Técnicas Computacionais em Probabilidade e Estatística I

Aula VIII

Chang Chiann
MAE 5704- IME/USP
1º Sem/2008

Análise de vários conjuntos de dados

1. Introdução:

Muito frequentemente temos várias variáveis associadas aos elementos de uma amostra.

Ex.: veículos: preço, marca, potência do motor, comprimento.

De um modo geral, temos p variáveis aleatórias

$(X_1, X_2, ..., X_p) \rightarrow P(X_1=x_1, X_2=x_2, ..., X_p=x_p)$: caso discreto

$\rightarrow f(x_1, x_2, ..., x_p)$: caso contínuo

Em algumas situações, podemos ter amostras de tamanhos diferentes para as variáveis de interesse.

No caso de mesmo número de observações,

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{np} \end{bmatrix}$$

n: no. de observações de cada variável;

p: no. de variáveis.

Em geral, temos um número grande de dados, np no total. Assim, serão muitos parâmetros a serem estimados: médias, variâncias, correlações, etc,...

As variáveis são usualmente correlacionadas, entre si, ou seja, embora os dados em cada coluna possam ser eventualmente independentes, as colunas, em geral, são correlacionadas.

Mas mesmo em cada coluna, os dados posem apresentar dependência: séries temporais.

Redução de número de variáveis → análise de componentes principais.

Redução de número de variáveis mantendo a maior parte da informação presente nos dados originais.

O recurso de representar graficamente variáveis e relações entre variáveis torna-se mais complicado quando temos mais de duas variáveis.

2. Gráficos com três variáveis

(i) "Draftsman's Display"

Nesta situação, fazemos

$$\binom{3}{2} = 3$$

Gráficos de dispersão bivariados.

$\binom{3}{2} = 3$

(ii) Gráficos de Dispersão Simbólicos

É útil quando temos duas variáveis quantitativas e uma qualitativa.

Assim para cada (x,y) utilizamos um símbolo para a variável qualitativa.

2. Gráficos com três variáveis

(iii) Partição e janeramento

Uma abordagem alternativa é:

a) Dividir as n observações disponiveis em subconjuntos de acordo com os valores de uma variável

b) Fazer um gráfico de dispersão para as outras duas variáveis, para cada subconjunto.

3. Gráficos com 4 ou mais variáveis