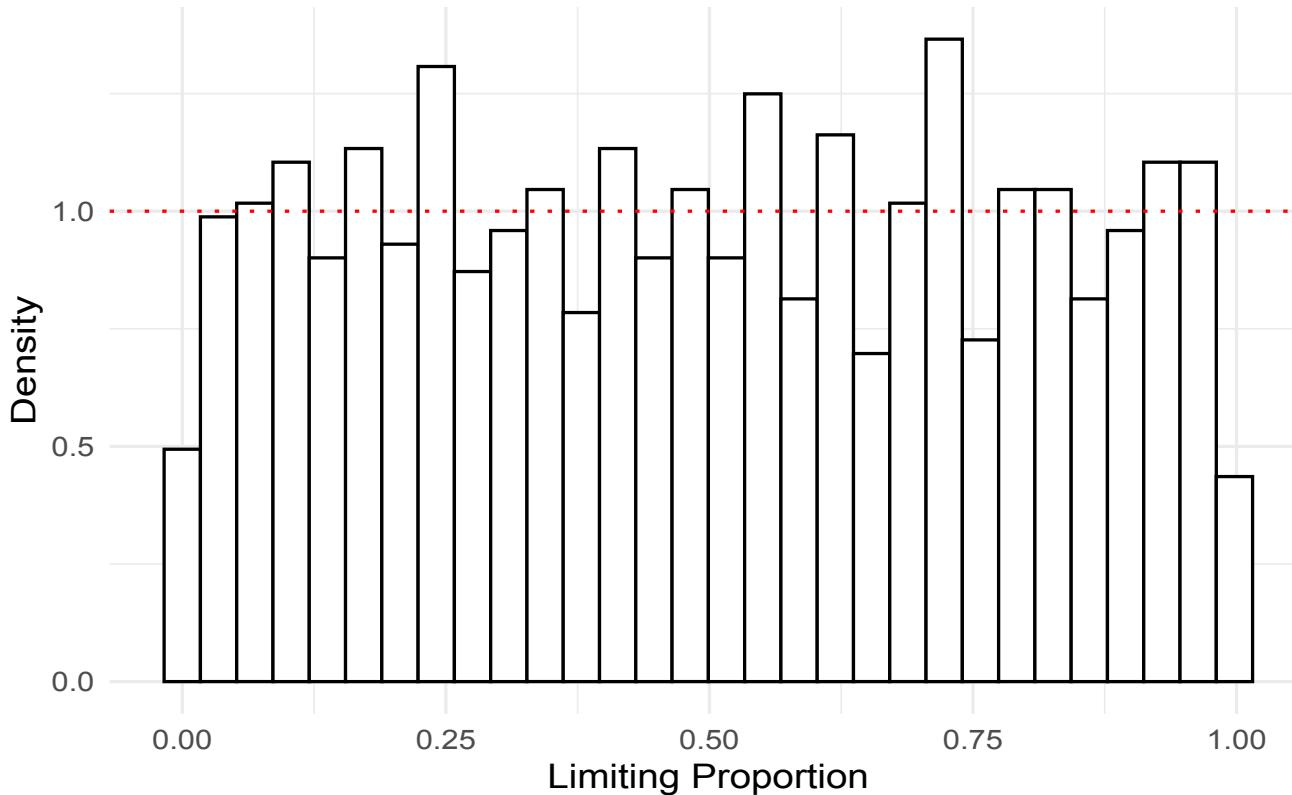
After 1000 draws, how many balls of each colour do you expect to be in the urn?

