Slide 2:

今天, 我将通过使用这本书 “An Introduction To Statistical Learning with Application in R” by Gareth James, Daniela Witten, Trevor Hastie, 去介绍统计学习.

什么是统计学习？

统计学习是指用于理解数据的大量工具。

这些大量工具集包括regression equation，discriminant functions和 decision tree.

Slide 3:

在统计学习中，有两类工具

le两类工具包括 supervised and unsupervised

Supervised tools 涉及建立一个统计模型来预测或估计一个output 基于一个或多个input

Unsupervised tool 是去了解一个或多个input但无supervised output的数据的关系和结构

Slide 4:

le一书介绍了三种可以去预测output的data。

le三种 data是

1. 工资数据 Wage data
2. 股市数据 Stock Market Data
3. 基因表达数据 Gene expression data

Slide 5:

Wage Data包含主题数据以及与其他标准的关系。

For example: 它包含同工资金额相关的 criteria 例如年龄，年份和教育水平等集合一起的数据

它是一个 regression Problem.

它涉及预测连续continuous或定量 quantitative的output value

(定量 quantitative是指 一di可以通过数字描述的数据)

仅有一个标准不太可能提供对主题的准确预测

(就像一个今年年份的工资数据不可能预测一个人的工资)

Slide 6:

股市数据 stock market data 包含一 di 可以用于预测非数值可能性的数据。

例如去使用过去5天的指数变化百分比来预测该指数在给定的一天中是增加还是减少

通过根据个人的特征对个人进行分组，我们可能会了解客户的类型彼此相似。 这称为clustering问题.

(Clustering可以被认为是最重要的 unsupervised learning problem

它涉及在未标记数据的集合中查找结构)

Slide 7:

基因表达 数据 Gene Expression Data 涉及一 di 观察input变量但没有相应output的情况

(目标不是预测output，而是了解关系）

For example, NCI60数据集包含64种癌细胞系, 每一种有6,830个基因表达测量值。 除了对特定的output去进行预测之外，我们可以去根据它们的基因表达测量结果确定 le一个细胞系内是否存在组或簇

Slide 8:

Statistical Learning 是一个相当新的名词，但早在很久以前就已经提出了许多构成该领域基础的概念

Le 一 di 概念 有:

1. Linear regression
2. Linear Discriminant analysis
3. Logistic regression
4. Generalized Linear model (GLM)
5. Classification and regression tree
6. Generalized additive model (GAM)

Slide 9:

19世纪，Le/gen/dren 和 Gauls发表了有关the method of least squares的论文，le 一个论文提出的the method of least squares 在将来构成最早期的 Linear Regression

Linear Regression用于预测quantitative value，例如个人的薪水

So easy in Excel

(go to excel file)

Estimate 显示方程式中使用的值。

Standard Error 显示了估计的不确定性

较低的Standard error告诉我Estimate是准确的。

Slide 10:

为了预测qualitative value，例如患者存活或死亡，Fisher提出了linear discriminant analysis。

(qualitative values 是通过单词描述的data)

Linear discriminant analysis 是一种用于查找特征的 linear combination的方法，这些描述或分离了两个或更多类对象或事件。

* Step 7: The discriminant function 是将对象分配到单独组中的分类规则

Possible questions:

What is the use of covariance matrix of group i?

It contains information about the sources that can be used for sources detection and separation.

What is prior probability distribution?

Prior probability distribution 是compound distribution的形式，这是根据一个random variable进行分布而产生分布而产生的概率分布。

Slide 11:

在1940年代，各种作者提出了另一种方法，即Logistic regression。

当dependent variable 为二进制时，Logistic regression 是进行合适regression分析。

它用于描述数据并解释一个dependent binary variable与一个或多个名义，有序，区间或比率级别的自变量之间的关系。

Slide 12:

在1970年代初期，Nel/der和Wed/der/burn为整个 Statistical Learning methods 去创造 Generalized Linear Models，其中包括Linear 和 Logistic Regression作为特例。

3 components in Generalized linear Model:

1. Linear predictor
2. Link function
3. Probability distribution

Slide 13

Linear predictor 只是 parameter（b）和 explanatory variable（x）的 linear combination.

Link function 是linear predictor和概率分布的 parameter 的"链接”

Probability distribution 产生variable y。

Slide 14:

在1980年代中期，Brei/man，Fried/man，Ol/shen和Stone引入了 Classification 和 Regression Trees，并且是最早展示一种详细practical implementation 方法的人之一。

它是用于描述用于Classification和Regression Learning task的decision tree的算法。

Two Tree:

1. Classification tree
2. Regression Tree

Slide 15:

Classification tree 是一种算法，其中目标变量是固定的或分类的

See example:

Only choose one out of two

Slide 16:

Regression Tree是指target variable位于其中的算法，该算法用于预测其值。

通常通过连续性dependent variable来预测subject

Example: Log salary is continuous not 固定的或分类

Slide 17:

Has/tie 和 Tib/shir/ran/ni在1986年为 Generalized linear model 的non-linear extension一类创造了 Generalized Additive Models.

可以使用多种方法来找到regression。并最终合并成GAM形式

Slide 18:

Possible question:

What is a transpose of a matrix?

Rows are the column of original matrix and columns are row of the original matrix.