南開大學

第六讲 主成分分析与聚类分析

提　纲

1. 主成分分析

【主要解决的问题】

在用统计分析方法研究多变量的课题时，变量个数太多就会增加课题的复杂性。人们自然希望变量个数较少而得到的信息较多。

在很多情形，变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，可以解释为这两个变量反映此课题的信息有一定的重叠。

主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

1. 主成分分析

【提出问题】

- 数据集**USJudgeRatings**包含了律师对美国高等法院法官的评分。数据框包含**43**个观测，**12**个变量。

变　量	描　述	变　量	描　述
CONT	律师与法官的接触次数	PREP	审理前的准备工作
INTG	法官正直程度	FAMI	对法律的熟稔程度
DMNR	风度	ORAL	口头裁决的可靠度
DILG	勤勉度	WRIT	书面裁决的可靠度
CFMG	案例流程管理水平	PHYS	体能
DECI	决策效率	RTEN	是否值得保留

如何以较少的变量来总结上述除**CONT**变量以外**11**个变量评估的信息？

1. 主成分分析

【基本概念】

信息太复杂是多变量数据处理的最大挑战之一，假如数据集有100个变量，想了解其中所有变量间的相互关系，需要考虑4950对相互关系，计算工作量非常大。有时候需要简化多变量数据集的复杂关系，节省计算资源，这个时候主成分分析（PCA）和探索性因子分析（EFA）是两种非常方便的方法。

探索性因子分析（EFA）的具体过程（同学们自学）

1. 主成分分析

主成分分析（PCA）

PCA的目标是用一组较少的不相关变量代替大量相关变量，同时尽可能保留初始变量的信息，这些推导所得的变量称为主成分，它们是观测变量的线性组合。如第一主成分为：

$$PC_1 = a_1X_1 + a_2X_2 + \ldots + a_kX_k$$

它是k个观测变量的加权组合，对初始变量集的方差解释性最大。第二主成分也是初始变量的线性组合，对方差的解释性排第二，同时与第一主成分正交（不相关）。后面每一个主成分都最大化它对方差的解释程度，同时与之前所有的主成分都正交。

1. 主成分分析

判断PCA中需要多少个主成分的准则：

- 根据先验经验和理论知识判断主成分数；
- 根据要解释变量方差的积累值的阈值来判断需要的主成分数
- 通过检查变量间$k \times k$的相关系数矩阵来判断保留的主成分数。

最常见的是基于特征值的方法。每个主成分都与相关系数矩阵的特征值相关联，第一主成分与最大的特征值相关联，第二主成分与第二大的特征值相关联，依此类推。

1. 主成分分析

【使用R的具体解决过程和结果解读】

- 主成分分析PCA（简易）

  - 使用 princomp(data, cor = T)函数进行主成分分析，cor = T是用相关系数进行主成分分析。

```
USJudge.pr <- princomp(USJudgeRatings[, -1], cor = T)
```

  - 观察主成分分析的详细情况，显示载荷矩阵

```
summary(USJudge.pr, loadings = T)
```

1. 主成分分析

```
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4      Comp.5      Comp.6     Comp.7      Comp.8       Comp.9      Comp.10
Standard deviation     3.1833029 0.65163561 0.50525195 0.302163952 0.193132512 0.140575118 0.1359209 0.0910572348 0.0780507762 0.0579730055
Proportion of Variance 0.9212198 0.03860263 0.02320723 0.008300278 0.003390924 0.001796488 0.0016795 0.0007537655 0.0005538112 0.0003055336
Cumulative Proportion  0.9212198 0.95982240 0.98302963 0.991329907 0.994720832 0.996517319 0.9981968 0.9989505847 0.9995043959 0.9998099295
                          Comp.11
Standard deviation     0.0457250007
Proportion of Variance 0.0001900705
Cumulative Proportion  1.0000000000

Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
INTG  0.289  0.574  0.118         0.375  0.510  0.230  0.285  0.145  0.103
DMNR  0.287  0.576 -0.177  0.240 -0.399 -0.514 -0.167  0.169         -0.105
DILG  0.304 -0.139  0.335  0.266  0.591 -0.298 -0.368         -0.355
CFMG  0.303 -0.310         0.478         -0.101  0.722                -0.207
DECI  0.302 -0.336         0.380 -0.399  0.448 -0.452  0.200  0.150  0.138
PREP  0.309 -0.125  0.229 -0.201         -0.336                0.717  0.252  -0.305
FAMI  0.307 -0.123  0.228 -0.524                              0.222         -0.544   0.452
ORAL  0.313                -0.229 -0.146         0.164 -0.274 -0.252  0.667   0.466
WRIT  0.311         0.148 -0.317 -0.237                      -0.493          -0.680
PHYS  0.281 -0.235 -0.820 -0.155  0.298                       0.272
RTEN  0.310  0.153 -0.201                0.234 -0.160 -0.797         -0.333
```

