大数据统计与计量分析——第四章作业

姓名：吴博成　班级：大数据91　　学号：2193211134

1 课本复现

1.1 数据背景

随着互联网的发展和Web2.0时代的到来,关于在线新闻这一信息全球化传播高速通道的研究兴趣也与目俱增。剖析网络新闻受关注程度的影响因素进而开展预测正在成为该研究领域的热点话问题的结论可以给作者和广告投放者提供重要的参考意见,甚至有助于舆情的考察和引导。

Kelwin Fernande和他的同事们通过新闻博客网站 Mashable 的数据,建立了预测新闻分享数量的智能决策支持系统 IDSS，是这一领域最新的杰作之一。 Mashable 是一个创办于2005年的互联网新闻博客,栏目包括社交媒体、科技、财经、环保等多个领域的新闻。 Mashable 现已成为世界上访问人数最多的博客之一,每月访问人数超过2000万,在各大社交网络上的粉丝人数超过1300万,《纽约时报》称其"已成为重要的新闻网站" Fernande 和他的同事们抓取了 Mashable 网站的海量博客新闻数据,包括39643篇文章、61个关于新闻的特征变量。本案例将基于这一数据。適过多种探索性数据分析的方法尝试降低自变量的维度，探索新闻特征及其受欢迎程度之间的关系。

Fernande提供的数据集抓取了2015年1月8日 Mashable 网站上所刊载的所有新闻博客文章,并提取出文章的基本信息。包括文章的分享数、标题包含的词汇数、正文包含的词汇数、文章包含的视频数等等。

1.2 数据描述性统计

在正式建模之前,应首先了解数据集的基本规律。使用R绘制变量之间的相关图。

得到的相关性图如图1所示。图中蓝色代表正相关,红色代表负相关,颜色越深代表相关度越高。从图中可以看到,变量之向有明显的相关性,部分变量基本可以归为一类。其中, n _ unique _ tokens , n _ non _ stop _ words , n _ non _ stop _ unique _ tokens 相关性极强,可以认为是一个指标,在后续的分析中可以只用其中一个代替; self _ reference_min_shares , self _ reference_max_shares , self reference_avg_shares 三个指标相关性很强,可以只用 avg 代替,消除共线性;描述文章所属频道的变量与主题模型 LDA 有一定的相关关系;描述文章主观性和情感强度的指标之间有较强的相关关系,可以考虑用一定的方法聚合。

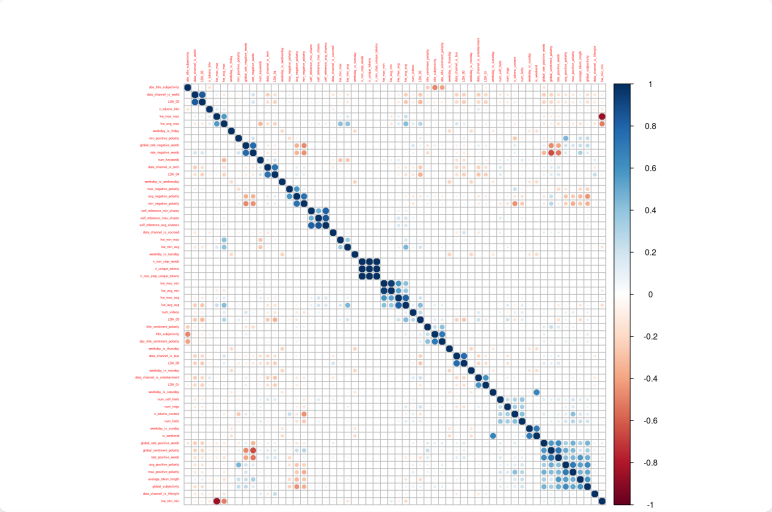

图1 各变量之间相关图

此外,通过观察发现文章频道、发布时间可以归为分类变量。针对这种情况，我们采取将定性变量与定量变量分开分析，并对定量变量进行投影降维的办法，对数据进行预处理。由于sample数据离散度过大，对其进行取log处理。书本代码有一些小错误均已修改。

```
library(corrplot)
library("REPPlab")
library(knitr) # 便于表格输出
data = read.csv("OnlineNewsPopularity.csv",sep=",")
mcor = cor(data[,3:60])  #绘制相关矩阵
corrplot(mcor,order = "hclust",tl.cex=0.3) #相关矩阵图，通过tl.cex参数调整label栏大小，美化图形
# 数据预处理
x_sc = scale(data[,3:60],center=F,scale = T)/sqrt(nrow(data[,3:60])-1)
C = t(x_sc)%*%x_sc
data.new=data
for (i in 14:19){
  data.new[,i]=factor(data.new[,i],levels = c(0,1),labels = c("no","yes"))}
for(i in 32:39){
  data.new[,i]=factor(data.new[,i],levels = c(0,1),labels = c("no","yes"))}
data.channel=data.new[,14:19]
data.weekday=data.new[,32:39]
data.num=data[,-c(1:2,14:19,32:39)]
```

