Final Project

Olivetti Faces Dataset(`http://cs.nyu.edu/~roweis/data.html`)包含了一些人的脸图, 该数据是AT&T剑桥实验室于1992年4月到1994年4月期间采集的. 该数据集包含了40位人的灰度脸图, 每张图的分辨率为64x64, 每个人共有10张不同姿态下的脸图. 有些人的脸图是在不同时间点,不同光照,以及不同脸部表情(睁/闭眼睛, 微笑/不笑)下拍摄的. 所有图片都是在黑色均匀背景下拍摄. 我们从该数据集提取出来变量faces的值并存储为*faces.txt*, 其为为400x4096矩阵. 本报告以分析该数据为目的.

该数据集被广泛用于验证统计学习方法的性能, 由于每个人仅有10次重复观测, 因此更适合用于无监督的或者半监督的统计学习方法中. 例如Bien, J., and Tibshirani, R. (2011), Verma et al. (2013), Calandriello et al. (2013)等等.

问题

- 选择合适的聚类方法对该数据集进行分析, 分析聚类方法的性能.
- 考虑使用判别与分类方法, 主成分方法对该数据进行分析, 评价你的分类器性能.

要求

- 分析报告应至少包含上述问题中的一个.
- 分析报告应按照学术论文格式和要求进行:
  - 标题, 作者, 摘要, 内容(简介, 方法, 分析结果), 总结, 参考文献
  - 内容简洁, 叙述清晰流畅, 分析过程严谨
- 分析报告的写作请遵守学术道德规范, 使用LaTeX软件书写, 独立完成.
- 分析方法以使用所学方法为基本, 但不限于此.
- 分析报告双面(黑白)打印, 页数不超过5页.
- 提交截止时间: **6月17日**
- 提交地点: 管理科研楼1006

评价标准

- 格式,内容的完整性, 报告写作的流畅性
- 分析过程的完整性, 结果表达的严谨性

初步聚类分析例

读取数据和图示所有脸图:

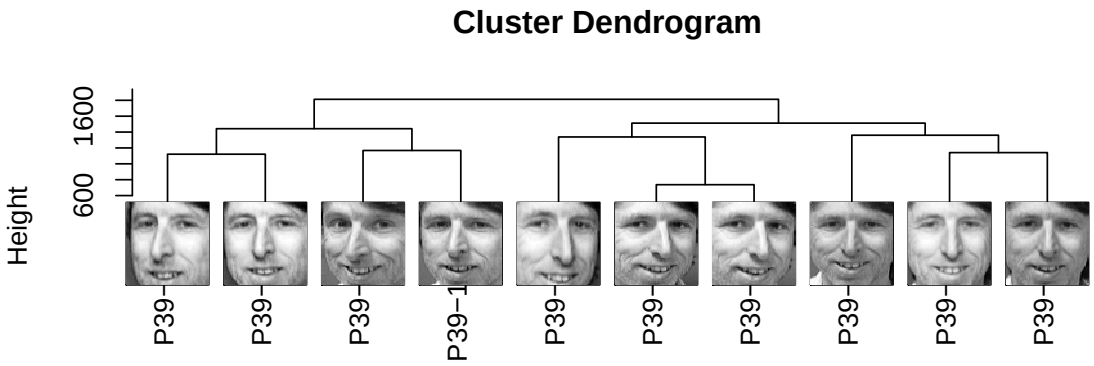
```
library(RColorBrewer)
showMatrix <- function(x, ...)
   image(t(x[nrow(x):1,]), xaxt = 'none', yaxt = 'none',
     col = rev(colorRampPalette(brewer.pal(9, 'Greys'))(100)), ...)

faces<-read.tables("faces.txt")
x<-as.matrix(faces)
par(mfrow=c(20,20),mar=rep(0,4))
for(i in 1:400)
showMatrix(matrix(x[i,],64,64))
rownames(x)<-paste("P",rep(1:40,each=10),"-",rep(1:10,40),sep="")
```

