数据科学重点

数据科学重点

1.数据清洗
2.ggplot2
3.变量转换（怎么转化成因子变量）
4.数据框的基本操作，头、尾、排序
5.flights数据集（作业要仔细看）
6.常识：R语言什么类别，存储在计算机哪个部分，内置包
7.混淆矩阵每个象限叫什么；准确率、召回率
8.数据的类别，定序，定比，定类——NOIR属性
9.统计值和描述值之间的差别
10.ggplot的代码——箱式图，柱状图，直方图
11.五大类算法
12.回归的诊断
13.summary()
14.逐步寻优，判断准则
15.cart、c4.5的判断指标
16.【计算题】会算激励系数
17.贝叶斯做分类数据比较好
18.神经网络
19.聚类距离
20.关联规则；推荐
21.时间序列

1.数据清洗

包的使用，合并、筛选、提取

两道大题：清理；

代码层；

2.ggplot2

七层代码

3.变量转换（怎么转化成因子变量）

as.factor(xxx)

转化为因子变量

4.数据框的基本操作，头、尾、排序

> 头

head()

> 尾

tail()

> 数据框的排序

1. dplyr（记得加载包）

   arrange(iris,iris[,1],-iris[,3])

   第一列升序，第三列降序

2. order（系统自带）

   iris[order(iris[,1],-iris[3]),]

   第一列升序，第三列降序

> 向量的排序

sort() 默认升序

sort(a,decreasing=True)降序

5.flights数据集（作业要仔细看）

繁忙率

误点

6.常识：R语言什么类别，存储在计算机哪个部分，内置包

> 类别

数据分析语言。免费开源，多系统兼容，高度可扩展。学习曲线陡峭。

数据处理的过程高度交互，螺旋式上升或迭代的过程。

> 存储

当R运行时，变量、数据、函数、结果等作为带名称对象的形式存储在计算机的**活动内存**（active memory）中

> 内置包

包名	是否默认加载	内容描述
base	是	基本函数，包括算术运算，I/O，编程支持等函数
datasets	是	一些数据集
graphics	是	基本绘图函数
stats	是	用于统计计算和随机树生成的函数，包括许多常用的统计检验，概率分布和建模工具。
utils	是	R的一些列使用功能，包括包管理，文件的读/写和编辑
grDevices	是	支持base和grid绘图的绘图设备，包括系统支持工具
methods	是	S4中引入的标准方法和类的实现

7.混淆矩阵每个象限叫什么；准确率、召回率

TP——真正

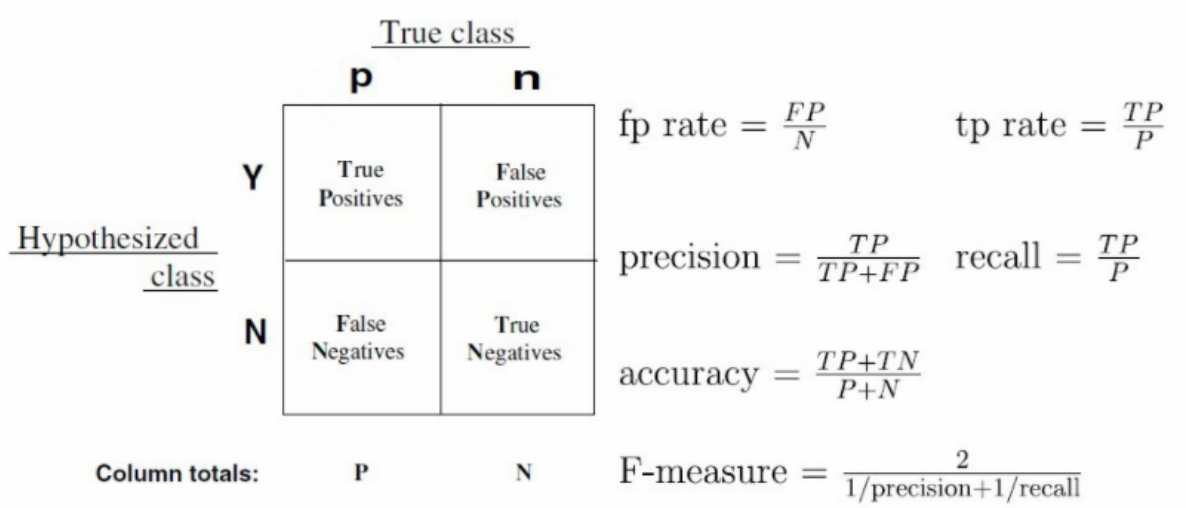
FP——假正

FN——假负

TN——真负

准确率（Precision）：TP/(TP+FP)

