Part 4

挖掘预处理
郎大为 J.D. Power

课程大纲

- 挖掘预处理[1小时]
    - 数据的类型
    - 数据的可能问题
    - 数据预处理的方法

数据的类型

什么是数据: Data Nomenclature

- 数据样本与属性的收集
- 一个属性就是一个样本的性质或者特征。
    - 属性也被称作变量，特征，性质等。
- 一系列特征的集合被称为样本。
    - 样本也被称作记录，样本点，例子等

```
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

数据中变量的类型

- 名义变量
  - 例子:ID，眼睛的颜色..
- 次序变量
  - 例子:排名（比如薯片味道的排名，1~10），级别，学历，身高（高，中，低）
- 间隔变量
  - 例子:日历上的时间，摄氏度，华氏度..
- 比率变量
  - 例子:开尔文温度，长度，时间....

数据中变量的类型

变量类型是取决于变量可以进行那种种类的运算

- 特异: =
- 排序: < >
- 加法/减法: + -
- 乘法/除法: * /

规则是这样的:

- 名义变量: 特异
- 排序变量: 特异 & 排序
- 间隔变量: 特异 ， 排序 & 加减
- 比率变量: 上述4种运算

名义变量

- 描述:
  - 名义变量的值只是名字上不同，或者说，名义变量只提供了足够判断两个观测是否不同。
- 例子
  - 工号，眼睛颜色，性别
- 分析工具:
  - 众数，熵，列联表相关性

次序变量

- 描述:
  - 次序变量提供了足够的信息供人们为观测值排序
- 例子
  - 一个矿产的优劣，成绩，门牌号
- 分析工具:
  - 中位数，百分位数，秩相关系数，符号秩检验

间隔变量

- 描述:
  - 对于间隔变量，变量之差的度量是有意义的:每个单位都是相同的
- 例子
  - 日历上的时间，摄氏度，华氏度
- 分析工具:
  - 均值，标准差，皮尔逊相关系数，t，F检验

比率变量

- 描述:
  - 对于比率变量，变量之间的差异与比率都是有意义的。
- 例子
  - 开尔文温度，货币数量，年龄，长度，次数
- 分析工具:
  - 几何均值，百分率，调和均值

离散变量和连续变量

- 离散变量
  - 只有有限或者可列个属性值
  - 例子:门牌号，次数，文档中的字
  - 经常可以看作整数变量
  - 二元变量是离散变量的特例（0-1）
- 连续变量
  - 有一个真实的数值来做属性值
  - 例子:温度，身高，体重
  - 实际上，连续变量只能有有限的数值或者小数
  - 连续变量可以作为浮点变量的特例

数据集的类型

- 记录
  - 数据矩阵
  - 文档
  - 交易数据
- 图数据
  - 社交网络数据
  - 分子式
- 排序
  - 空间数据
  - 时间数据
  - 基因序列数据

记录型数据

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