- This is an example of using Monte Carlo to simulate a process

Monte Carlo simulation of the Polya urn problem



Asymptotic Uniform distribution of the proportion of black balls



- In the limit the distribution is $U(0, 1)$

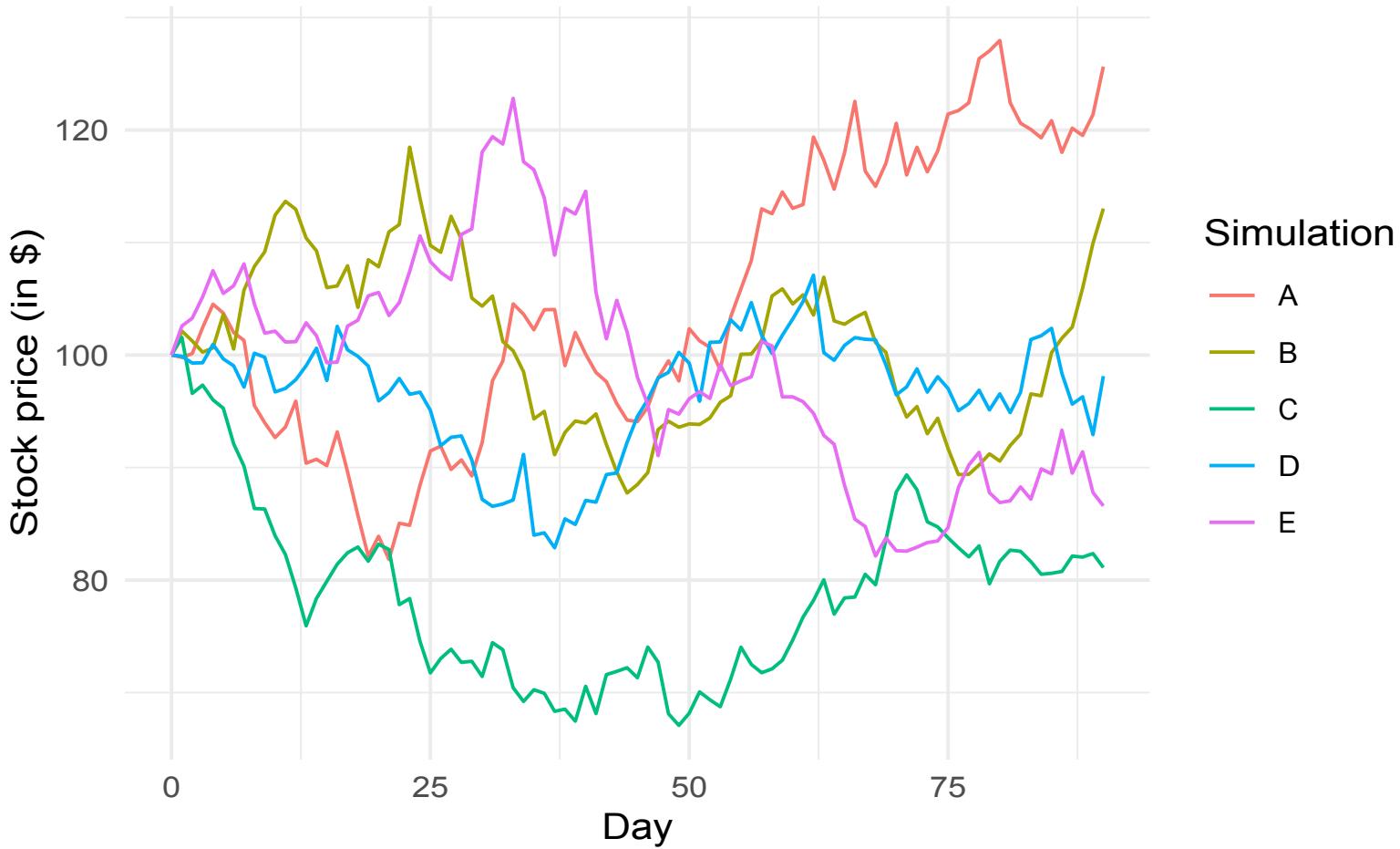
Simulate stock price as a random (stochastic) model

Equation for the future price S of a stock:

$$dS = S(\mu dT + \sigma \sqrt{dT} N)$$

- μ is the growth rate of the stock price
- σ is known as the volatility.
 - Think of this as the variance of the stock price
- N is a normal random variable
- dT is one unit of time

Examples of a stock price



Options

A **call** option gives you the right to buy a stock at a particular price at expiry

E.g. 90 day call option with \$105 strike. Let's say your stock is currently priced at \$100, if it hits \$108 at day 90, then you can exercise the option and make a \$3 profit. If its price is below \$105 at day 90, then you make nothing.