结果中的Comp.1、Comp.2、……、Comp.11是计算出来的主成分，Standard deviation代表每个主成分的标准差，Proportion of Variance代表每个主成分的贡献率，Cumulative Proportion代表各个主成分的累积贡献率。每个主成分都是11个变量的线性组合，并不是每个主成分的作用都非常关键，一般地，选择累积贡献率达到八成的前几个主成分即可（这个实例中直选1个即可，92%）。

1. 主成分分析

在得到主成分的基础上进行回归也好进行聚类也好，就不再使用原始的11个变量了，而是使用主成分的数据。但现在还没有各个样本的主成分的数据。

- 得到各个样本主成分的数据

```
pca_data <- predict(USJudge.pr)
```

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
AARONSON,L.H.	-0.59845117	0.30609639	-0.816725582	0.025176614	-0.009755941	0.036012469	-0.2348366851	-0.130996051	0.068847102
ALEXANDER,J.M.	2.40602604	0.72452812	-0.088022010	0.275385614	0.2	0	0.3555306213	-0.003854714	0.045762290
ARMENTANO,A.J.	0.22700541	0.19603194	0.018431386	0.116636631			2604011	-0.075510240	0.005492152
BERDON,R.I.	3.65863145	-0.16022318	0.260685996	-0.1655722			1515	0.091892662	0.083393846
BRACKEN,J.J.	-6.95286772	-0.87885607	1.168050345	0.2501			15	0.067358642	-0.038685905
BURNS,E.B.	2.47029760	0.55238532	0.12341135				3	-0.034896651	-0.027980492
CALLAHAN,R.J.	3.94790840	0.1574384	-0.2					-0.011207903	0.008920490
COHEN,S.S.	-8.09451081	-0.49074906	-1.173882077	-0.0379			4	0.154204604	-0.021009483
DALY,J.J.	3.70251770	0.11858607	0.018584698	0.30181			19	0.054764039	-0.110018552
DANNEHY,J.F.	1.06290294	-1.14546213	0.063826838	-0.3776206			1577	-0.085686686	-0.043218264
DEAN,H.H.	-0.34253639	0.14718925	-0.750913646	0.064917324			0031983	0.054804148	-0.042696660
DEVITA,H.J.	-1.46089767	0.81869959	0.444564651	0.100582722	-0.3180	1	0.0774628407	-0.121357234	-0.148645611
DRISCOLL,P.J.	-0.67945230	1.59011768	-0.759871543	-0.605747743	0.094639708	0.024540333	0.4484405349	0.071187889	0.131620934
GRILLO,A.E.	-3.26558547	0.25773707	0.938291097	-0.044447391	-0.335839901	0.328271558	-0.1135002451	-0.067726639	0.067341930
HADDEN,W.L.JR.	1.36589336	-0.23692972	-0.149863142	0.052678971	-0.196458425	-0.126813326	-0.0305239172	-0.016725504	-0.009376914
HAMILL,E.C.	-0.45055892	0.14383722	-0.200100464	0.039006404	0.296365151	-0.057498930	-0.0409752482	-0.085754294	-0.053015139
HEALEY.A.H.	-2.93029212	0.42309525	0.391546615	-0.202489965	0.430257385	-0.182356219	-0.0429475042	-0.145721914	0.016194160
HULL,T.C.	-0.72552521	-0.56957772	-0.256242324	-0.115281016	0.124082391	0.264054758	0.0555400528	-0.031096239	-0.203504942
LEVINE,I.	0.65174464	-0.10990281	0.134262313	-0.134455456	0.049224145	0.212200990	0.1267117702	-0.176180162	0.069594305
LEVISTER,R.L.	-4.40037232	-0.66464384	-0.342707665	0.456620168	0.270566309	-0.051158558	0.3495180361	0.039883674	0.098749170
MARTIN,L.F.	-1.84132491	1.32134885	-0.336819418	-0.093146099	0.282102551	-0.047199753	0.0133602860	-0.069886967	-0.112793494
MCGRATH,J.F.	-3.15554332	0.17639347	-0.475202996	0.279263472	-0.031137377	-0.168841388	-0.0548320789	0.017170217	-0.041071236
MIGNONE,A.F.	-6.43274239	1.54153721	1.647388365	-0.088319088	-0.094880939	-0.061112716	-0.0404109040	0.164429248	0.018417494
MISSAL,H.M.	-0.08730201	0.72868575	-0.045029510	0.197879846	0.073254640	-0.169017214	0.0631786944	0.075125192	-0.072646588
MULVEY,H.M.	3.39145598	-0.15700926	0.143299884	0.032712542	-0.075349166	-0.011219579	0.1370795316	-0.050543882	0.005660820