数据矩阵

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

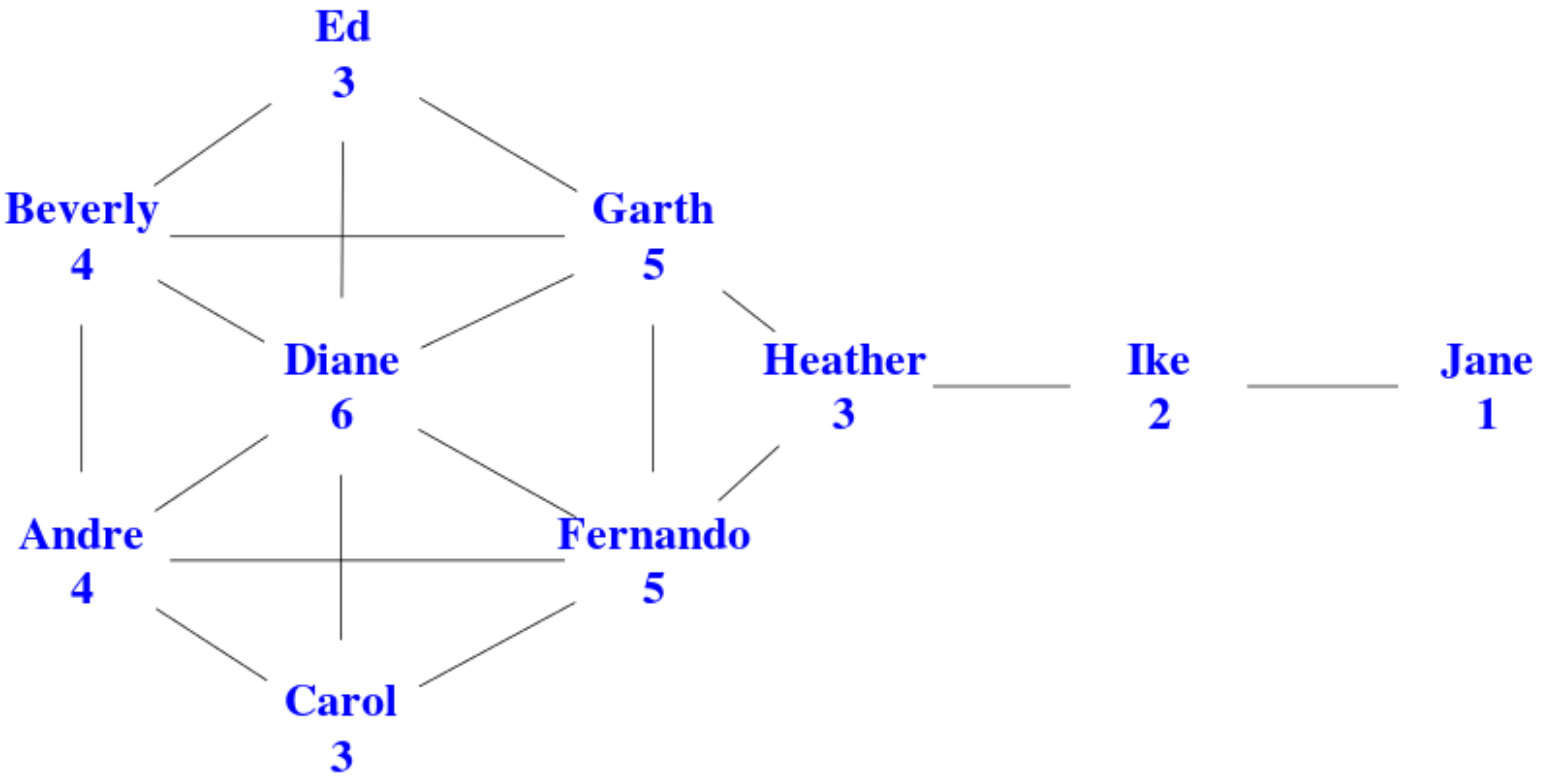
文档数据

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

交易数据

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

图数据


Ed
3
Beverly
4
Garth
5
Diane
6
Heather
3
Ike
2
Jane
1
Andre
4
Fernando
5
Carol
3

有序数据

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

数据的可能问题

为什么需要准备数据？

- 几乎所有建模方法都假定数据是一个矩阵形式并且是干净的数据。
- 从几个不同的数据源中的不同表格中提取数据是一项关键的工作。
- 最终模型的结果总依赖于数据的质量。

现实的数据都不是可以简单收集的

- 数据可能出现的问题？
- 我们怎么从数据中发现问题？
- 我们怎么解决这些问题?
- 可能会遇到的问题:
    - 噪声，离群点
    - 缺失值
    - 重复记录

基本流程

- 从业务系统或者数据仓库中提取数据
- 整合数据
- 探索数据并检验数据的完备性（离群值，缺失值）
- 转化数据并建立新的变量
    - 属性的转化，聚合，离散化，新变量的生成
- 删除不具预测能力或预测能力很低的冗余的变量

噪音

- 无法避免
- 更多的数据会使噪音减弱
- 可能通过一个良好的模型来减弱噪音的影响
- 例子
  - 信号很差的电话
  - 带雪花的电视频道

离群点

- 离群点是数据中一些观测值的特征与其他观测值相差较远的情况。
- 考虑是否要将离群点消除
- 例子
  - IQ
  - 跳水比赛打分

缺失值

- 缺失值产生的原因
  - 数据没有采集到（e.g人们倾向于拒绝提供自己的年龄和体重）
  - 属性不一定适合所有的观测值（年收入数据对于未成年人）
- 解决缺失值
  - 消除样本
  - 预测缺失值
  - 忽略缺失值来进行建模
  - 用其他数据代替（用频率的加权平均）
  - 通过联系业务人员恢复数据

数据源

- 业务系统
  - 原始数据（raw data）
  - 可能会有错误
  - 最能描述业务的数据源
- 数据仓库/数据集市
  - 理想的数据来源
  - 提供清洁、有条理的数据
- 合并数据可能会导致重复数据

数据预处理的方法

探索数据

- 检验各个数据集的统计性质
- 探索变量间的统计性质
- 比较数据集
- 确定数据完整性
    - 寻找无效值
    - 寻找离群值
    - 寻找缺失值

探索数据

保证训练集与验证集用同样的统计性质

- 确定所有的在第一个数据集的性质都在第二个数据集中体现

