SCC0173 – Mineração de Dados Biológicos

Análise Exploratória de Dados – Parte A:
Revisão de Estatística Descritiva Elementar

Prof. Ricardo J. G. B. Campello

SCC / ICMC / USP

Tópicos

- Análise Exploratória de Dados
- Estatísticas Descritivas
  - Dados univariados
    - Medidas de centralidade
    - Medidas de dispersão
  - Dados multivariados
    - Covariância
    - Correlação

Introdução

- Exploração preliminar e visualização dos dados facilita entendimento de suas características
- Principais motivações:
  - Pode ajudar na seleção da melhor técnica de pré-processamento e/ou mineração
  - Pode fazer uso da capacidade humana de reconhecer visualmente padrões
    - Muitas vezes difíceis de serem detectados automaticamente

Análise Exploratória de Dados

- **Exploratory Data Analysis** (EDA)
  - Área criada pelo estatístico John Tukey
  - Focada em Estatística e Visualização
  - Pode dar importante suporte a DM

Estatísticas Descritivas

- Descrevem os dados
- Quantidades que resumem características de um conjunto de dados, geralmente grande
  - Na maioria das vezes podem ser calculadas com uma simples passagem pelos dados
  - Exemplos:
    - Renda média dos alunos de uma turma
    - Porcentagem de alunos que se formam em 4 anos

Estatísticas Descritivas

- Assumem que os dados são gerados por um processo aleatório
  - Caracterizado por vários parâmetros
  - Podem ser vistas como estimativas dos parâmetros do processo que gerou os dados
    - Ex. Distribuição normal com média 0 e variância 1

Estatísticas Descritivas

- Podem capturar:
  - Frequência
  - Localização ou tendência central
    - Ex. Média
  - Dispersão ou espalhamento
    - Ex. Desvio padrão
  - Distribuição ou formato

Frequência

- Proporção de vezes que um atributo assume um dado valor
  - Para um determinado conjunto de dados
  - Muita usada para dados categóricos
  - Exemplo:
    - Em um conj. de dados médicos, 40% dos pacientes têm febre

Exemplo

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

66% das manchas são manchas grandes

Exemplo

- Seja o seguinte estudo:
  - Em uma pesquisa de opinião, 280 alunos de foram consultados a respeito de suas opiniões sobre o desempenho do professor de uma dada disciplina

Exemplo

- Tabela: Frequências observadas e freqs. relativas para cada categoria de resposta
  - Bom, Regular, Péssimo

Resposta	Freq.	Freq. Rel.
Bom	152	152/280 = 0,543
Regular	77	77/280 = 0,275
Péssimo	51	51/280 = 0,182
Total	280	280/280 = 1,000

Exemplo

- Gráfico: Frequências Relativas podem ser vistas no diagrama circular:

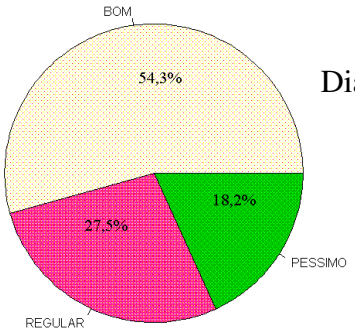
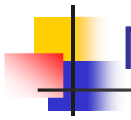
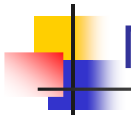


Diagrama de Torta



Medidas de Tendência Central

- Dados Categóricos
  - Moda
- Dados Numéricos
  - Média
  - Mediana
  - Percentil



Moda

- Valor mais frequente para o atributo nos dados
- Exemplo:

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

Moda para o atributo mancha: grande

Média e Mediana

- Medidas mais utilizadas para dados numéricos
  - Tendência central de um conjunto de pontos
- Considere um conjunto de N objetos e um atributo x
  - Seja $\{x_1, ..., x_N\}$ o valor do atributo para os N objetos

Média

- Pode ser calculada facilmente

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Problema: sensível a *outliers*

Mediana

- Valor que divide valores menores e maiores em quantidades iguais
- Como calcular:
  - Ordenar valores de *x*
  - Se N é impar, mediana = valor com ordem central
  - Senão, mediana = média dos dois valores centrais

Média e Mediana

- Média é um bom indicador do meio de um conj. de valores apenas se os valores estão distribuídos simetricamente
- Mediana indica melhor o meio
  - Se distribuição é oblíqua (assimétrica)
  - Se existem *outliers*
- Mas perde sentido de centro de área / massa

Média Podada

- *Trimmed Mean*
- Minimiza problema da média descartando exemplos extremos
  - Define porcentagem p dos exemplos a serem eliminados
  - Ordena os dados
  - Elimina (p/2)% dos exemplos em cada extremidade