保留Comp.1的数据即可

2. 主成分分析分析练习

【操作练习】**airdata.csv**是预测空气质量可能的相关因素，数据已经进行了归一化处理，**f1_f9**分别代表气压、气温、相对湿度、云量、日照时数、风速、**SO2**浓度、 **NO2**浓度、**PM2.5**浓度、**PM10**浓度。 对这**1758**条数据做主成分分析，并生成主成分数据。

【提示】

```
setwd("D:/R")
data<-read.csv("airdata.csv")
airdata.pr <- princomp(data, cor = T)
summary(airdata.pr, loadings = T)
pca_data <- predict(airdata.pr)
pca<-data.frame(pca_data)
write.csv(pca, 'airpca.csv')
```

2. 主成分分析分析练习

【操作练习提示】

```
setwd("D:/R")
data<-read.csv("airdata.csv")
airdata.pr <- princomp(data, cor = T)
summary(airdata.pr, loadings = T)
pca_data <- predict(airdata.pr)
pca<-data.frame(pca_data)
write.csv(pca, 'airpca.csv')
```

3. 聚类分析

机器学习算法主要就是分类和回归。

聚类(clustering)，是一个把数据对象划分成子集的过程，每个子集是一个簇(cluster)，使得簇中的对象彼此相似，但与其他簇中的对象不相似。聚类也称为自动分类，聚类可以自动的发现这些分组，这是其突出的优点。

本部分作为大数据分析入门，仅以聚类分析为例。感兴趣的同学可以参考《R语言预测实战《R语言实战：机器学习与数据分析》等参考资料，学习更多关于机器学习的东西。

【提示】聚类分析前要对数据进行标准化（scale），然后做主成分分析（降维）。

3. 聚类分析

- 聚类分析的几个应用场景
  （https://blog.csdn.net/liulingyuan6/article/details/53637812）
  - 基于用户位置信息的商业选址
  - 中文地址标准化处理
  - 国家电网用户画像
  - 非人恶意流量识别
  - 求职信息完善
  - 搜索引擎查询聚类以进行流量推荐
  - 生物种群固有结构认知
  - 保险投保者分组
  - 网站关键词来源聚类整和
  - 图像分割

3. 聚类分析

- 聚类分析是非监督学习的很重要的领域。所谓非监督学习，就是数据是没有类别标记的，算法要从对原始数据的探索中提取出一定的规律。
- 聚类分析就是试图将数据集中的样本划分为若干个不相交的子集，每个子集称为一个“簇”。

