Mnohorozměrná statistika

14.12.2012

Úvodní nastavení.

– Otevřete si R Studio. Z internetu si stáhněte data decathlon.txt. Jedná se o výsledky olympijského závodu v desetiboji z roku 2004.

Vícerozměrná data

1. Načtěte si data decathlon.txt. Dále si nastavíme jména řádků jako názvy ve sloupci name:

```
row.names(decathlon)=decathlon$name

# prohlidka dat:
head(decathlon,5)
str(decathlon)
summary(decathlon)
```

2. Budou nás zajímat proměnné udávající výsledky v jednotlivých kategoriích.

```
discipliny=data[,2:11]

plot(discipliny)
pairs(discipliny[,1:5])

library(car)
scatterplotMatrix(discipliny)
```

Číselný popis vztahů jednotlivých veličin:

```
cor(discipliny)
round(cor(discipliny)*10)

cov(discipliny)

#jak souvisi jednotlive disciplinny s celkovým počtem bodů
cor(discipliny, decathlon$Points)
```

Mnohorozměrná vizualizace – tzv. Chernoffovy tváře

```
install.packages("aplpack")
library(aplpack)
faces(discipliny)

faces(discipliny,face.type=2)
# pro vysvetleni ?faces

stars(discipliny)
```

Analýza hlavních komponent

3. Provedeme analýzu hlavních komponent na deseti proměnných, které udávají výsledky v jednotlivých kategoriích. Cílem analýzy je získat menší počet proměnných, které nám vysvětlí většinu variability.

   (a) Provedení výpočtu:

   ```
   p=princomp(discipliny , cor= TRUE) # volba cor - pocitame s normovanymi daty

   summary(p)
   plot(p,type="l")
   ```

   (b) Interpretace hlavních komponent

   ```
   loadings(p)
   # nebo mozna prehledneji
   round(loadings(p)*10)

   # graficky
   biplot(p,choices=1:2,cex=0.8)

   # jen pro promenne
   biplot(p1,choices=1:2,col=c("white","red"))
   biplot(p1,choices=2:3,col=c("white","red"))
   biplot(p1,choices=c(1,3),col=c("white","red"))
   ```

   (c) Jednotlivé hlavní komponenty (nové proměnné)

   ```
   pc1=p$scores[,1]
   pc2=p$scores[,2]
   pc3=p$scores[,3]
   # ty lze dale pouzit napr. v regresi atd.

   #  znazorneni  zavodniku:
   plot(pc1,pc2,type="n")
   text(pc1,pc2,row.names(discipliny),cex=0.8)
   # podobne podle pc1,pc3 a pc2,pc3
   ```

   Lze identifikovat nějaké odlehlé (netypické) závodníky?

   (d) Můžeme ověřit, že jednotlivé hlavní komponenty jsou nekorelované (kolmé):

   ```
   cor(pc1,pc2)
   cor(pc1,pc3)
   cor(pc2,pc3)
   ```

   (e) Můžeme se ptát, jak souvisí hlavní komponenty s celkovými body:

   ```
   cor(pc1,decathlon$Points)
   cor(pc2,decathlon$Points)
   ```

Faktorová analýza

4. Stejná data budeme analyzovat pomocí faktorové analýzy. Budeme uvažovat tři faktory.
   (a) Provedení odhadu:

```
f = factanal(discipliny, 3, rotation="varimax")

print(f, digits=2, cutoff=.3)
```

   (b) Interpretace faktorů viz výstup výše nebo graficky

```
f1 <- f$loadings[,1]
f2 <- f$loadings[,2]
f3 <- f$loadings[,3]

plot(f1,f2,type="n",xlim=c(-1,1))
text(f1,f2,names(discipliny),cex=.8)

# podobne pro f1,f3 atd
```

   Jakou interpretaci bychom navrhli nalezeným třem faktorům?

   (c) Podobně jako v analýze hlavních komponent si můžeme znázornit jednotlivé závodníky podle faktorů

```
f = factanal(discipliny, 3, rotation="varimax",scores="regression")

fs1=f$scores[,1]
fs2=f$scores[,2]
fs3=f$scores[,3]

plot(fs1,fs2,type="n",xlim=c(-2.2,2.8))
text(fs1,fs2,row.names(discipliny),cex=.8)
#podobne pro fs1,fs3 a fs2,fs3
```

Samostatná práce

5. Uvažujte data USArrests (načteme příkazem data(USArrests)). Data obsahují statistiky o počtech zatčených na 100 000 obyvatel pro 50 států USA. Více viz ?USArresrs.
   (a) Prohlédněte si data. Podívejte se na vztah jednotlivých proměnných.
   (b) Proveďte analýzu hlavních komponent.
   - Kolik komponent zvolíte?
   - Navrhněte možnou interpretaci hlavních komponent.
   - Jsou některé státy z pohledu trestných činů jiné než ostatní (odlehlé)?
   - Které státy jsou bezpečné a které naopak?