召回率（Recall）：TP/P=TPR



8.数据的类别，定序，定比，定类——NOIR属性

	定性数据（分类属性）		定量数据（数值属性）	
	定类数据Nominal	定序数据Ordinal	定距数据Interval	定比数据Ratio
定义	属性值仅仅区分彼此的标志，没有序次关系	属性表示个体在某个有序状态中所处的位置；	具有间距特征的变量，有单位或量纲，但一般没有绝对零点	数据具有测量价值
示例	邮政编码；国籍；姓名；性别；ID；真假逻辑值	品质；学习成绩；地震登记；满意评估	温度，日期	年龄；数量，长度，重量，面积，体积，.....
操作	逻辑比较运算：=，！	比较运算：=，！ 比较运算：>，<	比较运算： 加减运算：+，-	比较运算： 加减运算：+，- 乘除运算

> 统计数据主要可分为四类：

1. **定类数据**（Nominal）：名义级数据，数据的最低级，表示个体在属性上的特征或类别上的不同变量，仅仅是一种**标志**，**没有序次关系**。例如，**"性别"**，"男"编码为1，"女"编码为2。
2. **定序数据**（Ordinal）:数据的中间级，用数字表示个体在某个有序状态中所处的**位置**，**不能做四则运算**。例如，**"受教育程度"**，文盲半文盲=1，小学=2，初中=3，高中=4，大学=5，硕士研究生=6，博士及其以上=7。
3. **定距数据**（Interval）:具有**间距特征**的变量，有**单位**，**没有绝对零点**，可以做**加减运算**，**不能做乘除运算**。例如，**温度**。
4. **定比数据**（Ratio）:数据的最高级，既有**测量单位**，也有**绝对零点**，例如**职工人数**，**身高**。

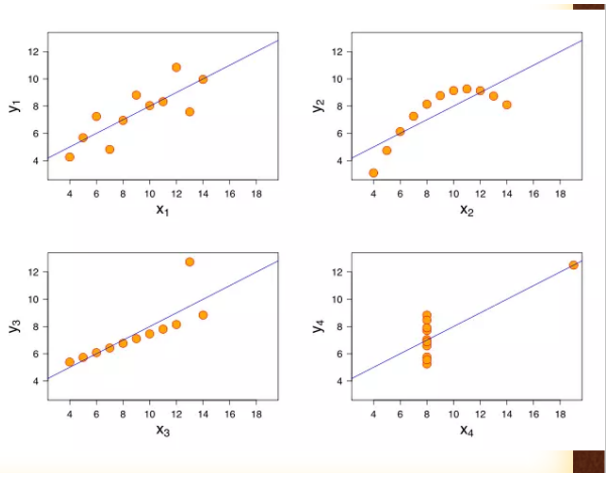
9.统计值和描述值之间的差别

统计数据和大数据的区别——**安斯库姆四重奏**

说明：**数据统计值不能完全代表真实的数据分布**

四组不同的数据做线性相关，得到相关系数一样（0.816），但实际作图可以看出很大差异

性质	数值
x的平均数	9
x的方差	11
y的平均数	7.50（精确到小数点后两位）
y的方差	4.122或4.127（精确到小数点后三位）
x与y之间的相关系数	0.816（精确到小数点后三位）
线性回归线	y=3.00+0.500x（分别精确到小数点后两位和三位）



