

Preparação de Dados

Prof. Jefferson T. Oliva

Aprendizado de Máquina e Reconhecimento de Padrões (AM28CP)
Engenharia de Computação
Departamento Acadêmico de Informática (Dainf)
Universidade Tecnológica Federal do Paraná (UTFPR)
Campus Pato Branco

- Atributo
- Exploração de Dados

- Conjuntos de dados
 - Estruturados
 - Facilmente analisado por métodos de aprendizado de máquina
 - Exemplos: planilhas e tabelas atributo-valor
 - Não estruturados
 - Mais facilmente analisados por humanos, mas não por métodos de aprendizado de máquina
 - Exemplos: texto, imagens, áudio ...
 - Para viabilizar a aplicação de algoritmos de aprendizado, os dados não estruturados devem ser convertidos para estruturados

Introdução

- Exemplo de conjunto de dados estruturados

Atributos				Alvo/Classe
sepalwidth	sepalwidth	petalwidth	petalwidth	specie
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
...				
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
..				
7.1	3.0	5.9	2.1	Iris-virginica
6.3	2.9	5.6	1.8	Iris-virginica
6.5	3.0	5.8	2.2	Iris-virginica

Atributo

- Tipo de atributos
 - Nominal
 - Exemplos: cor, nome, profissão
 - Ordinal
 - Dias da semana, tamanho (pequeno, médio, grande)
 - Intervalar
 - Exemplos: data, temperatura
 - Racional/proporção
 - Exemplos: peso, salário

- Atributos também se distinguem pela quantidade de valores (quantitativos)
 - Discretos: o conjunto de valores possíveis é finito ou enumerável
 - Exemplos: idade, quantidade de alunos, tamanho da frota de uma organização
 - Contínuo: o conjunto de valores são contínuos, como números reais
 - Exemplos: peso, renda, temperatura
- Atributos ordinais e nominais são denominados atributos categóricos

Exploração de Dados

- Exploração preliminar dos dados facilita entendimento de suas características
- Ajuda na seleção de métodos adequados de pré-processamento e de aprendizado de máquina
- Principais recursos para exploração de dados
 - Estatística descritiva
 - Visualização gráfica

- Estatística descritiva tem o objetivo de de descrever dados
- Para isso, são calculados valores que caracterizam um conjunto de dados
- Essas medidas podem ser de
 - Frequência: proporção de vezes que um atributo assume determinado valor (por exemplo, em um determinado conjunto de dados, 70% dos egressos de engenharia de computação da UTFPR residem na região sul do Brasil)
 - Tendência central: média, mediana, percentil, moda
 - Dispersão ou espalhamento: desvio-padrão, variância, intervalo interquartil

- Média

- Por mais que possa ser facilmente calculada, é sensível a *outliers*
- É um bom indicador do centro de um conjunto de valores, desde que estejam simetricamente distribuídos

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

- Mediana
 - Menos sensível a *outliers*
 - Mas os dados devem ser ordenados
 - Se n for ímpar:

$$Median = x_{\frac{n+1}{2}}$$

- Se n for par:

$$Median = \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$$

- Quartis
 - Mediana representa o meio de um conjunto ordenado
 - Outras medidas utilizam pontos de divisão diferentes
 - Quartis dividem os dados ordenados e quartis
 - 1º quartil (Q1): valor para o qual 25% dos elementos são menores ou igual a ele
 - 2º quartil (Q2): mediana
 - 3º quartil (Q3): valor para o qual 75% dos elementos são menores ou igual a ele
- Percentis

- Medidas de espalhamento
 - Medem dispersão ou espalhamento de um conjunto de valores
 - Indicam se os dados estão amplamente distribuídos ou concentrados em torno de um ponto (e.g. média)
 - Exemplos de medidas: intervalo, variância, desvio-padrão

- Medidas de espalhamento

- intervalo

- Não é uma boa medida caso a maioria dos valores estejam concentrados próximos de um ponto e alguns valores estejam próximos aos extremos

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

- Variância

- Principal medidas utilizada na análise do espalhamento dos dados

$$\sigma(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- $n - 1$: correção de Bessel, para melhorar a estimativa da variância verdadeira
 - $\sqrt{\sigma}$: desvio-padrão

- Momento

- Utilizado para obter diversas medidas

$$\mu_k = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^k$$

- $k = 1$: primeiro momento em torno da origem (primeiro momento central)
- $k = 2$: variância (segundo momento central)
- $k = 3$: obliquidade (terceiro momento central)
- $k = 4$: curtose (quarto momento central)

- Obliquidade

- Determina a simetria da distribuição dos dados em torno da média

$$\mu_3 = \frac{1}{(n-1)\sigma^3} \sum_{i=1}^n (x_i - \bar{x})^3$$

- A divisão por σ^3 tem a finalidade de tornar a medida independente de escala

- Curtose

- Determina o achatamento da distribuição dos dados

$$\mu_4 = \frac{1}{(n-1)\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^4$$

- Conjunto de dados multivariados
 - Possuem mais que um atributo, onde cada é considerado uma variável
 - Medidas podem ser obtidas separadamente para cada atributo, como média, desvio-padrão, etc
 - Por exemplo, a média de um conjunto de dados pode ser representada na seguinte forma: $\bar{x} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_m\}$
 - Variáveis contínuas: o espalhamento pode ser (melhor) mensurado por uma matriz de covariância, onde o elemento r_{ij} representa a covariância entre os atributos i e j