3. 聚类分析

- 聚类的经典算法——层次聚类法

最开始的时候将所有数据点本身作为簇，然后找出距离最近的两个簇将它们合为一个，不断重复以上步骤直到达到预设的簇的个数。

步骤：

- 定义问题与选择分类变量
- 选择聚类方法
- 确定群组数目
- 聚类结果评估结果的描述、解释

3. 聚类分析

R语言中使用

hclust(d, method = "complete", members=NULL)

来进行层次聚类。

method表示类的合并方法，有：

single	最短距离法
complete	最长距离法
median	中间距离法
mcquitty	相似法
average	类平均法
centroid	重心法
ward	离差平方和法

3. 聚类分析

聚类分析过程

1. 数据的标准化

数据的标准化（normalization）是将数据按比例缩放，使之落入一个小的特定区间。在某些比较和评价的指标处理中经常会用到，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。

最典型的就是数据的归一化处理，即将数据统一映射到[0,1]区间上，常见的数据归一化的方法有：

3. 聚类分析

- **min-max标准化(Min-max normalization)**

也叫离差标准化，是对原始数据的线性变换，使结果落到[0,1]区间，转换函数如下：

$$x^* = \frac{x - min}{max - min}$$

其中max为样本数据的最大值，min为样本数据的最小值。这种方法有一个缺陷就是当有新数据加入时，可能导致max和min的变化，需要重新定义。

3. 聚类分析

- log函数转换

通过以10为底的log函数转换的方法同样可以实现归一下，具体方法如下：

$$x^* = log_{10}(x)/log_{10}(max)$$

- atan函数转换

用反正切函数也可以实现数据的归一化：

$$x^* = atan(x) * 2/\pi$$

使用这个方法需要注意的是如果想映射的区间为[0,1]，则数据都应该大于等于0，小于0的数据将被映射到[-1,0]区间上。

3. 聚类分析

z-score 标准化(zero-mean normalization)

并非所有数据标准化的结果都映射到[0,1]区间上，其中最常见的标准化方法就是Z标准化，也是SPSS中最为常用的标准化方法：

也叫标准差标准化，经过处理的数据符合标准正态分布，即均值为0，标准差为1，其转化函数为：

$$x^* = \frac{x - \mu}{\sigma}$$

其中μ为所有样本数据的均值，σ为所有样本数据的标准差。

3. 聚类分析

【问题提出】

如何对下列的新疆地区进行分类?

place	height	waterfall	icesoildepth	windday
哈巴河	532.6	173.8	150	61.8
阿勒泰	735.1	191.5	146	37.7
克拉玛依	427	114.4	197	75.4
巴楚	1116.5	41.6	64	7.6
莎车	1231.2	42.5	93	11
于田	1427	46.4	81	1.4