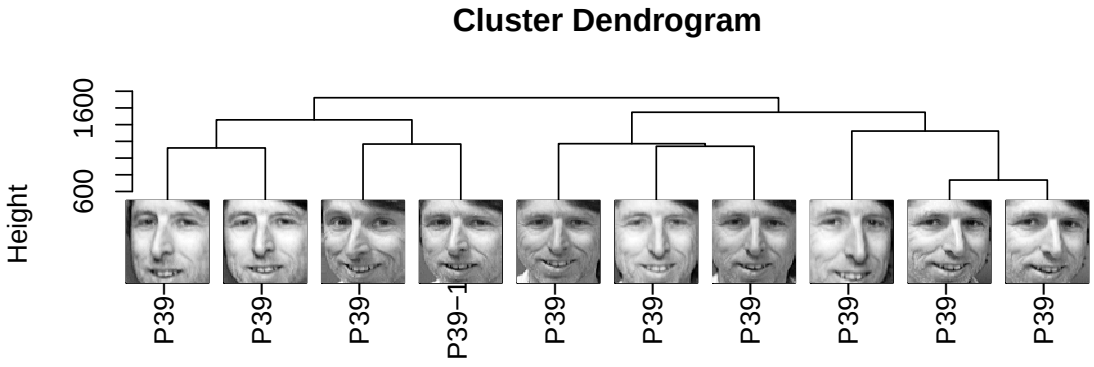
我们以第39个人的脸图为例,

```
xs<-x[381:390,]
xd<-dist(xs)
par(mfrow=c(2,5),mar=rep(0,4))
for(i in 1:10)
showMatrix(matrix(xs[i,],64,64))
```

使用层次聚类方法和Bien and Tibshirani (2011)的方法对其进行聚类

```
par(mfrow=c(2,1))
#Hierachical clustering
hh<-hclust(xd,method="average")
plot(hh,hang=-1)
for(i in 1:nrow(xs))
subplot(showMatrix(matrix(xs[hh$order[i],],64,64)),i,70,size=c(0.5,0.5))
#perform minimax linkage clustering by Bien, J., and Tibshirani, R. (2011).
library(protoclust)
library(TeachingDemos)
hm<-protoclust(xd)
plot(hm,hang=-1)
for(i in 1:nrow(xs))
subplot(showMatrix(matrix(xs[hm$order[i],],64,64)),i,70,size=c(0.5,0.5))
```





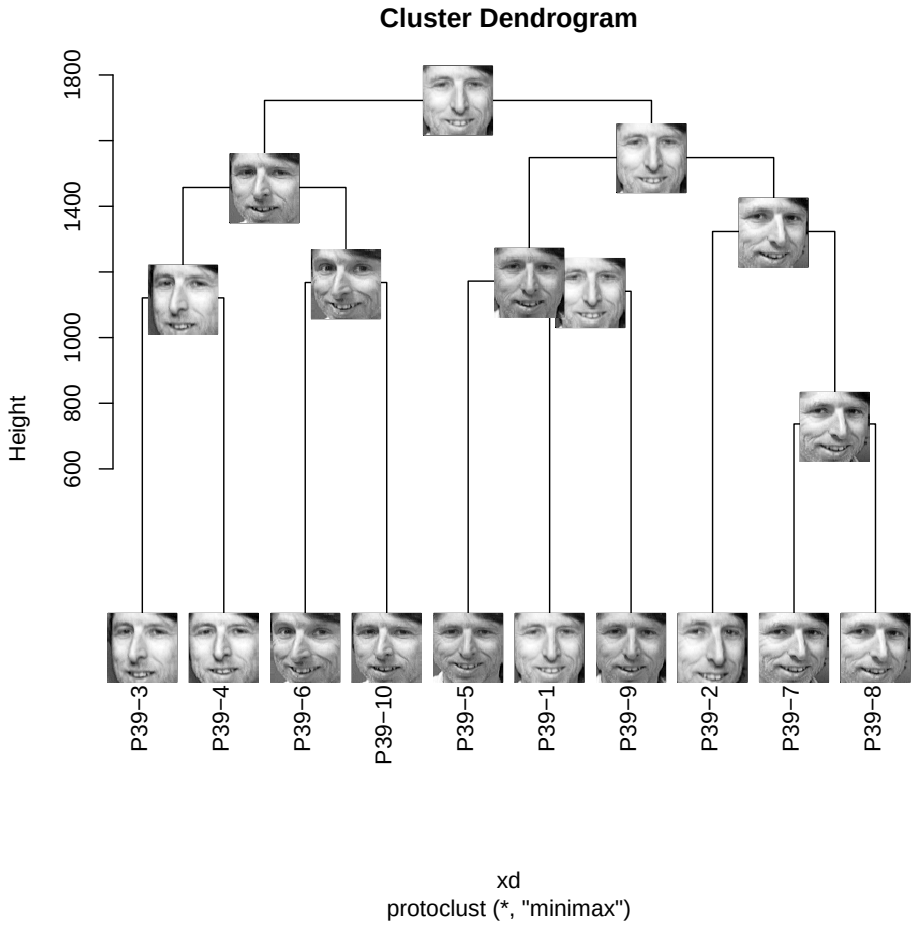
可以看出, 两种聚类方法都很好的将这10张脸图进行了聚类, 每个类表示了不同的鼻子方向.

