MCLUST包

主讲人：赵凯阳

R语言中实现聚类的包有很多

聚类算法	软件包	主要函数
K-means	stats	kmeans()
K-Medoids	cluster	pam()
系谱聚类（HC）	stats	hclust(), cutree(), rect.hclust()
密度聚类（DBSCAN）	fpc	dbscan()
期望最大化聚类（EM）	mclust	Mclust(), clustBIC(), mclust2Dplot(), densityMclust()

EM(最大期望算法)原理

- EM算法是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐藏变量，最大期望经常用在机器学习和计算机视觉的数据聚类领域。
- 第一步：计算期望E，利用对隐藏变量的现有估计值，计算其最大似然估计值。
- 第二步：最大化，最大化在E步上求得的最大似然值来计算参数的值，M步上找到的参数估计值被用于下一个E步计算中，这个过程交替进行。

MCLUST应用举例1

- GMM 模型常用于基于模型的聚类分析。 GMM中的每一个高斯分布都可以代表数据的一类，整个数据就是多个高斯分布的混合。
- 在R中的Mclust包中的Mclust函数可以用来进行基于GMM的聚类分析。
- 下面即是以最常用的iris数据集为例。

第一步：MCLUST 包的加载与安装

```
library(mclust)
```

第二步：构建EM算法模型

- model.EM<-Mclust(subset(iris,select=-Species))

MCLUST包应用举例

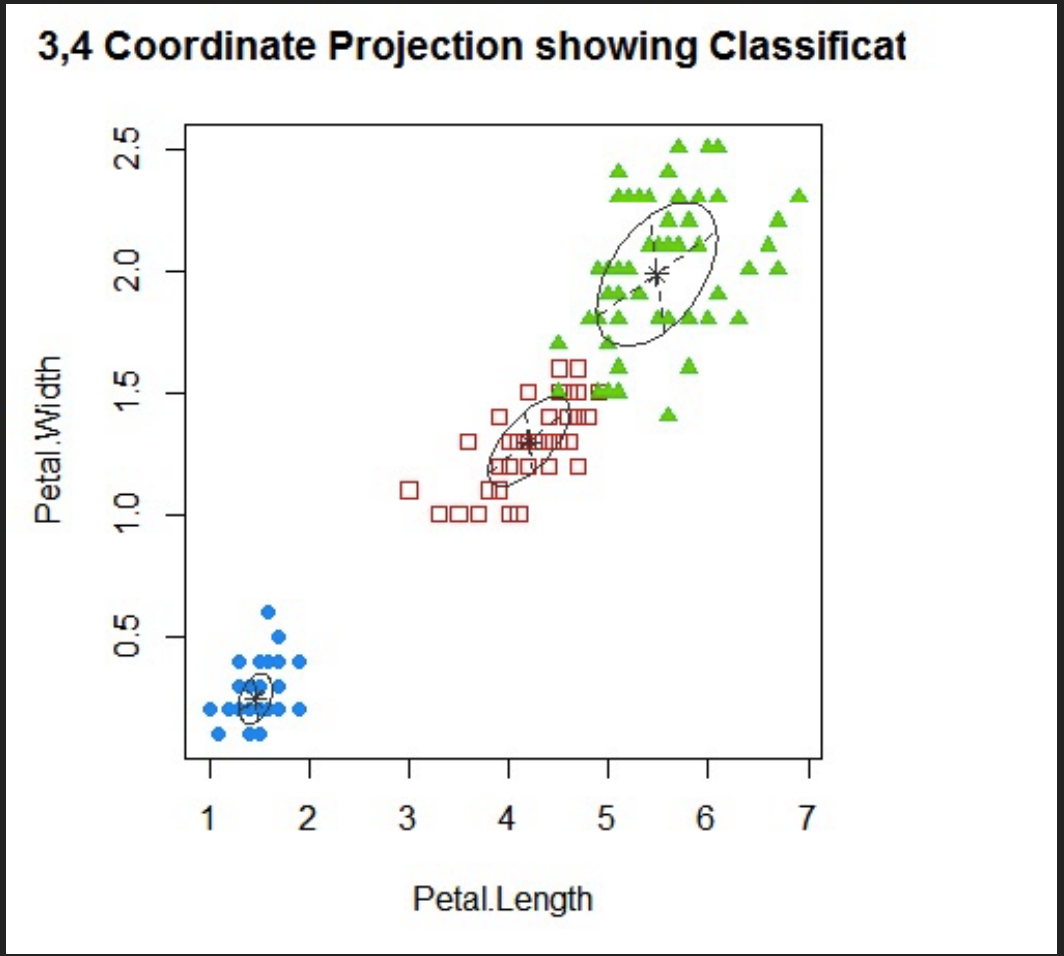
通常在机器学习中，我们将学习一组参数，
使我们观察的数据的可能性最大化。
但是，如果我们没有观察到数据中存在一些隐藏的变量，
该怎么办？ 期望最大化是使用参数来估计隐藏变量的
概率分布的一种非常常见的技术，计算预期的可能性，
然后找出将使预期可能性最大化的参数。 可以解释如下