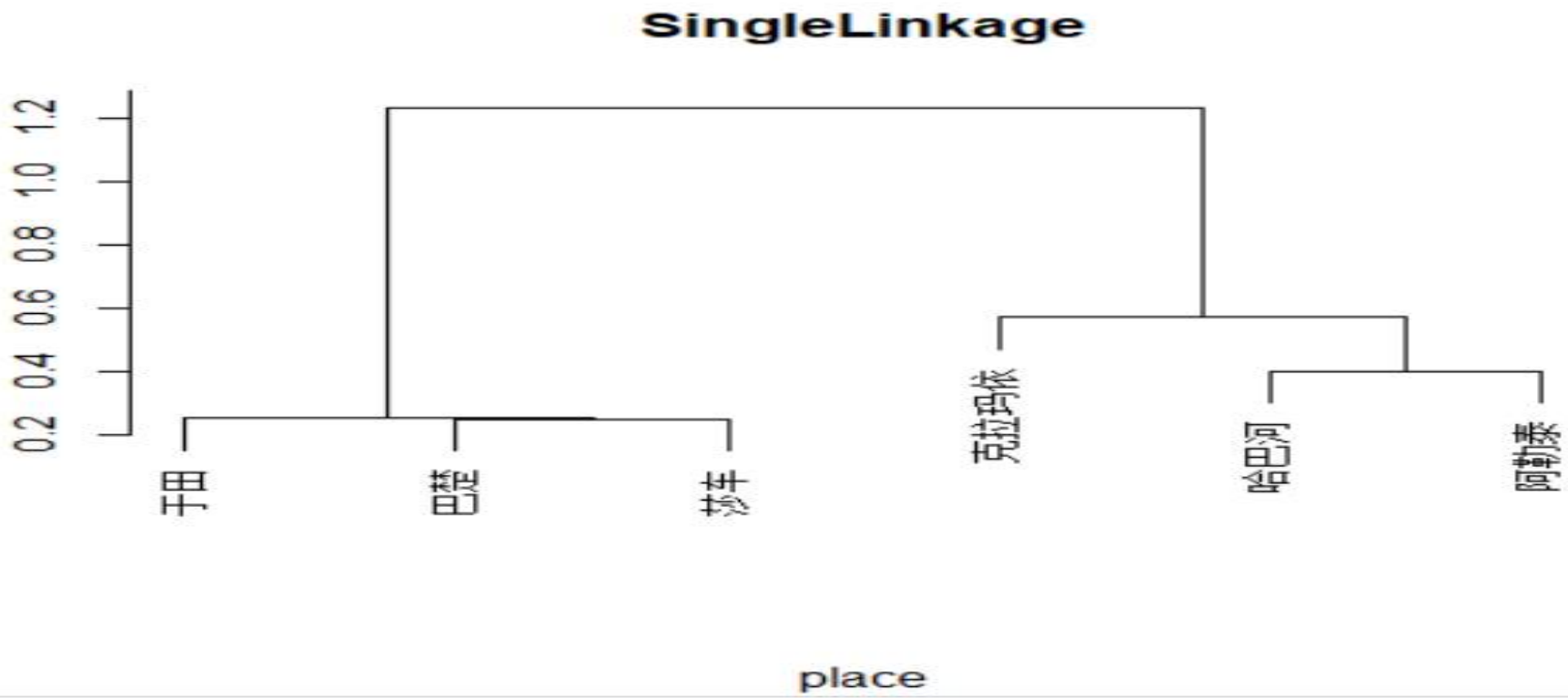
3. 聚类分析

【R操作及解读】

```
#读入数据
xinj<-read.csv("xinjiang.csv",header = TRUE)
#数据归一化 use method "min-max"
fun <- function(x) (x-min(x))/(max(x)-min(x))
xj<- apply(xinj[,2:5], 2, FUN=fun)
xj<-data.frame(xinj[,1],xj)
#最短距离法聚类
d<-dist(xj[2:5])  #计算各观测值之间的欧式距离（默认）
hc.single=hclust(d,method = "single")
```

3. 聚类分析

```
#制作聚类图
plot(hc.single,main = "SingleLinkage",
        xlab="",labels=xj$xinj...1.,ylab="",
        sub = "place",cex=.9)
```



