Clustering

Statistics 4868/6610 Data Visualization

Prof. Eric A. Suess

3/2/2016

Introduction

Today Chapter 7.

- Parallel Coordinate Plots
- Dimension Reduction - Multidimensional Scaling
- Cluster Analysis
- Outliers

The idea of Clustered data

"Although Chernoff Faces and star charts can make it easier to spot observations (units or rows) that are different from the rest of the observations in a dataset, it is a challenge to identify groups or how variables could be related."

The author of the book does a good job describing the main idea behind clustering methods.

Today we are going to discuss methods to group obsevations in datasets.

- Find similar basketball players.
- Find similar cities.
- Find similar cars.

Parallel Coordinate Plots

Parallel Coordinate Plots plot data as described.

Each variable is listed with a parallel bar and each observation's values for the variables are connected. Patterns can be seen. The application of color to similar observations make the plot appear to show the similarities better.

The lattice package in R has the function **parallel** and there is another R package **ggparallel** that makes parallel coordinate plots more colorful.

Excellent Visualization blog

- eagereyes
- Parallel Coordinates
- Parallel Sets

See the ggparallel package for similar plots.

Parallel Coordinate Plots

R Examples:

- Mastering Parallel Coordinate Plots
- Quick-R Interactive
- R-blogger Parallel Coordinates
- R-statistics Clustergram

Parallel Coordinates Plot

D3 Examples:

- D3 Parallel Coordinates See the video at the bottom of this page. Cool! Kai Chang- Visualizing
- D3 Parallel Coordinates
- D3 Parallel Coordinates
- Nutrient Content - Parallel Coordinates
- Titanic - Parallel Coordinates

Reducing Dimensions

The author talks about

- Distance
- Scaling
- Clustering

Reducing Dimensions

You know about Reducing Dimension.

You know about computing Descriptive Statistics, such as the sample mean and sample standard deviation. You know about stepwise linear regression.

You know about comparing groups of data, two-sample t-tests and ANOVA.

Note that the groups are known in these cases.

The groups are really groups of **rows** in your dataset.

Reducing Dimensions

In regression you select the imporant explanatory variables that are correlated with your response variable. The explanatory variables that are not statistically significant are dropped. These are variables that have low correlation with the response variable.

Note that there are two main groups here, they are the predictor variables included in the model and those that are not included.

Note that the groups are not known before the analysis.

Reducing Dimensions

What does stepwise regression do? Reduce the dimension of the data.

Also, . . .

You also may know about sufficient statistics.

Reducing Dimensions

So there are three steps to visualize clusters of data in smaller dimension.

Example: The education data has 6 dimension: reading, math, etc. and states.

Step 1: To plot the state names on an x-y plot using all of the variables and grouped by similarity, measured by **distance**, the author introduced the **dist()** function in R.

Reducing Dimensions

Step 2: To scale the distance matrix so that it can be plotted on an x-y graphs, Multidimensional Scaling (MDS) is introduced. The **cmdscale()** fuction is used.

This is like projecting 5-dimensions onto 2-dimensions.

And the state names are put on the plot.

Reducing Dimensions

Step 3: To identify the clusters in the data model-based clustering is mentioned. The **mclust** library is called and the function **Mclust()** is used.

Here the clusting is done after the MDS. There are many kinds of clustering methods and clustering can be preformed on datasets with many variables.

The main question when clustering is how many clusters are there in the data?

Reducing Dimensions

In this example, the author uses . . .

MDS is used to reduce the number of **columns** in the dataset.

Clustering is used to reduce the numbers of **rows** in the data.

Reducing Dimensions

To learn more about Reducing Dimensions in data, see the Quick-R website.

See Advanced Statistics

See Principle Components and read about Facor Analysis

See Multidimensional Scaling

See Tree Based Methods and read about CART, MARS, Random Forests

R Machine Learning

Reducing Dimensions

Small correction to the author's code

```r
# Clustering
library(mclust)
ed.mclust <- Mclust(ed.mds)
par(mfrow=c(2,2))
plot(ed.mclust)    # remove the data=ed.mds
```

Outliers

Use boxplots!!!