1.3 建模过程

由于原始变量间相关性较强,将数据集中的星期和频道这两个定性变量进行单独分析,通过图形描述、假设检验等方法进行探索性分析。

```
ch=matrix(0,nrow(data.channel),1)
shares=log(data$shares) #书上的样例图片应该是进行了log变换的
for(i in 1:6){
  ch[data.channel[,i]=='yes',] = i}
boxplot(shares~ch,main="Boxplot of Shares vs. Channel")
ch1=data.frame(shares,as.factor(ch))
aov1 = aov(shares~ch,data=ch1)
summary(aov1)

wk=matrix(0,nrow(data.weekday),1)
shares=log(data$shares)
for(i in 1:7){
```

```
  wk[data.weekday[,i]=='yes',] =i
}
boxplot(shares~wk,main="Boxplot of Shares vs. Weekday")
wk1=data.frame(shares,as.factor(wk))
aov2 = aov(shares~wk,data=wk1)
summary(aov2)
```

我们首先绘制了文章关注度关于发布频道这一因素的箱线图(如图2所示,由于shares数据离散程度过大，进行了log调整，以1-6表示6种不同的频道,0表示没有包含在这6种频道中的其他文章),然后开展方差方析,可以直观地看出。不同频道的文章的关注度水平存在一定差异。方差分析的结果则证明了这一差异是显著的。

结合图2和表1,不同发布频道的文章对应的关注度有显著差异。其中, Social Media 频道的文章关注度高于其他，World 频道文章关注度低于其他，反映了人们对不同主题的关注程度。 Social Media 类的文章会让人们更愿意去分享和转发，而 World，Tech类的文章人们更愿意“阅后即焚”。

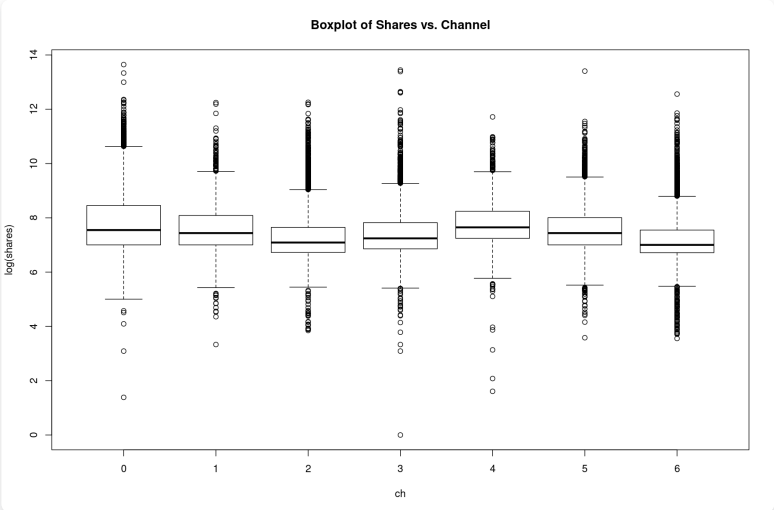

图2 文章关注度关于频道种类的箱线图

表1 文章关注度关于频道种类的方差分析表

	自由度	平方和	均方3	F值	P值
频道	1	609	609.0	716	<2e-16
残差	39642	33714	0.9		

对于星期因素,我们同样绘制了箱线图(见图3)并进行了方差分析(见表2)。结果表明,不同星期日时的文章分享数存在显著差异。文章发布时间也是影响关注度的潜在因素。下图显示了从周一到周日的分享量的分布。可以看出,周末的分享量高于其他,比较符合人们周末上网时间长的事实。