3. 聚类分析

【练习1】数据文件为city1999.csv是1999年全国31个省市自治区的城镇居民全年消费性支出的八个主要指标，这八个变量是

x1 食品
x2 衣着
x3 家庭设备用品及服务
x4 医疗保健
x5 交通与通信
x6 娱乐教育文化服务
x7 居住
x8 杂项

分别用最长距离法，类平均法，对各地区做层次聚类分析。

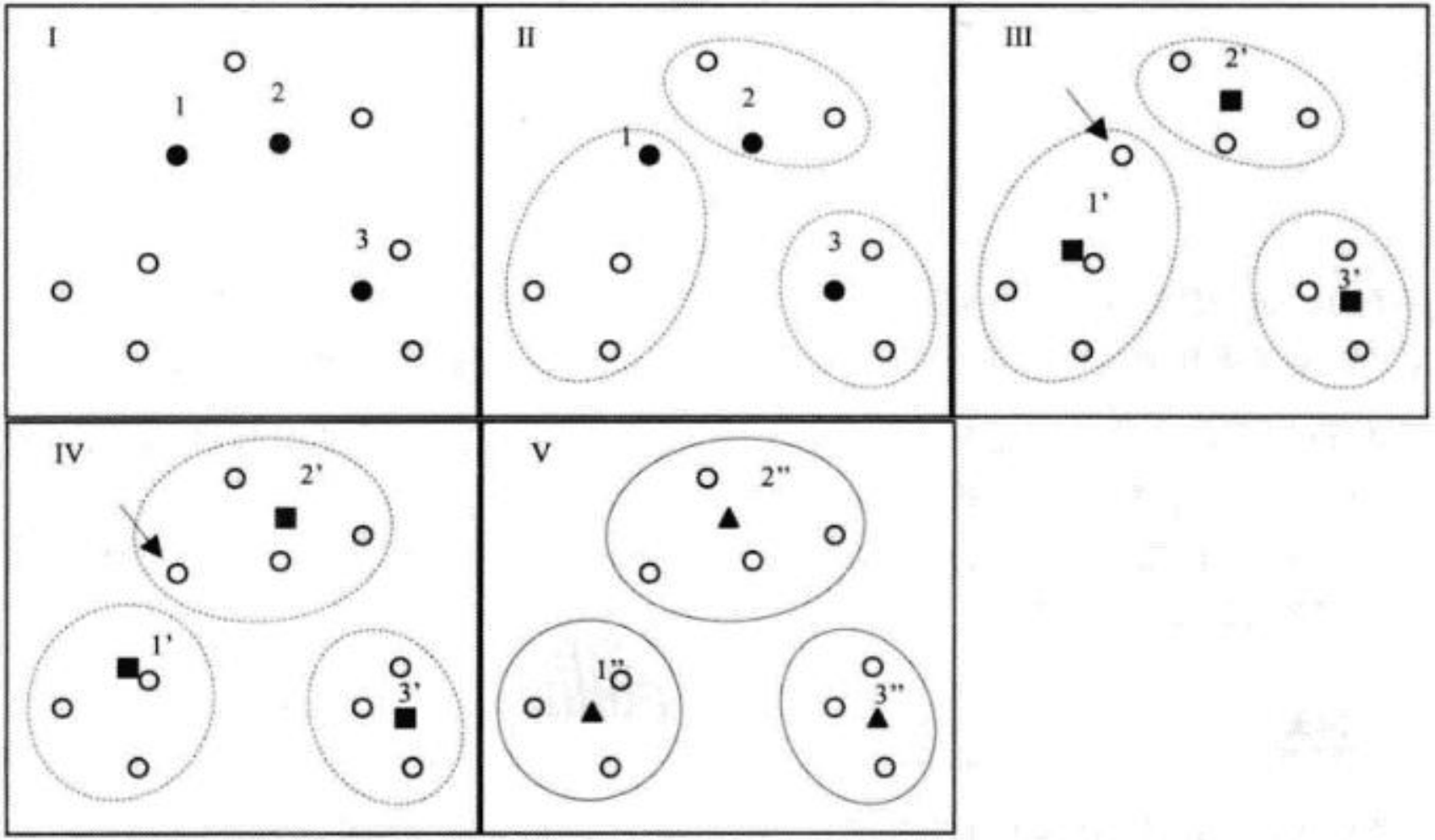
3. 聚类分析

- 聚类的经典算法——K-Means

K-Means算法是输入聚类个数k，以及包含 n个数据对象的数据库，输出满足方差最小标准k个聚类的一种算法。K-Means 算法接受输入量 k ；然后将n个数据对象划分为 k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。