A **put** option gives you the right to sell a stock at a particular price at expiry

Option pricing with Monte Carlo

- How much would you pay for a call option? Think of this as the amount of insurance premium.
- Using Monte Carlo simulation, simulate many realisations of the stock price. For each realisation, calculate the option payoff
- The price you should pay for the option is the expected value of the profit you should make

References

Shonkwiler, R. W. and F. Mendivil (2009). *Explorations in Monte Carlo Methods*. Springer Science & Business Media.

STAT5003

Week 11 : Markov Chain Monte Carlo

Dr. Justin Wishart



Markov Chain Monte Carlo

https://blog.csdn.net/huang1024rui/article/details/113949629?utm_medium=distribute.pc_relevant.none-task-blog-baidujs_title-0&spm=1001.2101.3001.4242

https://blog.csdn.net/qq_42415326/article/details/104056711



THE UNIVERSITY OF
SYDNEY

Markov Chain Monte Carlo

- Markov Chain Monte Carlo (MCMC) is a Monte Carlo sampling technique for generating samples from an arbitrary distribution
- The difference between MCMC and Monte Carlo simulation from last week is that it uses a Markov Chain
- Two popular implementations of MCMC are
 - Metropolis-Hastings algorithm (core by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and generalization by Hastings (1970))
 - Gibbs samplers.

MCMC方法是用来在概率空间，通过随机采样估算兴趣参数的后验分布

是一种用于从任意分布中生成样本的蒙特卡洛抽样技术。区别在于它使用了马尔科夫链。MCMC的两种流行的实现方式是Metropolis-Hastings算法和Gibbs采样器

Markov Chains



THE UNIVERSITY OF
SYDNEY

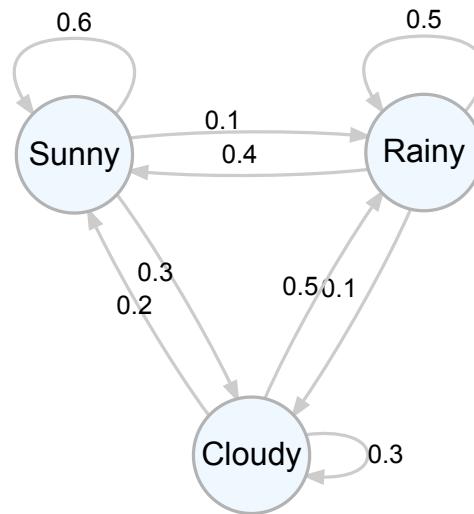
What are Markov Chains?

- Markov chain is a stochastic process that follows the Markov property
- Markov property means that the future state of the process only depends on the current state
 - Consider a **dependent sequence** where each point only depends on the immediate past.
 - Sequence $\{X_1, X_2, \dots, X_n\}$
 - Probabilities $P(X_n | X_{n-1}, X_{n-2}, \dots, X_1) = P(X_n | X_{n-1})$
- Almost Memory-less system

Markov Chain是一个遵循Markov Chain特性的随机过程。
Markov Chain属性意味着过程的未来状态只取决于当前状态。考虑一个依赖性的序列，其中每一个点只取决于刚刚过去的时间。

Markov state diagrams

- Represent states as the vertices of the graph
- Edges represent the probability of moving from one state to another state
 - e.g. if it is sunny today, 10% chance of being rainy tomorrow
- Can use this state diagram to construct a sequence of states

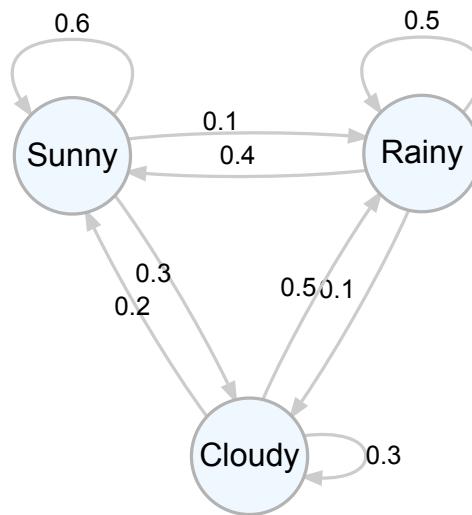


Transition Probability Matrix

$$P = \begin{pmatrix} 0.6 & 0.1 & 0.3 \\ 0.4 & 0.5 & 0.1 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}$$