Exercício

- Dado o conjunto {1, 2, 3, 4, 5, 80}, calcular:
  - Média
  - Mediana
  - Média podada com p = 33%

Quartis e Percentis

- Mediana divide os dados ao meio
- Outras medidas usam pontos de divisão diferentes
  - Quartis dividem um conj. ordenado de dados em quartos
    - 1º quartil, Q_1, é o valor da amostra que tem 25% das observações abaixo de seu valor
    - Segundo quartil é a mediana

Percentil

- Seja x um atributo numérico ou ordinal e p um valor entre 0 e 100
  - O p^o percentil é um valor x_i do conjunto de valores de x tal que p% dos valores no conj. de dados são menores que x_i
  - Exemplos:
    - 40^o percentil do atributo x é o valor $x_{40\%}$ tal que 40% dos valores de x são menores que $x_{40\%}$
    - 25º percentil = 1º quartil, 50º percentil = mediana

Exemplo

- Obter os quartis e o 95º percentil para o conjunto de dados abaixo:

6.2 7.67 8.3 9.0 9.4 9.8 10.5 10.7 11.0 12.3

Achar Q_1: 10x1/4 = 2.5
usar o terceiro valor: Q_1 = 8.3
Achar Q_2: 10x1/2 = 5
usar a média entre o 5º e o 6º valores: Q_2 = (9.4 + 9.8)/2 = 9.6
Achar Q_3: 10x3/4 = 7.5
usar o oitavo valor: Q_3 = 10.7
Achar $P_{0.95}$: 10x0.95 = 9.5
usar o décimo valor: $P_{0.95}$ = 12.3

Percentis

Seja N o número de observações, calcular o pº percentil:

1. *Ordenar as observações da menor para a maior*
2. *Determinar o produto N × p e chamar este produto de k*
3. *Se k não for um inteiro*
   Então Arredondar k para o próximo inteiro
   Retornar o valor da posição k na sequência ordenada
   Senão
   Calcular a média entre as kº e (k+1)º observações ordenadas
   Retornar o valor calculado

Exercício

Dados os números abaixo, calcular a mediana, o 1º quartil e o 2º quartil

23, 7, 12, 6, 10, 23, 7, 12, 6, 10, 7

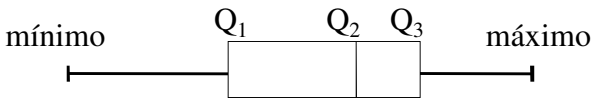
Exercício

- Obter os quartis e a 95ª percentil para o conjunto de dados:

3,20	11,70	13,64	15,60	15,89	28,44	29,07
37,34	41,81	43,35	43,94	49,51	49,82	51,20
51,43	52,47	53,72	53,92	54,03	56,89	63,80
66,40	68,64	70,15	70,98	74,52	76,68	77,84
80,91	84,04	85,70	86,48	88,92	89,28	91,36
91,62	98,79	102,39	104,21	124,27		

Boxplot

- Um resumo das informações dos quartis é apresentado em um gráfico chamado **boxplot**



- Um boxplot modificado
  - Linha exterior vai até a maior (menor) observação apenas se não for muito distante do 3º (1º) quartil

Medidas de Espalhamento

- Medem dispersão ou espalhamento de um conjunto de valores
- Indicam se os dados estão
  - Amplamente espalhados ou
  - Relativamente concentrados em torno de um ponto
- Medidas comuns
  - Intervalo
  - Variância
  - Desvio padrão

Intervalo

- Medida mais simples
  - mostra espalhamento máximo
- Sejam $\{x_1, ..., x_N\}$ os valores do atributo x para N objetos. Então:

$$r(x) = \max(x) - \min(x)$$

- Pode não ser uma boa medida...
  - P. ex. se maioria dos valores forem concentrados, com um pequeno número de valores extremos

Variância

- Medida preferida para analisar espalhamento

$$\text{var}(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

  - Denominador *N-1*: correção de Bessel, usada para uma melhor estimativa da variância verdadeira
  - Desvio padrão σ_x: raiz quadrada da variância

Variância

- Assim como a média, a variância pode ser distorcida por *outliers*
  - quadrado da diferença entre os valores e a média...
- Estimativas mais robustas também usadas:
  - Desvio médio absoluto
    - *Absolute Average Deviation* – AAD
  - Desvio mediano absoluto
    - *Median Absolute Deviation* – MAD
  - Intervalo interquartil
    - *Interquartil Range* – IQR