在Bien and Tibshirani(2011)的minimax linkage方法中, 类的中心通过一个代表元表示, 因此, 层次聚类过程中每步合并得到的类的中心可以在树状图上表示出来. 下面我们找出聚类过程中每步的中心脸图的坐标, 据此将代表元在树状图上表示出来. 实现Bien and Tibshirani(2011)的图5.

```
#compute the coordinates of nodes
absi <- function(hc, level=length(hc$height), init=TRUE){
  if(init){
    .count <<- 0
    .topAbsis <<- NULL
    .heights <<- NULL
  }
  if(level<0) {
    .count <<- .count + 1
    return(.count)
  }
  node <- hc$merge[level,]
  le <- absi(hc, node[1], init=FALSE)
  ri <- absi(hc, node[2], init=FALSE)
  mid <- (le+ri)/2
  .topAbsis <<-  c(.topAbsis, mid)
  .heights <<- c(.heights, hc$height[level])
```

```
    invisible(mid)
}
absi(hm)
ord<-order(.heights)
yo<-.heights[ord]
xo<-.topAbsis[ord]
for(i in 1:(nrow(xs)-1))
subplot(showMatrix(matrix(xs[hm$proto[i],],64,64)),xo[i],yo[i],size=c(0.5,0.5))
```

最终得到



初步判别与分类分析例

由于数据为400 × 4096维矩阵, 变量个数远远大于样本量, 这里我们先采用主成分降维, 然后使用线性判别分析对数据进行判别与分类. 以每个人最后一张图为检验脸图,其他图为训练脸图.

```
id<-rep(1:40,each=10)
faces.data.frame<-data.frame(cbind(id=id,faces))
#pick out the last photo for each person as the testing photo
testid<-seq(10,400,by=10)
train.faces<-faces.data.frame[-testid,]
test.faces<-faces.data.frame[testid,]
```

计算主成分方向, 选择主成分个数以及计算主成分得分.

```
xc<-scale(train.faces[,-1],scale=FALSE)
A<-t(xc)/sqrt(360-1)
# 4096x360
# thus, the covariance matrix is A%*%t(A)
A.egn<-eigen(t(A)%*%A)
# 360x360
pc<-A%*%A.egn$vectors
pc<-apply(pc,2,function(i) i/sqrt(sum(i*i)))
# normalize the pc
n<-80
sum(A.egn$value[1:n])/sum(A.egn$value)
# 92%, thus we use the first  80 pcs
pcs<-pc[,1:n]
# pc scores for training data
yt<-xc%*%pcs
```

在主成分空间使用线性判别分析训练分类器, 并对训练样本进行判别. 最后发现我们非常完美的对训练样本进行了判别.

```
ft<-data.frame(cbind(id[-testid],yt))
flda<-lda(X1~.,ft)
fl<-predict(flda,ft)$class
diag(table(fl,id[-testid]))
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
9  9  9  9  9  9  9  9  9  9  9  9  9  9  9 9  9  9  9  9  9  9  9  9  9
 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
 9  9  9  9  9  9  9  9  9  9  9  9  9  9  9
```

对检验集,计算其在主成分空间下的得分, 据此使用训练好的分类器进行分类, 最后发现我们非常完美的对检验样本进行了分类.

```
#pc scores for testing data
xv<-as.matrix(test.faces[,-1])
xvs<-scale(xv,scale=FALSE)
yv<-xvs%*%pcs
fv<-data.frame(cbind(id[testid],yv))
```

```
pl<-predict(flda,fv)$class
diag(table(pl,id[testid]))
```

参考文献

[1] Bien, J., and Tibshirani, R. (2011), Hierarchical Clustering with Prototypes via Minimax Linkage, The Journal of the American Statistical Association

[2] Verma, T.; Sahu, R.K., (2013) PCA-LDA based face recognition system & results comparison by various classification techniques, Green High Performance Computing (ICGHPC), 2013 IEEE International Conference on , vol., no., pp.1,7, 14-15

[3] Calandriello D., Niu G., Sugiyama M. (2013), Semi-Supervised Information-Maximization Clustering