Reshape2包

李茂宇1600022705
闫骏达1600022708

摘要

- Reshape2简介
- 宽数据 VS 长数据
- 主要函数介绍

Reshape2包介绍

- reshape2是一个强大的数据处理操作的R包。对于重塑数据格式很有用。

- 特性：

  改进算法，使计算与内存使用效能增强；

  用变量名来设定边际参数；

  删除cast中的一些特性，因为他确认plyr包能更好的处理；

  所有的melt函数族都增加了处理缺失值的参数

宽数据 VS 长数据

- 对于宽型数据，每列代表一个不同的变量。例如R内置的airquality数据集就是宽型数据：

#	ozone	wind	temp
# 1	23.62	11.623	65.55
# 2	29.44	10.267	79.10
# 3	59.12	8.942	83.90
# 4	59.96	8.794	83.97

宽数据 VS 长数据

- 对于长型数据，一列包含了所有可能的变量，另一列是对应的取值。上面的数据可以用长型数据来表示：

- 长数据有一列数据是变量的类型，有一列是变量的值。长数据不一定只有两列。ggplot2需要长类型的数据，plyr也需要长类型的数据，大多数的模型(比如lm(), glm()以及gam())也需要长数据。

```
#    variable   value
# 1     ozone  23.615
# 2     ozone  29.444
# 3     ozone  59.115
# 4     ozone  59.962
# 5      wind  11.623
# 6      wind  10.267
# 7      wind  8.942
# 8      wind  8.794
# 9      temp  65.548
# 10     temp  79.100
# 11     temp  83.903
# 12     temp  83.968
```

函数介绍

- reshape2包中两个主要的函数：melt和cast
- melt ——将宽型数据融合成长型数据。这是一种把多个类别列“融合”为一行的结构重组。我们来通过代码理解它是怎么运行的。
- cast——将长型数据转成宽型数据。它始于融合后的数据，然后把数据重新构造为长格式。它就是melt函数的反向操作。它包括两个函数，即dcast 和acast。dcast返回数据框作为输出结果。acast返回向量/矩阵/数组作为输出结果。

Melt函数

- melt函数对宽数据进行处理，得到长数据
- Melt（data，id.vars = null，variable.name = “variable”，
  value.name = “value”）
- 参数：data-----数据集
  id.vare-----设为主键的变量
  variable.name-----变量要改成的名称
  value.name-----变量值要改成的名称

Melt函数示例

- 此处用R内置的airquality数据集，首先将列名改成小写，然后查看相应的数据：

```
> names(airquality) <- tolower(names(airquality))
> head(airquality)
```

	ozone	solar.r	wind	temp	month	day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6

Melt函数示例

- 直接用metl函数处理上述的数据，得出结果

```
> aql <- melt(airquality) # [a]ir [q]uality [l]ong format
> head(aql)
```

```
 variable  value
1    ozone    41
2    ozone    36
3    ozone    12
4    ozone    18
5    ozone    NA
6    ozone    28
```

Melt函数示例

- 然后再看看末尾的几个数据

```
> tail(aql)
```

```
    variable value
913      day    25
914      day    26
915      day    27
916      day    28
917      day    29
918      day    30
```

Melt函数示例

- 默认情况下，melt认为所有数值列的变量均有值。很多情况下，这都是我们想要的情况。在这里，我们想知道每个月(month)以及每天(day)的ozone，solar.r, wind以及temp的值。因此，我们需要告诉melt，month和day是"ID variables"。ID variables就是那些能够区分不同行数据的变量，类似于数据库中的主键。

```
> aql <- melt(airquality, id.vars = c("month", "day"))
> head(aql)
```

	month	day	variable	value
1	5	1	ozone	41
2	5	2	ozone	36
3	5	3	ozone	12
4	5	4	ozone	18
5	5	5	ozone	NA
6	5	6	ozone	28

Melt函数示例

- 接下来修改长数据中的列名：

```
> aql <- melt(airquality, id.vars = c("month", "day"),
      variable.name = "climate_variable",
      value.name = "climate_value")
> head(aql)
```

	month	day	climate_variable	climate_value
1	5	1	ozone	41
2	5	2	ozone	36
3	5	3	ozone	12
4	5	4	ozone	18
5	5	5	ozone	NA
6	5	6	ozone	28

Cast函数

- cast函数用于把长格式数据转化成宽格式数据，有2个主要的函数：

  dcast——输出时返回一个数据框。

  acast——输出时返回一个向量/矩阵/数组
- cast（data， formula，fun.aggregate， na.rm = TRUE）
- 参数：

  data（数据集）

  formula（铸造公式，一般用来定义主键变量）

  fun.aggregate（聚合函数）

  na.rm = TRUE（删除空值）

Cast函数示例

- dcast借助于公式来描述数据的形状，左边参数表示"ID variables"，而右边的参数表示measured variables。可能需要几次尝试，才能找到合适的公式。
- 这里，我们需要告知dcast，month和day是ID variables，variable则表示measured variables，进而从宽格式数据变换到长格式的数据：

```
> aql <- melt(airquality, id.vars = c("month", "day"))
> aqw <- dcast(aql, month + day ~ variable)
> head(aqw)
```

	month	day	ozone	solar.r	wind	temp
1	5	1	41	190	7.4	67
2	5	2	36	118	8.0	72
3	5	3	12	149	12.6	74
4	5	4	18	313	11.5	62
5	5	5	NA	NA	14.3	56
6	5	6	28	NA	14.9	66

Cast函数示例

- 最后调整变量的顺序，恢复到原始数据：

```
> head(airquality) # original data
```

	ozone	solar.r	wind	temp	month	day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6

Cast函数示例

- 当每个单元有多个数据时，需要告诉dcast如何聚合(aggregate)这些数据，比如取均值(mean)，计算中位数(median)，或者简单的求和(sum)。比如，在这里，我们简单的计算下均值，同时通过na.rm = TRUE删除NA值：

```
> dcast(aql, month ~ variable, fun.aggregate = mean, na.rm = TRUE)
```

	month	ozone	solar.r	wind	temp
1	5	23.61538	181.2963	11.622581	65.54839
2	6	29.44444	190.1667	10.266667	79.10000
3	7	59.11538	216.4839	8.941935	83.90323
4	8	59.96154	171.8571	8.793548	83.96774
5	9	31.44828	167.4333	10.180000	76.90000

练习题

- 练习一：利用datasets包中的mtcars数据集做melt()和cast()的处理
- 练习二：将表1与表2中长宽数据进行互相转换

表1：长数据

车辆	道路等级	日均覆盖里程
车辆1	高速	100
车辆2	高速	111
车辆1	快速路	125
车辆2	快速路	110
车辆1	主要道路	30
车辆2	主要道路	76

表2：宽数据

车辆	高速覆盖里程	快速路覆盖里程	主要道路覆盖里程
车辆1	100	125	30
车辆2	111	110	76