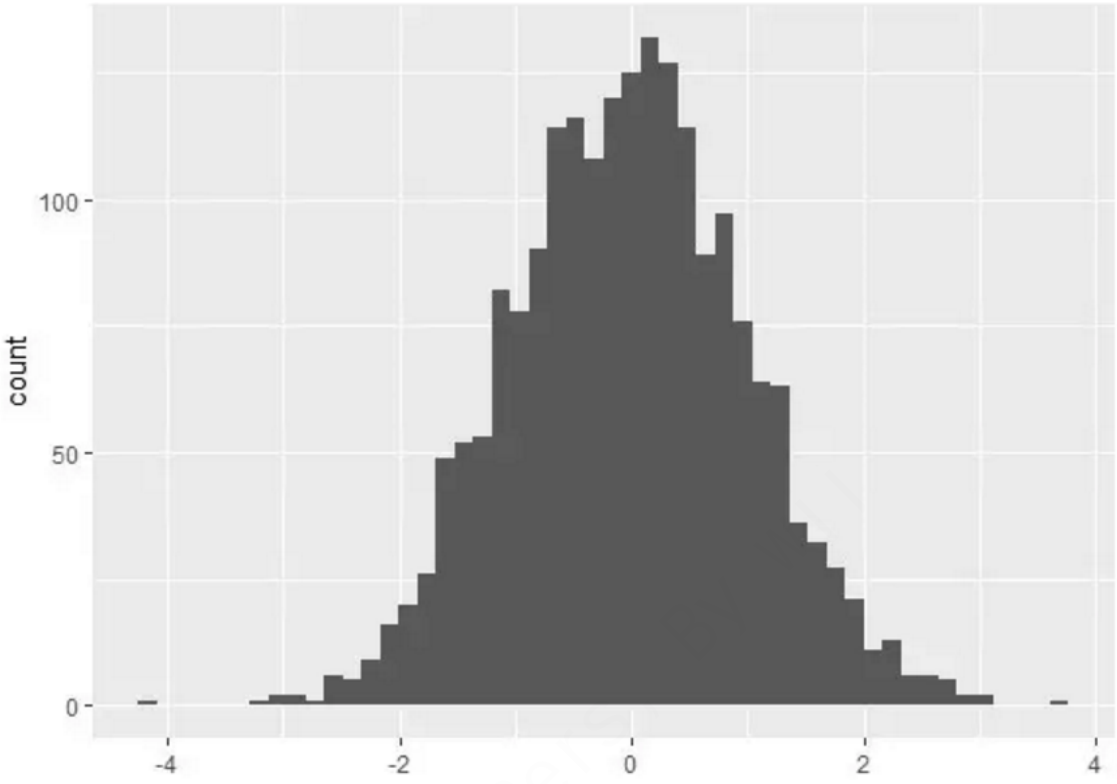
10.ggplot的代码——箱式图，柱状图，直方图

- **柱状图**

  geom_bar

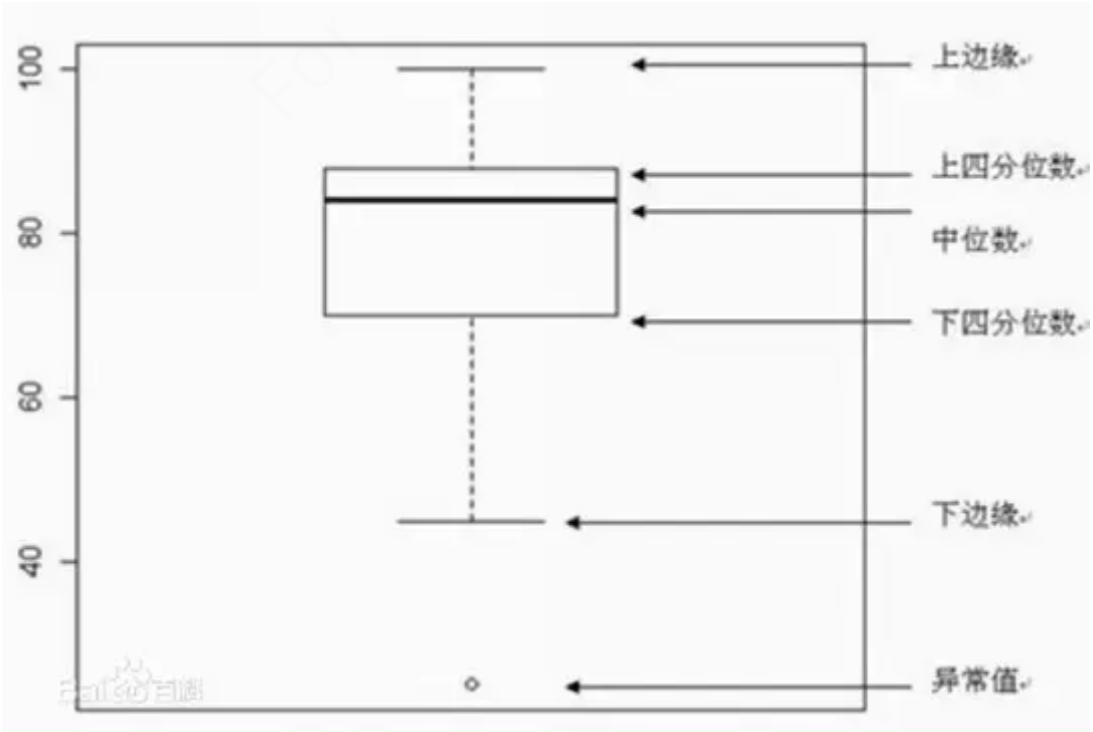
- **直方图**

  geom_histogram



- **箱式图**

  geom_boxplot



11.五大类算法

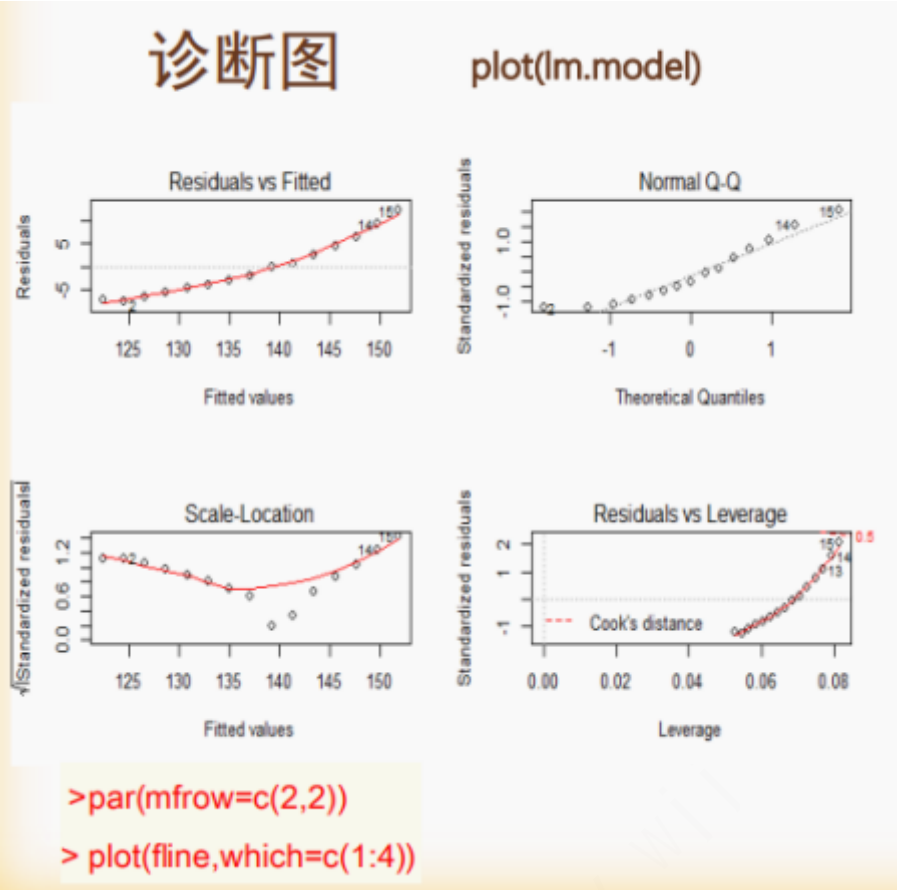
对应的包、函数、参数

注意大小写！！！！

模型	包	函数
线性回归	stats	lm()
广义线性回归模型	\	glm()
C4.5	party	ctree()
CART	\	tree()
神经网络	nnet	nnet()
朴素贝叶斯	e1071	naiveBayes()
	klaR	NaiveBayes()
支持向量机(SVM)	e1071	svm()
K均值	stats	kmeans()
K-medoids	cluster	pam()
密度聚类	fpc	dbscan()
关联规则(Apriori)	arules	apriori()
ECLAT	arules	eclat()
推荐算法	Recommenderlab	recommender()

12.回归的诊断

四象限的诊断



1. 残差图	检验方程是否有异方差。如果残差图上的点的分布存在一定趋势，如随横坐标增长而增大或减小，则可判断存在异方差。
2. QQ图	检验样本是否符合正态分布。如果是标准正态分布，QQ图上的点近似在Y=X直线上。

3.标准化残差方根散点图——类似残差图

4. Cook距离图	判断强影响点是否为Y的异常值点。如果D<0.5则不是，D>0.5则是异常点

13.summary()

```
➢ summary(lm.model) # 输出模型的统计信息
➢  Call: lm(formula = weight ~ height - 1, data = women)
➢ Residuals:
➢  Min   1Q     Median  3Q    Max
➢ -7.461 -5.235 -2.118 3.498 12.115
➢ Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
➢ height    2.10951   0.02452    86.05      <2e-16 ***
➢ ---
➢ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

➢  Residual standard error: 6.185 on 14 degrees of freedom
➢ Multiple R-squared: 0.9981  Adjusted R-squared: 0.998
➢ F-statistic: 7404 on 1 and 14 DF, p-value: < 2.2e-16
```

14.逐步寻优，判断准则

以AIC信息统计量为准则，通过选择最小的AIC信息统计量，来达到删除或增加变量的目的。

最终选择**AIC最小的方程**

• logit.step <- step(glm, direction = "both") # 逐步寻优法

• summary(logit.step)

