- Sumário:
  - Fundamentos de Aprendizagem Automática
    - Conceitos básicos
    - Preprocessamento dos dados

Objectivos

- Adquirir uma noção acerca dos conceitos básicos utilizados em AA
- Adquirir a noção de preprocessamento dos dados
- Descrever as principais técnicas de preprocessamento

Exemplos – problema do tempo

Determinação das condições favoráveis para a prática de um desporto não especificado

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

Exemplos – conjunto de dados *Iris*

- Determinação do tipo de planta a partir de um conjunto de características
- Derivado do trabalho do eminente estatístico R. A. Fisher (1890 – 1962)

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

Conceito

- A coisa que se pretende aprender
- A saída produzida pelo algoritmo de aprendizagem se designa como a descrição do conceito

Classe

- Possíveis valores que pode ter um conceito dado

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

Instancia

- É um objecto individual e independente, representante do conceito a ser aprendido
- Também designado como *exemplo*, *padrão, observação, mostra…*

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

Características

- Cada instancia é determinada pelos valores atribuídos a um conjunto fixo e predeterminado de *características* ou *atributos*
- Cada *vector de características* identifica de forma unívoca a uma instancia individual

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

Características

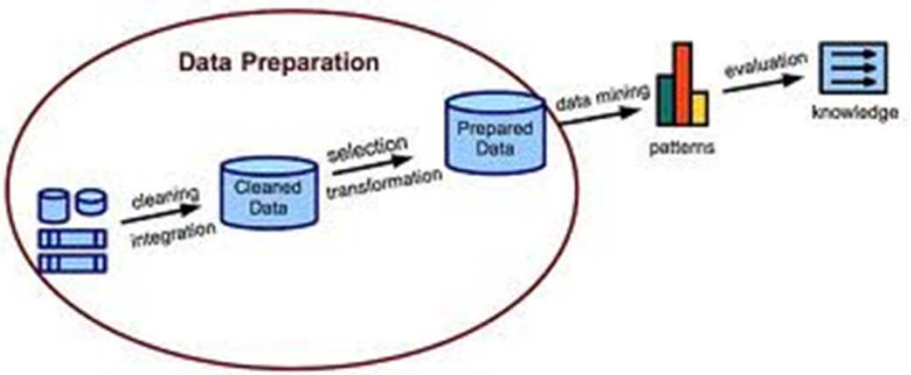
- A maioria dos sistemas de aprendizagem automática utilizam dois níveis de medida nas características: nominal e ordinal
- Os atributos nominais correspondem aos tipos de dados enumerativos utilizados nas linguagens de programação
  - Neste caso não existe implícita uma relação de ordem
- Os valores ordinais se codificam como dados numéricos
  - Tem como caso especial a escala dicotómica, que apresenta somente dois membros (booleana)

Ruído

- Erros não intencionados com carácter aleatório e esporádico existentes nos dados de entrada
- Devido a factores diversos
  - Erros nos dispositivos de recolha de dados
  - Erros cometidos ao etiquetar os dados
  - Atributos equivocados…

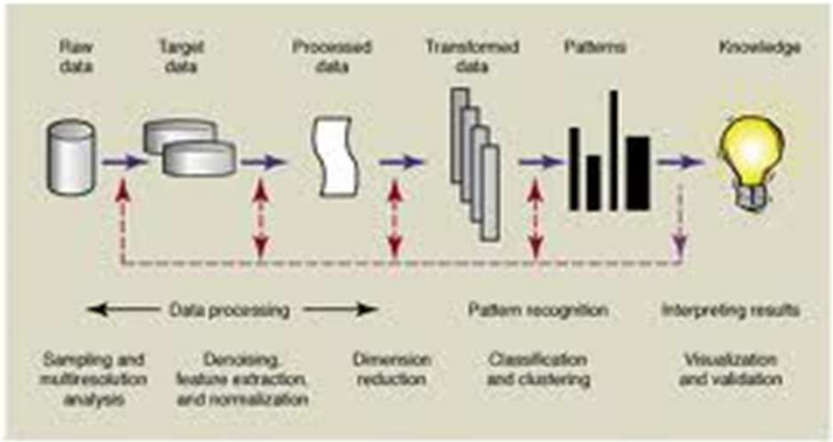
Preparação dos dados

- A preparação dos dados utilizados na solução de um problema é uma tarefa complexa e geralmente consome a maior parte do esforço investido no processo
- É uma tarefa que geralmente é dependente da aplicação concreta



Preprocessamento dos dados

- Em muitas circunstancias é necessário efectuar certas operações de processamento aos dados antes da sua utilização
- Entre essas se encontram
  - A remoção de *outliers*
  - A normalização dos dados
  - O tratamento de dados omissos



Sapphire, a data-mining infrastructure developed at Lawrence Livermore, is an iterative and interactive process designed to help scientific researchers uncover patterns in large data sets.

Outliers

- Um *outlier* é um ponto cujo valor se encontra muito distante do valor médio da correspondente variável aleatória
- A distancia é medida com relação a um limiar determinado, geralmente um múltiplo do desvio padrão

Remoção de *outliers*

- Se a quantidade de *outliers* é pequena, os mesmos são geralmente descartados
- Caso contrario é necessário empregar métodos de aprendizagem que sejam pouco sensíveis à sua presença

Normalização dos dados

- Em muitas circunstâncias os valores das características se encontram em diferentes intervalos
- Ao utilizar certos métodos isto pode provocar um predomínio da influência de umas características sobre outras nos resultados obtidos

Normalização dos dados

- Este problema pode ser resolvido normalizando os dados
- Uma técnica utilizada consiste em limitar os valores das características a um intervalo dado ([0, 1] ou [-1, 1]) através de uma transformação adequada

$$\hat{x}_{ik} = \frac{x_{ik} - x_{k\min}}{x_{k\max} - x_{k\min}}, \qquad \hat{x}_{ik} = \frac{2(x_{ik} - x_{k\min})}{(x_{k\max} - x_{k\min})} - 1, \quad k = 1,2,\ldots,l$$

Normalização dos dados

- Outra técnica muito utilizada consiste em aplicar uma transformação de forma tal que todas as características apresentem um valor médio igual a 0 e desvio padrão igual a 1
- Também designada como *padronização (standardization)*

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k^{\,2}}, \quad k = 1,2,\ldots,l$$

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}, \qquad \sigma_k^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$$

Dados omissos

- Na prática muitos conjuntos de dados possuem valores omissos
- Podem dever-se a diversas razões, falhas no equipamento de medição, alterações ao desenho experimental, junção de vários conjuntos de dados semelhantes mas não idênticos…

Dados omissos

- Possíveis soluções
  - Descartar os vectores de dados que apresentem valores omissos, se a quantidade de dados disponíveis o permitir
  - Preencher os casos de omissão com valores fora do rango
  - Completar os casos omissos com o valor médio calculado com os dados disponíveis para a correspondente característica…

Bibliografia

- Witten
  - Exemplos de conjuntos de dados, pg. 9 – 21
  - Conceitos, pg. 38 - 51
- Theodoridis
  - Preprocessamento, pg. 164 - 166