Medidas de Espalhamento

$$AAD(x) = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \bar{x}\right|$$

$$MAD(x) = mediana\left(\left\{\left|x_1 - \bar{x}\right|,\ldots,\left|x_N - \bar{x}\right|\right\}\right)$$

$$IQR(x) = x_{75\%} - x_{25\%}$$

Exercício

- Dados os valores {1, 2, 3, 4, 5, 80}, calcular:
  - Intervalo
  - Variância
  - AAD
  - MAD
  - IQR

Dados Multivariados

- Aqueles que possuem vários atributos
- Medidas de tendência central
  - Podem ser obtidas calculando medida de cada atributo separadamente
    - Ex.: média, mediana, ...
  - Média dos objetos de um conjunto de dados com n atributos x_1, ..., x_n é dada por:

$$\overline{\mathbf{x}} = [\overline{x}_1 \ldots \overline{x}_n]$$

Dados Multivariados

- Medidas de espalhamento
  - Podem ser calculadas para cada atributo independentemente dos demais
    - Usando qualquer medida de espelhamento
  - Variáveis numéricas
    - Espalhamento de um conjunto de dados é melhor capturado por uma **matriz de covariância**
      - Cada elemento é a covariância entre dois atributos

Dados Multivariados

- Matriz de covariância S para um conjunto de dados com N objetos e n atributos x_1, ..., x_n

$$s_{ij} = \mathrm{cov}(x_i, x_j)$$

$$\mathrm{cov}(x_i, x_j) = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

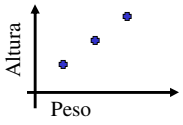
onde:
$\bar{x}_i$: valor médio do i-ésimo atributo
x_{ki}: valor do i-ésimo atributo para o k-ésimo objeto

  - Note que cov(x_i , x_i) = variância (x_i) !
    - Valores na diagonal da matriz = variância dos atributos !

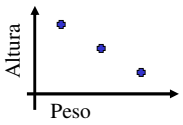
Exercício

- Calcular as matrizes de covariância para os seguintes dados de três pessoas:

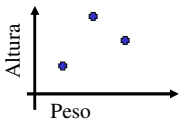
Peso	Altura
60	170
70	180
80	190



Peso	Altura
60	190
70	180
80	170



Peso	Altura
60	170
70	190
80	180



Exercício

- Calcular a matriz de covariância para o conjunto de dados:

Peso	altura	temperatura
73,2	170	37,5
67,5	165	38
90	190	37,2
49	152	37,8

Dados Multivariados

- Covariância de dois atributos
  - Mede grau com que os atributos variam juntos
    - Depende da magnitude dos atributos
    - Valor próximo de 0:
      - Atributos não têm um relacionamento linear
    - Valor positivo:
      - Atributos diretamente relacionados
      - Quando o valor de um atributo aumenta, o do outro também aumenta

Dados Multivariados

- Covariância x Correlação
  - É difícil avaliar a **força** do relacionamento entre dois atributos olhando apenas a covariância
    - valor depende dos espalhamentos de cada atributo
  - Correlação de Pearson é mais apropriada para medir a força da relação linear entre atributos
    - covariância normalizada pelos desvios padrão

Dados Multivariados

- Correlação
  - Indica força da relação entre dois atributos
  - Matriz de correlação R

$$r_{ij} = \mathrm{corr}(x_i, x_j) = \frac{\mathrm{cov}(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}}$$

x_i: i-ésimo atributo
σ_{xi}: Variância do atributo x_i

  - Note que corr(x_i, x_i) = 1 (elementos da diagonal)
    - corr(x_i,x_j) $\in$ [–1,+1]

Exercício

- Calcular a matriz de correlação para o conjunto de dados:

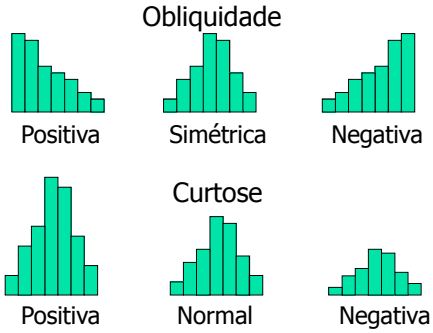
Peso,	altura,	temperatura
73,2	170	37,5
67,5	165	38
90	190	37,2
49	152	37,8

Outras Estatísticas...

- Outros momentos além de média e variância:
  - Obliquidade / Skewness (3º momento central)
    - captura simetria da distribuição dos dados
  - Curtose (4º momento central)
    - Captura achatamento / pico da distribuição
  - ...

Histograma

- Poderosa ferramenta para verificar visualmente características dos dados



Visualização de Dados

- Em vários casos, a forma mais fácil de entender aspectos mais complicados dos dados é ver os seus valores graficamente
  - Por exemplo, histogramas
- Vide próxima aula...

Perguntas