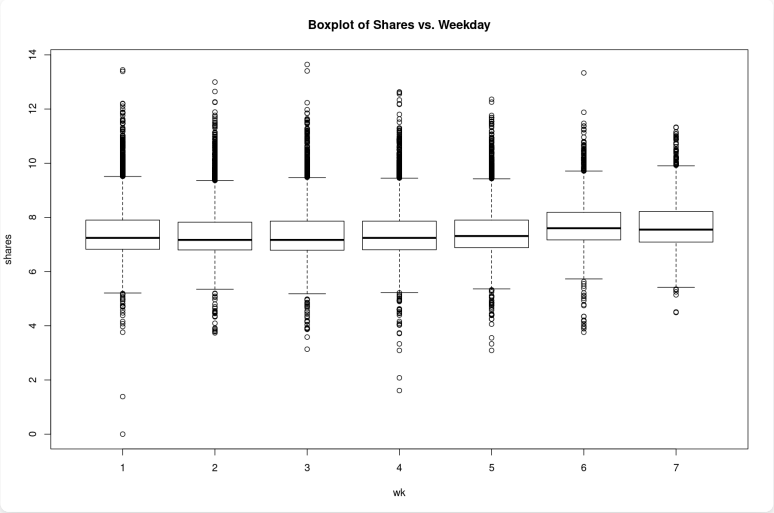

图3 文章关注度关于星期的箱线图

表2 文章关注度关于星期的方差分析表

	自由度	平方和	均方3	F值	P值
频道	1	236	236.07	274.5	<2e-16
残差	39642	34087	0.86		

由于定量自变量个数众多,以下将通过投影寻踪分析对自变量的数据规律进行探索。根据前面的分析,我们已经知道自变量之向存在较为复杂的相关关系,显现出若干相关性较强的群组,因而在 R 中选择 KurtosisMin 这个投影指标,帮助我们研究自变量间的群组特征、实现变量降维。使用 PSO 算法,经过20次仿真,得到使得峰度最小的排名前5个投影方向。对目标变量和5个投影变量做回归分析。

书本代码中缺少对y变量sample的赋值定义，已经予以增补。此外，对于回归模型的选取方面，由于PSO算法仿真得到的投影方向含有负值，不适合用log进行映射变换，故本文采取y~log(ep5)的模型进行拟合。（书本结果疑似通过y~ep5进行拟合）

```
library(REPPlab)
newdata_set = data[,-c(1,2,14:19,32:39,61)]
epplab = EPPlab(newdata_set,PPindex ="KurtosisMin",PPalg
="PSO",maxiter=50,n.simu=20,sphere=T)
coef.result=coef(epplab)
coef.result.first=coef.result[,1]
coef.result.second=coef.result[,2]
coef.result.third=coef.result[,3]
coef.result.forth=coef.result[,4]
coef.result.fifth=coef.result[,5]
ep5 = fitted(epplab,which=c(1:5))
y = data$shares
modelofEp5=lm(y ~ ep5)
kable(summary(modelofEp5)$coefficients,format = "markdown")
colnames(newdata_set)[c(order(abs(coef.result.first), decreasing = T)[1:5])]
colnames(newdata_set)[c(order(abs(coef.result.forth), decreasing = T)[1:5])]
```

结果如表3所示。可以发现方向1,3和4显著影响分享量(方向1，3影响为正,方向4影响为负),可以猜测,方向1和方向4聚合了对分享量影响较大的因素。

表3 投影方向与文章分享数的线性回归分析结果

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3395.3802	58.28492	58.254862	00.000000e+00
ep5Run1	635.7596	65.17178	9.755136	1.858779e-22
ep5Run2	205.7284	65.43253	3.144130	1.667042e-03
ep5Run3	444.1937	64.19883	6.919031	4.616353e-12
ep5Run4	-337.7450	59.98677	-5.630326	1.810842e-08
ep5Run5	293.4185	63.41782	4.626752	3.726193e-06

表4 与投影方向1相关性最强的原始变量

方向一	方向三	方向四
LDA_00	LDA_00	LDA_01
title_sentiment_polarity	LDA_02	LDA_04
LDA_03	title_sentiment_polarity	LDA_02
average_token_length	LDA_01	LDA_03
LDA_04	LDA_03	average_token_length

通过对投影方向各变量贡献排序,得到表4。考察原始变量对投影方向1和3，4的贡献,可以发现对分享量影响最大的一组变量正是文章的主题,其次是文章题目的情感性和正文平均词长。

1.4 结论和建议

综上,我们可以归纳出影响文章关注度的因素。

- 文章标题及文章主题,读者在点击或分享一篇文章时,往往先关注的是文章的标题或主题。文章标题的情感对立性越强,文章越容易获得较高的关注度，与特定主题模型相关度高的文章也往往可以获得更高关注。对应的网站如果想提升其文章的关注度,可以先在文章的题目和主题上做文章。
- 文章正文的长度来看，越长的文章越容易被分享推荐。

除了以上结论,探索性分析也为后续的分析提供了方向和思路,投影降维的结果可以作为后续分析的基础,在探索性分析中显示出显著影响的一些变量也应该在后续的分析(如机器学习)中予以适当关注。

总结本章的内容,可视化技术最直观地向我们展现数据的结构和形态,为进一步分析奠定了基础,投影寻踪让我们从量化的角度了解数据的结构特征,独立成分分析成功地解决了混合数据的分离问题,这三个步骤都是处理高维度大数据的关键步骤,经过这几步的探索,数据使用者对数据已经有了进一步的了解,探索性数据分析案例和探索性数据分析综合应用的介绍,加深了大家对大数据分析的认识。