```
v1 <- factor(sample(letters[1:2],50,replace=TRUE))
v2 <- factor(sample(letters[1:3],50,replace=TRUE))
levels(v1)
```

```
[1] "a" "b"
```

```
levels(v2)
```

```
[1] "a" "b" "c"
```

探索数据

- 确定各个性质在两个数据集中的比例近似相等

```
v2 <- factor(sample(letters[1:2],100,replace=TRUE,prob=c(0.4,0.6)))
mat <- rbind(table(v1),table(v2))
chisq.test(mat)
```

```
    Pearson's Chi-squared test

data:  mat
X-squared = 0, df = 1, p-value = 1
```

探索数据

- 确定数据的集中程度和离散程度相似

```
v1 <- runif(100)
v2 <- rnorm(100)
mean(v1);sd(v1)
```

```
[1] 0.49891
```

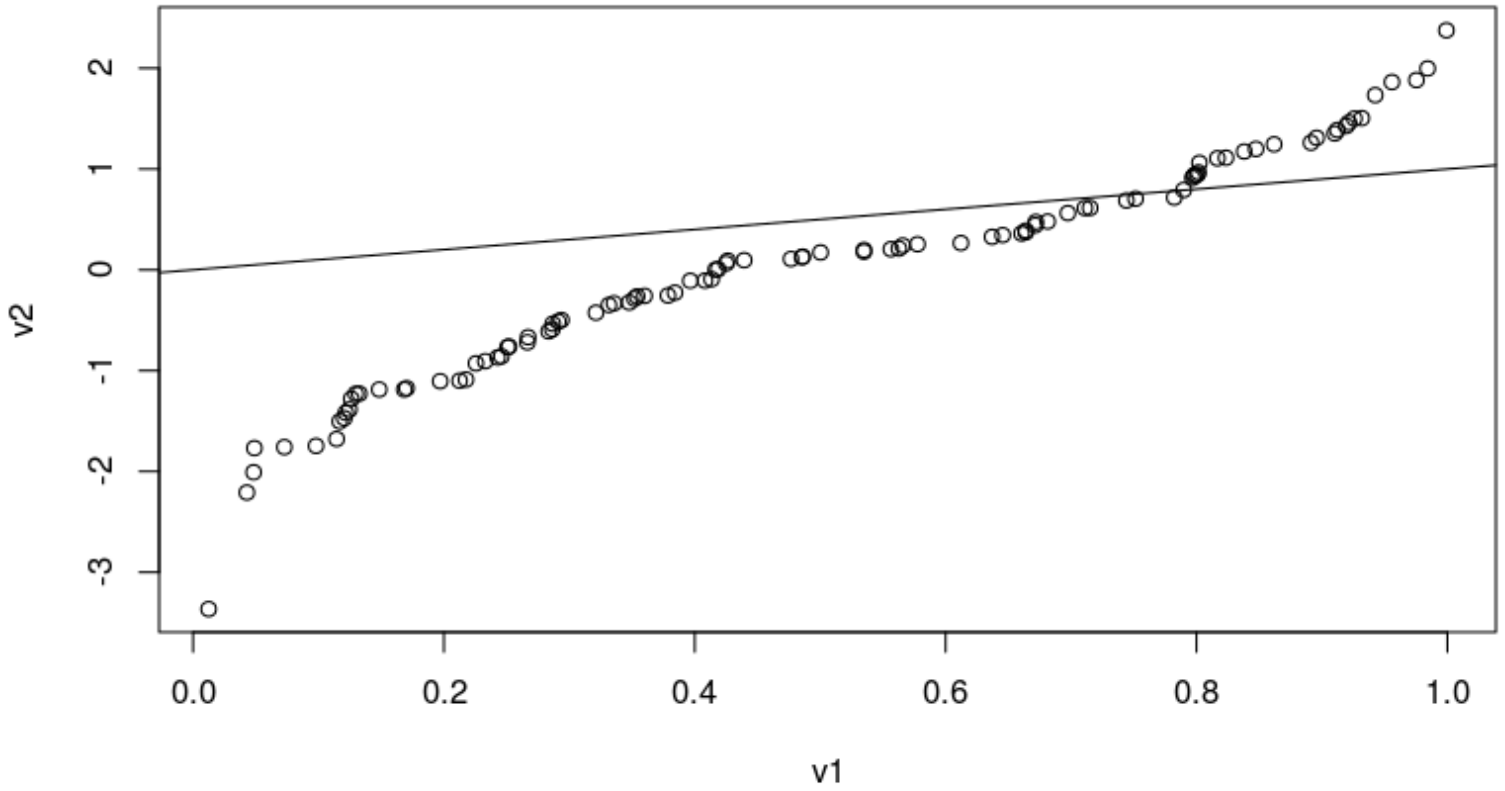
```
[1] 0.28732
```

```
mean(v2);sd(v2)
```

```
[1] -0.00197
```

```
[1] 1.0709
```

```
qqplot(v1, v2)
abline(0, 1)
```



离群点的探测

有三种主要的方法来确定离群点

- 确定一个基于业务或者统计的区间来定义离群值
- 用一个候补模型和一系列统计量来确定
- 用聚类的方法把数据集分割成较小的子集

对于离群点的定义是主观的，因此，当我们在数据挖掘中标记了一些变量是离群点，或许不一定意味着他们应当被删除

缺失数据

什么是缺失数据，他们是怎么出现的？

- 未知值
- 错误值
- 在一个数值变换中没有定义的值

处理缺失值的方法:

- 忽略这个观测记录
- 用一个其他值代替:
    - 0/NA
    - 均值/中位数/众数
    - 常数
- 插补数据（EM）
- 现实方法

缺失值的模式