- Rows represent current state
- Columns represent next state

P_{ij} = probability of transitioning from state i to state j



Transition Probability Matrix

- Start with a sunny day on day 0

$$p_0 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$

$$p_1 = p_0 P = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.6 & 0.1 & 0.3 \\ 0.4 & 0.5 & 0.1 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}$$

$$p_1 = \begin{pmatrix} 0.6 & 0.1 & 0.3 \end{pmatrix}$$

$$0.6*0.4 + 0.1*0.4 + 0.3*0.2 = 0.46$$

$$0.6*0.1 + 0.1*0.5 + 0.3*0.5 = 0.26$$

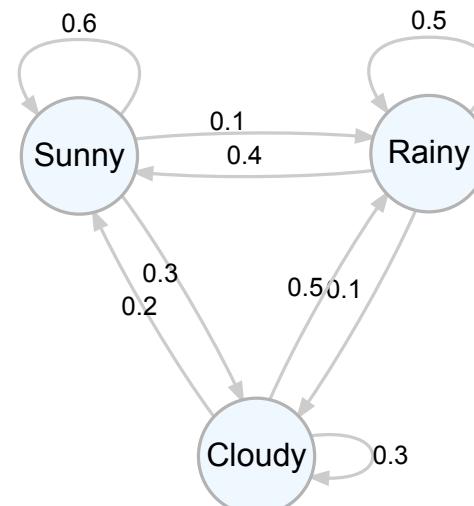
$$0.6*0.3 + 0.1*0.1 + 0.3*0.3 = 0.28$$

$$p_2 = \begin{pmatrix} 0.46 & 0.26 & 0.28 \end{pmatrix}$$

$$p_3 = \begin{pmatrix} 0.436 & 0.316 & 0.248 \end{pmatrix}$$

$$p_4 = \begin{pmatrix} 0.4376 & 0.3256 & 0.2368 \end{pmatrix}$$

- Eventually converges to an invariant distribution



Invariant distribution

- For regular Markov chains, the probability vector p_t converges to the invariant distribution π in the limit
- Can also be represented as:

$$\pi = \pi P$$

- This is satisfied if the Markov chain is :
 1. **Irreducible** - i.e. there is a path from every vertex to every other vertex
 2. **Aperiodic** – i.e. there are no loops in the Markov chain. If this is not satisfied, then the system will oscillate
 1. 不可还原. Irreducible -- 即从每个顶点vertex到另一个顶点都有一条路径
 2. 非周期性Aperiodic -- 即马尔科夫链中没有循环loops。如果这一点没有得到满足，那么系统将会震荡oscillate

MCMC - Metropolis-Hastings algorithm

[https://mp.weixin.qq.com/s?
__biz=MzU3MjA2NTQzMw==&mid=2247484003&idx=1&sn=86d3562a93bd8fc488b26af34a2d7d7bb&chksm=fcd7d195cba058835dd17830f99d7655008a4d91474b0c0535c74c033e660be677d05f3bda78&token=1706835683&lang=zh_CN#rd](https://mp.weixin.qq.com/s?__biz=MzU3MjA2NTQzMw==&mid=2247484003&idx=1&sn=86d3562a93bd8fc488b26af34a2d7d7bb&chksm=fcd7d195cba058835dd17830f99d7655008a4d91474b0c0535c74c033e660be677d05f3bda78&token=1706835683&lang=zh_CN#rd)



THE UNIVERSITY OF
SYDNEY

Metropolis-Hastings algorithm - Intuition

- Travelling politician problem
- Imagine you are a politician trying to visit all the town halls in your electorate and you want to spend time proportional to the number of voters in each town hall
- You start at a random town hall
- Choose the next town hall to visit
 - If the new town hall has more voters than your current town hall, then go there
 - If not, then go there with a probability that is equal to Number of people in new town hall / Number of people in current town hall

Metropolis-Hastings algorithm

- Similar to the acceptance-rejection method
 - it simulates a trial state
 - accepts or rejects it according to some random mechanism
- Uses the Markov chain because each trial state depends on the previous state – almost memoryless system.
- Aim is to construct a Markov chain $X_t, t = 0, 1, \dots$ such that the limiting distribution is $f(x)$