Shluková analýza

6. Budeme analyzovat data z desetiboje pomocí shlukové analýzy. Zajímá nás, zda lze závodníky rozdělit do několika skupin tak, aby v těchto skupinách byli závodníci nějakým způsobem podobní (podle toho, které disciplíny jim více vyhovují apod.).

   (a) Pro další analýzu je vhodné data normovat:

   ```
   disc=scale(discipliny)
   ```

   (b) Nejprve použijeme hierarchický přístup ke shlukové analýze.

   ```
   d = dist(disc, method = "euclidean")  # spocte vzdalenosti

   #provede shlukovou analyzu dle vzdalenosti a Wardovou metodou:
   shl = hclust(d, method="ward")

   plot(shl)  # tzv. dendrogram

   rect.hclust(shl, k=4, border="red")    #napr. 4 skupiny

   groups = cutree(shl, k=4)
   groups
   sort(groups)
   ```

   Jak lze charakterizovat jednotlivé shluky? Lze nalézt nějakou souvislost s celkovým pořadím na olympiádě?

   (c) Proveďte totéž pomocí jiné volby vzdálenosti mezi shluky: změna method na 'single' nebo 'complete' nebo 'centroid' atd. Výsledky porovnejte.

   (d) Vyzkoušíme také metodu k-means. Řekněme, že bychom chtěli závodníky roztřídit do čtyř skupin (shluků).

   ```
   fit = kmeans(disc, 4)
   print(fit, digits=3)

   #skupiny:
   sort(fir$cluster)

   #skupiny graficky:
   pairs(discipliny,col=fit$clust)

   # nebo podrobneji napr. pro sprint 100m a disk
   plot(discipliny$X100m,discipliny$Discus,type="n")
   text(discipliny$X100m,discipliny$Discus,row.names(discipliny),col=fit$cluster)
   ```

Samostatná práce

6. Aplikujte metodu shlukové analýzy na data o kriminalitě v USA.

Diskriminační analýza

7. Uvažujte data iris, která obsahují informace o třech druzích kosatců. Na každé květině byla provedena čtyři měření: šířka a délka kališních lístků a okvětních lístků. Otázkou je, zda je možné na základě informace o těchto čtyřech veličinách zařadit květinu do příslušné třídy.

   (a) Nejprve si data načteme a prohlédneme

   ```
   data(iris)

   summary(iris)
   pairs(iris[,1:4], col=iris[,5])

   par(mfrow=c(2,2))
   for(i in 1:4)
   boxplot(iris[,i]~iris[,5],main=names(iris)[i])
   ```

   (b) Na základě všech dat vytvoříme diskriminační pravidlo:

   ```
   library(MASS)

   da<- lda(Species ~Sepal.Length + Sepal.Width + Petal.Length+
               Petal.Width, data=iris)
   # nebo zkracene: da= lda(Species ~., data=iris)
   print(da)
   ```

   (c) Zhodnocení přesnosti klasifikace (hodnoceno na trénovacích datech)

   ```
   pred=predict(da)

   table(iris$Species, pred$class)
   prop.table(table(iris$Species, pred$class),1)
   prop.table(table(iris$Species, pred$class))
   ```

   (d) Obrázek:

   ```
   plot(da,col=as.numeric(iris$Species))

   plot(da,dimen=1)
   ```

   (e) Nyní na základě modelu budeme klasifikovat nové pozorování s hodnotami 6,2,1.7,0.8

   ```
   new=data.frame(Sepal.Length=6, Sepal.Width=2,Petal.Length=1.7,Petal.Width=0.8)
   predLDA <- predict(fit,newdata=new)
   predLDA
   ```

Samostatná práce

8. Uvažujte data crabs. Cílem je vytvořit diskriminační pravidlo, pomocí něhož lze určit pohlaví kraba pouze na základě morfologických měření (FL, RW, CL, CW). Zjistěte, jak je taková diskriminace úspěšná.

   ```
   data(crabs)
   ?crabs
   ```