MCLUST包举例

- library(mclust)
- mc<-Mclust(iris[,1:4,3]
- plot(mc,data=iris[,1:4],what="classification,dimens=c(3,4)
- table(iris$Species,mc$classification)

MCLUST

	1	2	3
setosa	50	0	0
versicolor	0	45	5
virginica	0	0	50



MCLUST应用举例2

- 假设现在有100个人的身高数据，而且这100条数据是随机抽取的。男女生的身高满足正态分布，但这两个分布的参数不同。我现在不仅不知道男女身高分布的参数，甚至不知道这100条数据哪些来自男生，哪些来自女生。这正符合聚类分析的假设（除了数据之外，并不知道其他任何信息）。而我们的目的正是推断每个数据应该属于哪个分类。所以对于每个样本，都有两个需要被估计的项
- 一个就是它到底来自男生身高的分布，还是女生身高的分布
- 另外一个就是，男女身高分布的参数各是多少

MCLUST应用举例2

- 在开始状态下，二者都是未知的，但如果知道A的信息就可以得到B的信息，反之亦然。

- 所以首先赋予A某种初值，以此得到B的估计，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到熟练为止。

MCLUST应用举例2

- > my.em<-Mclust(countries)
- >summary(my.em)
- Mclust EVI(diagonal,equal volume, varying shape) model with 2 components:
- log.likelihood n df BIC ICL
- -179.2962 30 8 -385.802 -385.8023