- Conjunto de dados multivariados
 - Covariância

$$r_{ij} = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

- Caso $i = j$, o que é calculada é a variância do atributo x_i
- Dessa forma, a diagonal principal da matriz r contém as variâncias dos atributos
- A interpretação do relacionamento entre atributos utilizando covariância é considerada complexa, pois o valor é influenciado pela magnitude dos atributos
- Opcionalmente, pode ser utilizado coeficiente de correlação, que possui maior popularidade em comparação com a covariância

- Correlação linear
 - Covariância normalizada
 - Determina a força da relação entre dois atributos

$$s_{ij} = \frac{r_{ij}}{\sigma(x_i)\sigma(x_j)}$$

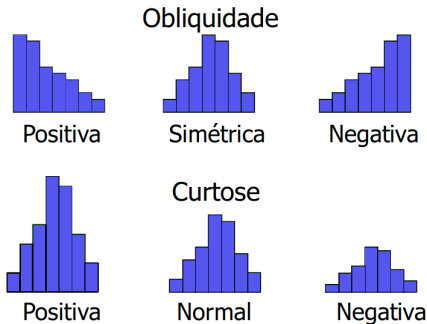
- Caso $i = j$, então $s_{ij} = 1$, ou seja, os elementos da diagonal principal matriz de correlação possui valor igual a 1
- Os demais elementos da matriz s possuem valores entre -1 e +1

Exploração de Dados

Visualização gráfica

- Uso de imagens para a compreensão de dados
- Simplifica a análise (por humanos) de grandes volumes de dados
- Diversos recursos podem ser utilizados, tais como:
 - Histograma
 - Boxplot
 - Gráfico de linha
 - Gráfico de pizza
 - Gráfico de dispersão
 - Mapa de calor

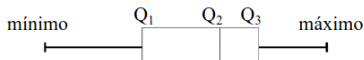
- Histograma
 - Comumente para verificar a distribuição de dados
 - Simplifica a verificação gráfica de medidas de curtose e de obliquidade



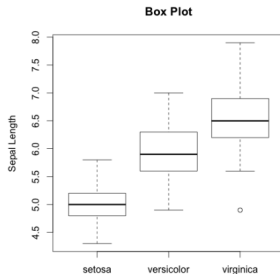
Exploração de Dados

Visualização gráfica

- Boxplot
 - Visualização gráfica dos quartis
 - Caso houver, os pontos antes do mínimo e depois do máximo são considerados *outliers*



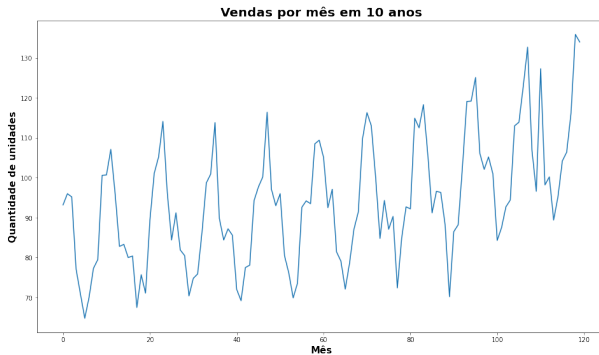
- Exemplo:



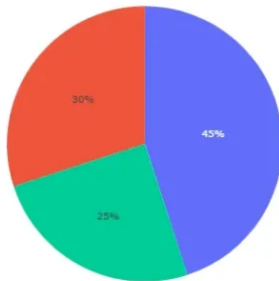
Exploração de Dados

Visualização gráfica

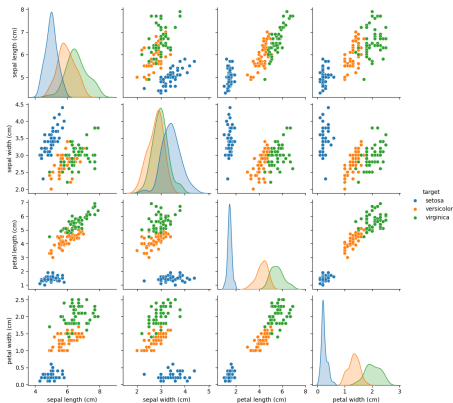
- Gráfico de linha
 - Comumente utilizado em dados sequenciais (e.g. séries temporais)
 - A ordem dos dados é importante



- Gráfico de pizza
 - Utilizada para verificação de proporções (frequências) para classes em um diagrama circular



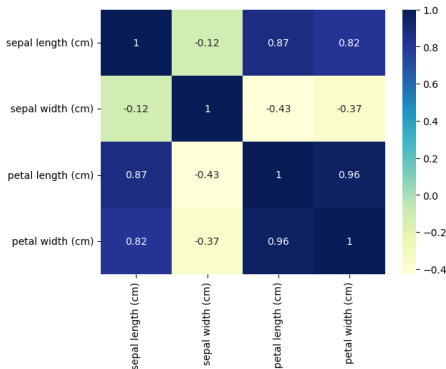
- Gráfico de dispersão (*scatter plot*)
 - Utilizado para ilustrar correlação entre pare de atributos



Exploração de Dados

Visualização gráfica

- Mapa de calor
 - Outro recurso comumente utilizado para verificar correlações entre atributos
 - Para isso, é necessária a matriz de correlações





BISHOP, C. M.

Pattern Recognition and Machine Learning.

Springer, 2006.



DE CARVALHO, A. P. L. F.

Preparação de Dados. Aprendizado de Máquina.

Slides. Ciência de Computação e Matemática Computacional.

ICMC/USP, 2015.



DUDA, Richard O.; HART, Peter E.; STORK, David G.

Pattern classification.

2nd ed. New York, NY: J. Wiley & Sons, 2001.



RASCHKA, S.; MIRJALILI, V.

Python Machine Learning.

Packt, 2017.