K-Means算法原理：https://www.cnblogs.com/nxld/p/6376496.html

3. 聚类分析



3. 聚类分析

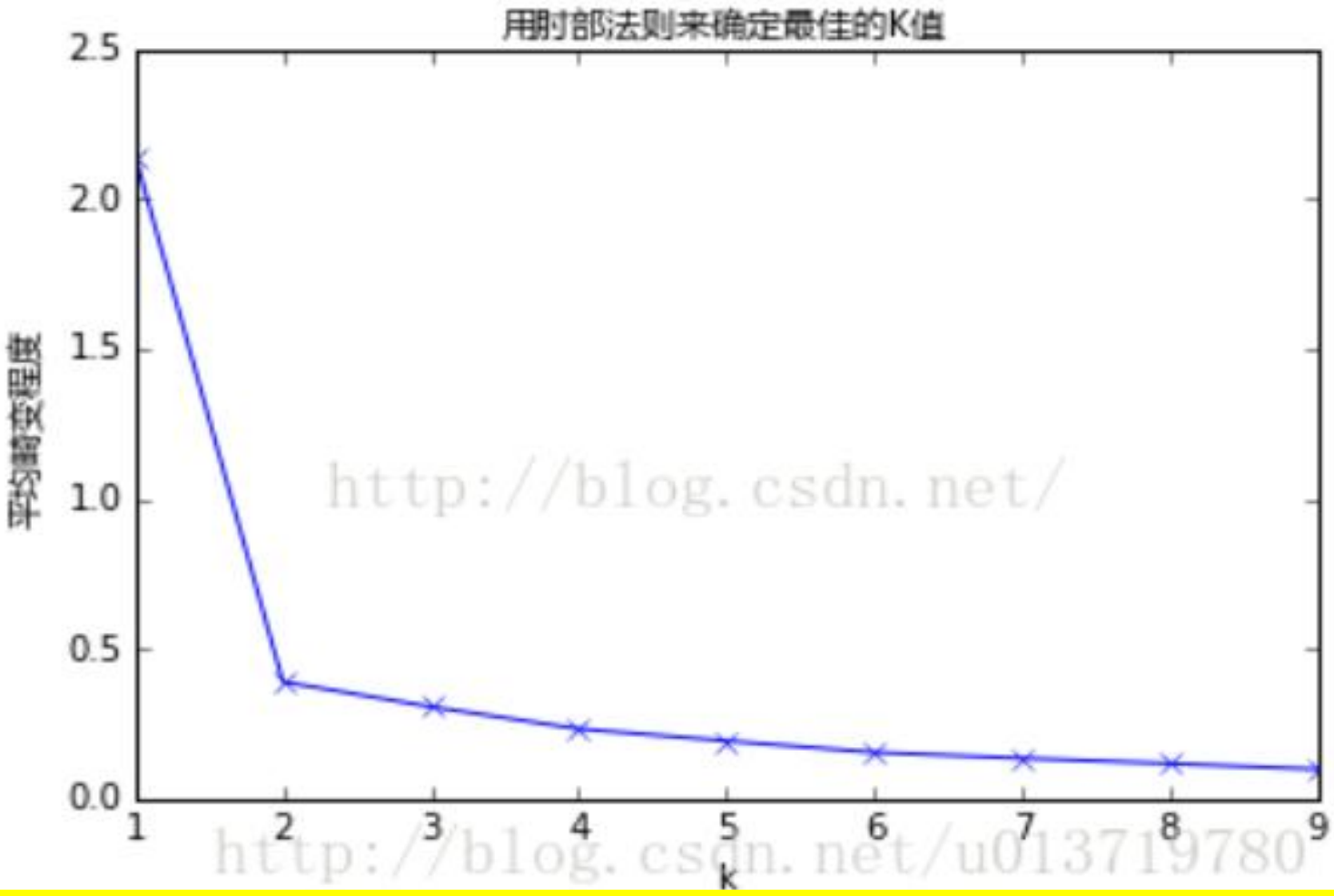
- K值的确定

1.根据问题内容确定——具体问题具体分析

2.肘部法则

肘部法则会把不同k值的成本函数值画出来。随着k值的增大，平均畸变程度会减小；每个类包含的样本数会减少，于是样本离其重心会更近。但是，随着k值继续增大，平均畸变程度的改善效果会不断减低。k值增大过程中，畸变程度的改善效果下降幅度最大的位置对应的k值就是肘部。下面让通过一张图用肘部法则来确定最佳的k值。下图数据明显可分成两类。