Metropolis-Hastings algorithm

Initialise state to X_0 . Require as input a target pdf $f(x)$ and a proposal pdf $q(x, y)$

For $t = 0, 1, \dots, N - 1$ do:

- Draw $Y \sim q(x|X_t)$
- Calculate acceptance probability $\alpha(X_t, Y)$
- Define $\alpha(x, y) = \min \left\{ \frac{f(y)q(x|y)}{f(x)q(y|x)}, 1 \right\}$
- Draw $U \sim U(0, 1)$
- if $U \leq \alpha$ then $X_{t+1} \leftarrow Y$ else $X_{t+1} \leftarrow X_t$

Return X_1, X_2, \dots, X_N

Proposal function

- If the proposal density function is symmetric,
 - $q(y|x) = q(x|y)$
 - the acceptance probability has a simpler form.
 - the MCMC algorithm is also known as a **random walk sampler**.
- One common choice of a symmetric proposal function is just the Gaussian function i.e.
 $q(x) \sim N(x_t, \sigma)$
- The choice of σ affects how quickly the state space is explored.

Where would you use MCMC

- One common application of MCMC is to draw from the **posterior distribution** in Bayesian statistical methods.

$$P(A | B) = \frac{P(B | A)}{P(B)} P(A)$$

- **Posterior**: The likelihood of A occurring given B has occurred.
- **Likelihood ratio**: The support B provides for A
- **Prior**: The probability of A before any data is gathered.

The posterior distribution

- Can use the Bayes rule for modelling and data.

$$P(\phi | D) = \frac{P(D | \phi)}{P(D)} P(\phi)$$

- **Posterior**: The likelihood of ϕ occurring given the data D .
- **Likelihood ratio**: The support D provides for ϕ
- **Prior**: The probability of ϕ before any data is gathered.
- Typically $P(D)$ is a difficult integral to evaluate.

$$P(D) = \int P(D | \phi) P(\phi) d\phi$$

Example involving Regression

$$P(\phi|D) = \frac{P(D|\phi)}{P(D)} P(\phi)$$

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon; \quad \varepsilon \sim N(0, \sigma^2)$$

- $\phi = \{\beta_0, \beta_1, \dots, \beta_p, \sigma^2\}$
- $D = \{Y_1, Y_2, \dots, Y_n, X_{11}, X_{12}, \dots, X_{np}\}$
- $P(D|\phi)$ will be Gaussian.
- $P(D)$ is not so easy to compute.
- $P(\phi)$ what do I choose for the prior.

Estimating posterior with MCMC

- In the Metropolis-Hastings algorithm, we only need to calculate

$$\alpha = \frac{P(\phi' | D)}{P(\phi | D)} = \frac{P(D | \phi') P(\phi')}{P(D | \phi) P(\phi)}$$

- Since $P(D)$ doesn't depend on ϕ , it cancels out on the right hand side of the above formula and hence it isn't included in the formula.