MCLUST应用举例2

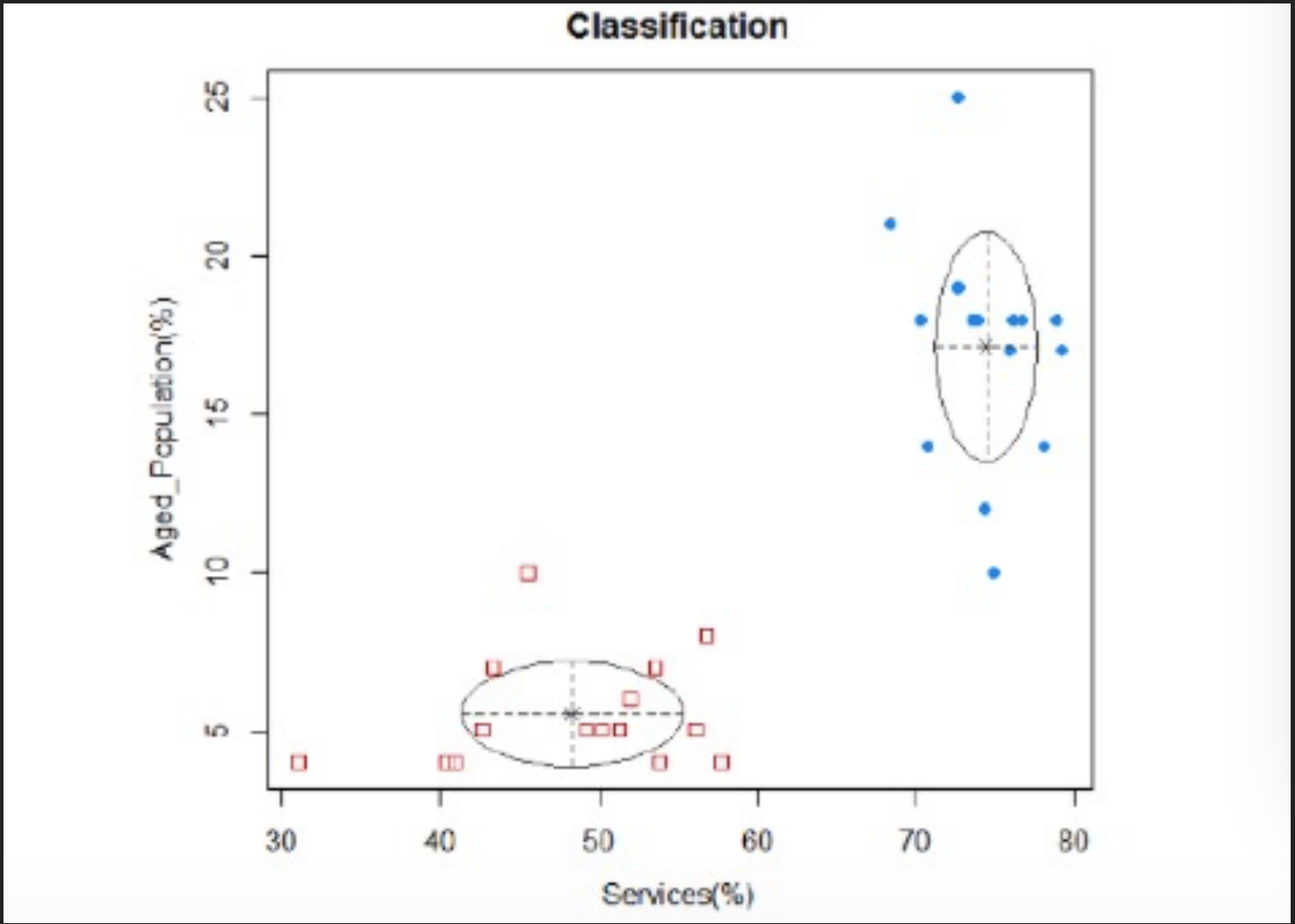
- Clustering table:
- 1 2
- 15 15

MCLUST应用举例2

- 用mclust软件包里面的mclust2Dplot()来把数据图像化

- > mclust2Dlpot(countries,parameters=my,em$parameters
- +　　z=my.em$z,what="classification", main=TRUE)

MCLUST应用举例2



MCLUST应用举例2

- 如果把上述代码中的参数值“classification”修改成“uncertainty”,那么绘制出的图所展示的是分类结果的不确定性情况。颜色越深，面积越大，表示不确定性越高。