3. 聚类分析



http://blog.csdn.net/
http://blog.csdn.net/u013719780

从图中可以看出，k值从1到2时，平均畸变程度变化最大。超过2以后，平均畸变程度变化显著降低。因此最佳的k是2。

3. 聚类分析

初始质心的选取

1. 随机的选取初始中心，但是这样簇的质量常常很差。处理选取初始质心问题的一种常用技术是：多次运行，每次使用一组不同的随机初始质心，然后选取具有最小SSE(误差的平方和)的簇集。

2. 取一个样本，并使用层次聚类技术对它聚类。从层次聚类中提取k个簇，并用这些簇的质心作为初始质心。

3. 随机地选择第一个点，或取所有点的质心作为第一个点。然后，对于每个后继初始质心，选择离已经选取过的初始质心最远的点。

4. canopy算法。

3. 聚类分析

聚类效果评估

K-Means是一种非监督学习，没有标签和其他信息来比较聚类结果。但是，有一些指标可以评估算法的性能。例如类的畸变程度的度量方法。

另一种聚类算法效果评估方法称为轮廓系数(Silhouette Coefficient)。轮廓系数是类的密集与分散程度的评价指标。它会随着类的规模增大而增大。彼此相距很远，本身很密集的类，其轮廓系数较大，彼此集中，本身很大的类，其轮廓系数较小。轮廓系数是通过所有样本计算出来的，计算每个样本分数的均值，计算公式如下：

$$s=(a-b)/max(a,b)$$

a是每一个类中样本彼此距离的均值，b是一个类中样本与其最近的那个类的所有样本的距离的均值。

3. 聚类分析

- 使用R语言可以轻松实现聚类分析，stats、cluster、fpc和mclust是常用的四个聚类分析软件包。
  - stats主要包含一些基本的统计函数，如用于统计计算和随机数生成等；
  - cluster专用于聚类分析，包含很多聚类相关的函数及数据集；
  - fpc含有若干聚类算法函数，如固定点聚类、线性回归聚类、DBSCAN聚类等；
  - mclust主要用于处理基于高斯混合模型，通过EM算法实现的聚类、分类以及密度估计等问题。

3. 聚类分析

在R语言，使用内置的kmeans函数：

kmeans(x, centers, iter.max = 10, nstart = 1,
algorithm = c("Hartigan-Wong", "Lloyd", "Forgy","MacQueen"), trace=FALSE)

其中，

x为进行聚类分析的数据集；

centers为预设类别数k；

iter.max为迭代的最大值，且默认值为10；

nstart为选择随机起始中心点的次数，默认取1

algorithm则提供了4种算法选择，默认Hartigan-Wong算法。

3. 聚类分析

【问题提出】 使用K-means，基于city1999.csv数据，将全国31个省市自治区分为5类。

```
#读入数据
getwd()
setwd("e:/")
city<-read.csv("city1999.csv")
summary(city)
```

3. 聚类分析

```
#使用x1_x8这些属性进行聚类
citydata1<-city[,2:9]
#进行kmeans聚类前，使用scale函数进行数据标准化
citydata=scale(citydata1)
#加载fpc包，使用pamk函数估计聚类个数
install.packages("fpc")
library(fpc)
pamk.result=pamk(citydata)
pamk.result$nc # 结果是2，根据国情，下面实际用5
```

3. 聚类分析

```
#使用kmeans函数,iter.max是最大迭代次数，可写可不写，但是数据量很大的时候，一定要写防止死机。
#kmd=kmeans(citydata,centers =
      pamk.result$nc,iter.max = 100)
kmd=kmeans(citydata,centers = 5,iter.max = 100)
#输出聚类结果
type=kmd$cluster
#查看聚类结果分布
table(type)
#聚类中心结果输出
centerver=kmd$centers
centerver
```

3. 聚类分析

```
#将聚类中心结果写入本地excel中,用excel雷达图来描述聚类因子特征，找出每类的优势特征和劣势特征
write.csv(centerver, 'citycenterver.csv')
#把聚类结果添加到原始数据中
city<-cbind(city,type=kmd$cluster)
#把聚类结果添加到原始数据文件中
write.csv(city, 'city1999new.csv')
```

4. 聚类分析练习

【操作练习】参考给出的文档，完成南开大学本科助教聚类分析问题。