• logit.step <- step(glm, direction = "forward") # 前向选择法

• summary(logit.step)

• logit.step <- step(glm, direction = "backward") # 后向选择法

```
> # 逐步寻优法
> logit.step <- step(glm, direction = "both")
Start:  AIC=569
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

       Df Deviance    AIC
- x2    1   551.54 567.54
- x8    1   551.79 567.79
- x5    1   552.25 568.25
<none>      551.00 569.00
- x1    1   554.95 570.95
- x6    1   555.93 571.93
- x4    1   573.03 589.03
- x7    1   594.37 610.37
- x3    1   635.21 651.21
```

图中内容为：删除变量后对应的AIC值

如，第一行指：删除x2后，方程的AIC为567.54

`<none>`指，没有删除变量时，对应方程的AIC值

因此，优先删除x2

以实现方程的AIC最小

15.cart、c4.5的判断指标

分类的判断指标

ID3——信息增益

C4.5——信息增益比

CART——Gini指数（数据分割的特征）

16.【计算题】会算基尼系数

不同的分类方案，要会算对应的激励系数

以这门课的排序而不是成绩

69算优良?

79算优良?

哪种的基尼系数更好

两个班分别生成一个基尼系数

激励系数是否一致

17.贝叶斯做分类数据比较好

1. 量化数据怎么处理——**分类化**or**高斯转化**
2. 0概率事件——**拉普拉斯修正**

   公式：**分子+1/分母+自由度**

18.神经网络

激活函数：**sigmoid函数**

主要作用：加入非线性因素，解决线性模型的表达、分类能力不足的问题

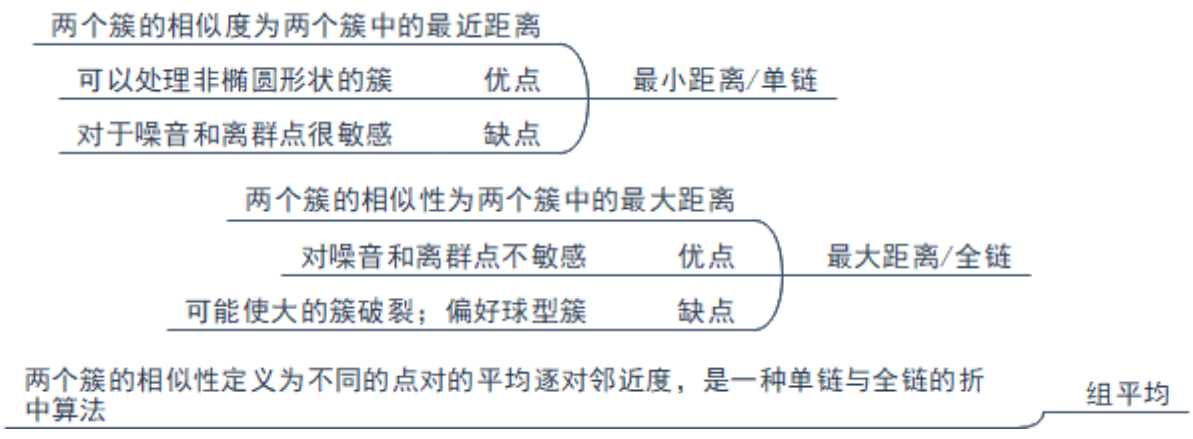
$$f(z)=\frac{1}{1+\exp(-z)}.$$

完美的**可导性**、**二分性**、**连续性**

20.kmeans（B卷出了；A卷没出）

19.聚类距离

1. 单链和全链的区分【选择】



全链；单链；平均；重心；wd

距离的度量方式——单链和全链要区分；优缺点是是什么（选择题）

【计算题】

不考层次聚类

就业类型——就业类型；企业；签约类型

定类数据。聚类的时候，gower？×。**离散值的距离计算方法——简单匹配系数**或**jaccard**

8个对象，两两距离

层次聚类；

只可能用**单链**做。

20.关联规则；推荐

置信度与支持度的计算问题，没有算提升度；

数据的格式问题——垂直/水平

优先法则

aprior——包

调用要清楚

数据格式要清楚

超过全班平均分算优秀？

二值化清单——0，1矩阵

做推荐

21.时间序列

ts()

decompose()

HoltWinters()

> 功能和作用；识别；只考了一个选择题

【拖尾和截尾的判断】

ARIMA的pdq的判断方法

选填：15+20（1）——35分

计算：50分

【计算题】

1.贫困

2.做图

3.就业类型

4.编程

5.奖学金评定——激励系数测二分类