Example

Observe a series of coin flips

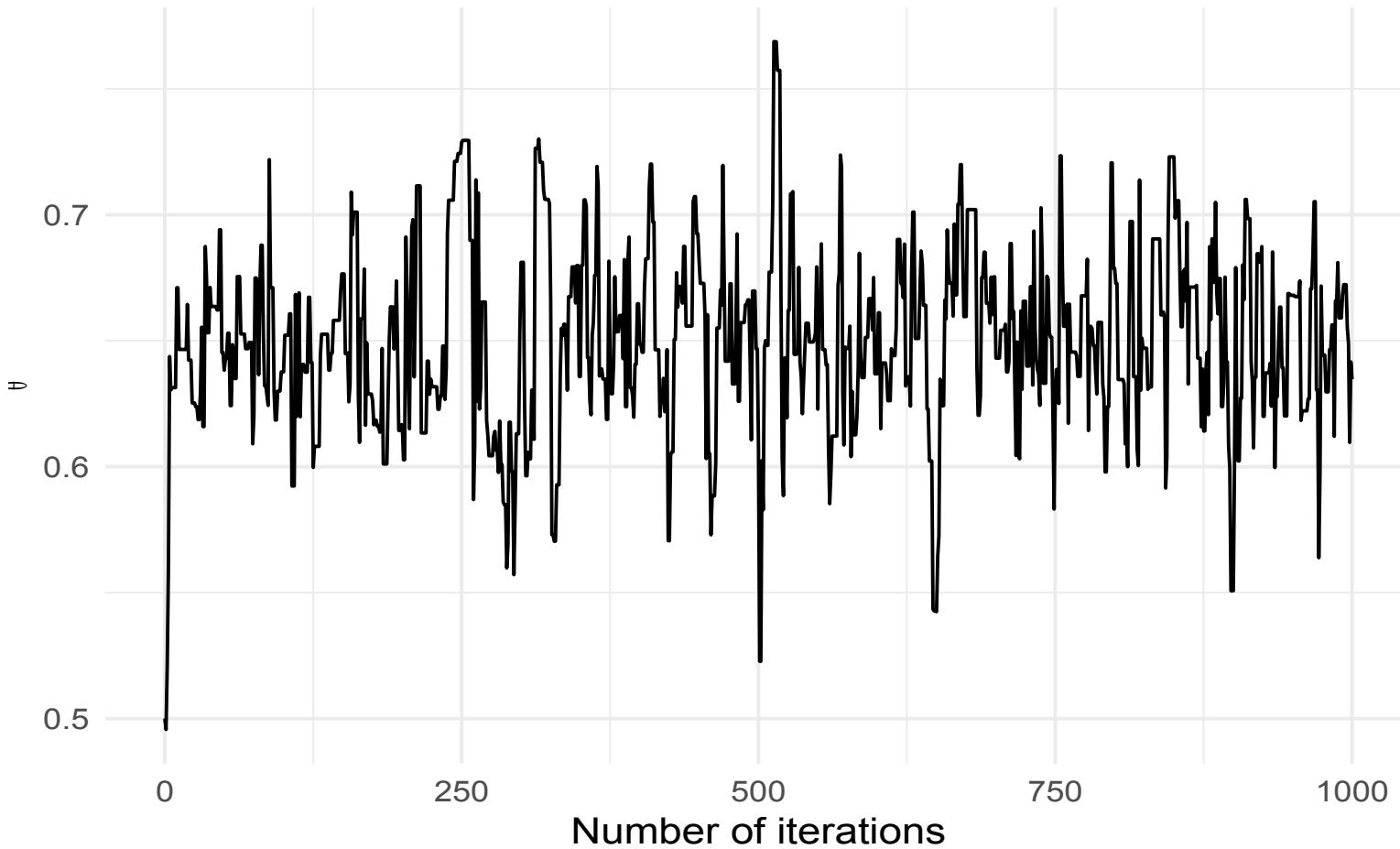
H, T, H, H, T, H, T, H, H, T, H, H, H, T, H, H, H, T, H, H, ...

Can you estimate the $P(\text{Head})$ of this coin?

Assume you don't know anything about this coin and it could be biased!

- $\theta = P(\text{Head})$
- Frequentist perspective: θ is fixed but unknown
- Bayesian perspective: θ is a random variable, can compute the likelihood of θ given the observed data.
- θ_0 set as an initial point.
- Compute θ_t as a Markov chain
- Use MCMC algorithm to estimate the posterior distribution of θ .

Estimate of $P(\text{Head})$ using MCMC



MCMC Practical considerations

- The samples at the start of the MCMC chain, before the algorithm converges to the true distribution are known as the **burn-in** period.
 - It should be discarded
- The samples generated by MCMC are correlated since they are from a Markov chain.
 - Previously, many practitioners advocated **thinning** the samples by taking say every k^{th} sample.
 - This was done for a few reasons historically
 - Reduce correlations and compute standard errors more easily
 - Less space needed to store the chain.

References

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97-109. ISSN: 0006-3444. DOI: [10.1093/biomet/57.1.97](https://doi.org/10.1093/biomet/57.1.97). eprint: <https://academic.oup.com/biomet/article-pdf/57/1/97/23940249/57-1-97.pdf>. URL: <https://doi.org/10.1093/biomet/57.1.97>.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, et al. (1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087-1092. DOI: [10.1063/1.1699114](https://doi.org/10.1063/1.1699114). eprint: <https://doi.org/10.1063/1.1699114>. URL: <https://doi.org/10.1063/1.1699114>.