```
library(VIM)
data(sleep)
which(is.na(sleep$Dream))
```

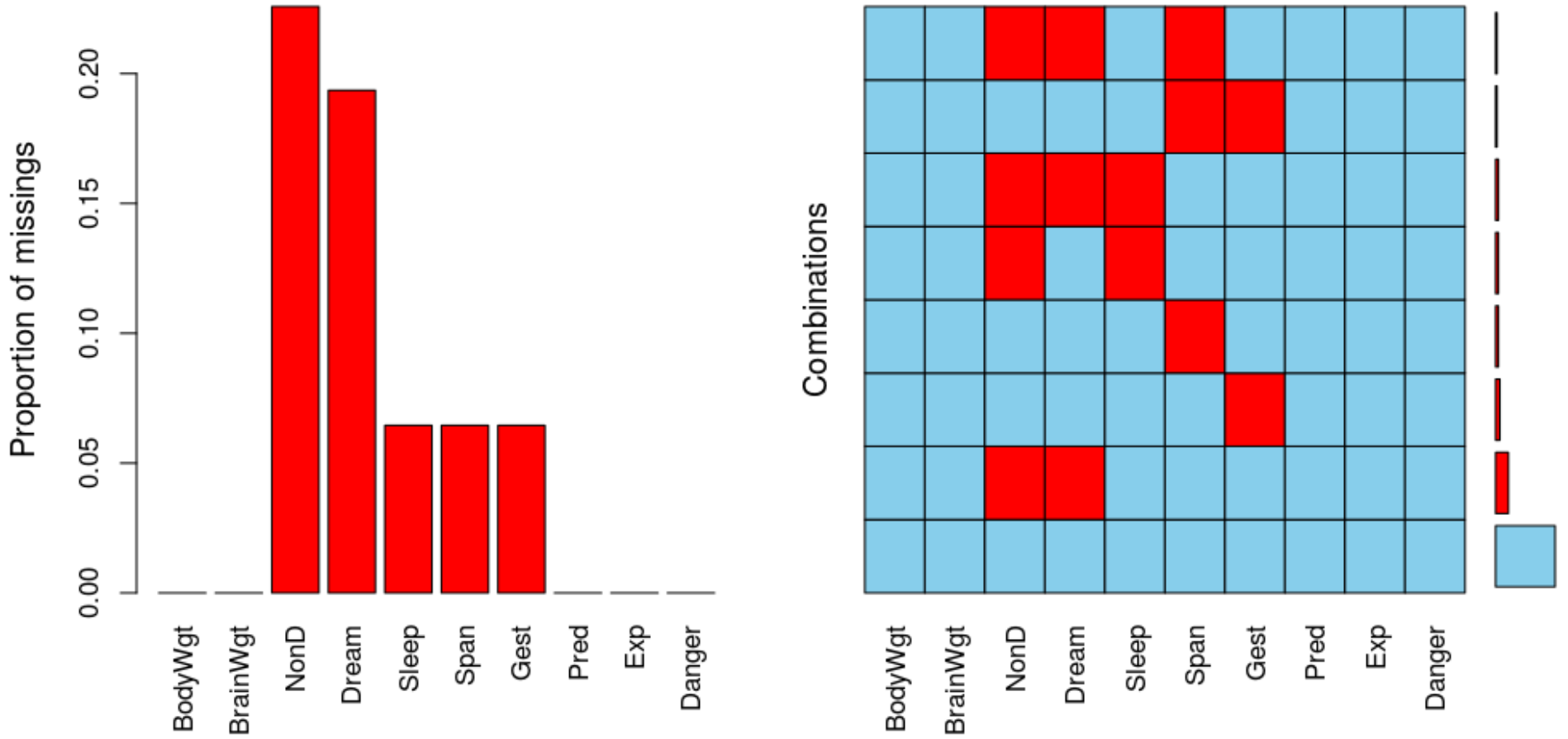
```
 [1]  1  3  4 14 24 26 30 31 47 53 55 62
```

```
complete.cases(sleep)
```

```
 [1] FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
[17]  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
[33]  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
[49]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
```

识别缺失值的模式

```
aggr(sleep)
```



用VIM包来插入缺失值

```
sleep_knn <- kNN(sleep)  #knn impuration
```

```
Time difference of 0.022922 secs
```

```
aggr(sleep_knn, delimiter="_imp")
```