MCLUST应用举例2

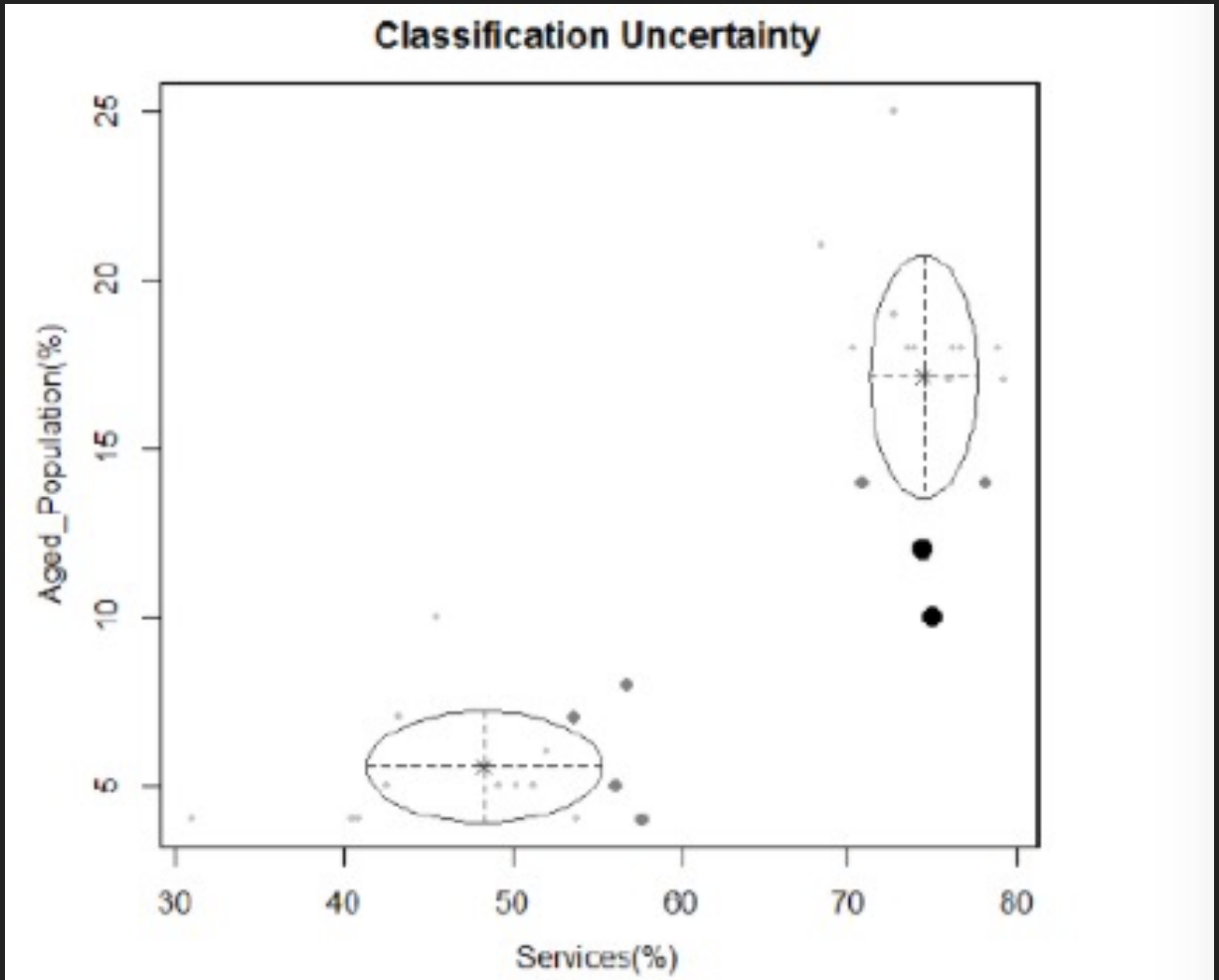


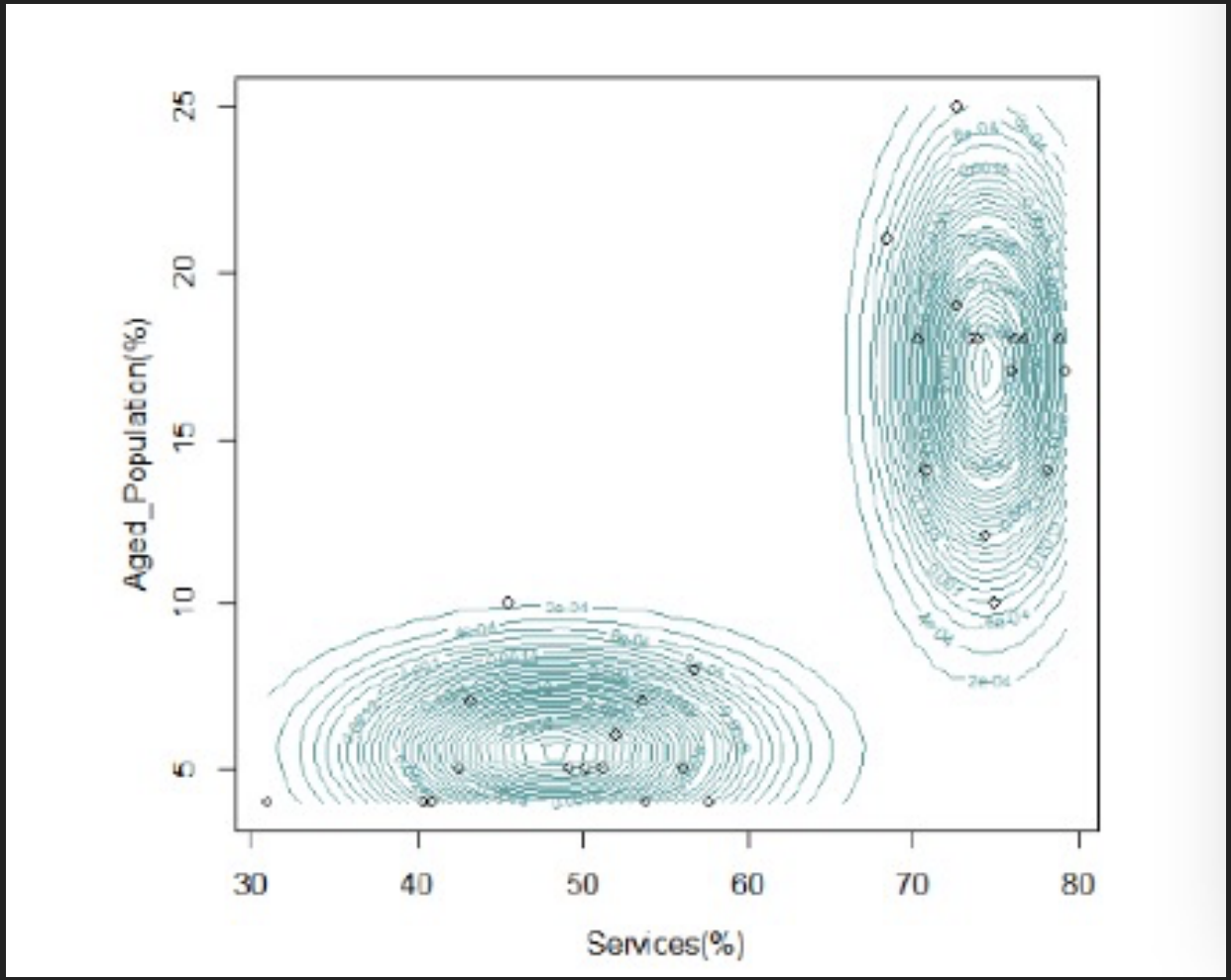
图 13-7　高斯混合模型聚类的不确定性

MCLUST应用举例2

- 借助densityMclust()函数，还可以绘制出高斯混合的密度图。

- > model density <- densityMclust(countries)
- >plot(model_density,countries,col=cadetblue",
- nlevels =25 , what = "density")

MCLUST应用举例2



MCLUST应用举例2

- >plot(model_density,what="density",type="persp",theta=25)
- theta用于控制三维图像水平方向上的旋转角度