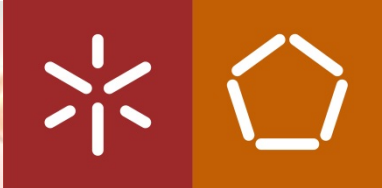




Universidade do Minho
Escola de Engenharia
Departamento de Informática

Classificação para Extração de Conhecimento

Cesar Analide, Paulo Novais, José Neves



Classificação para Extração de Conhecimento

O que é Classificação?

- Um sistema de Classificação **mapeia** cada **objeto** numa **classe**;
- Classificar significa **atribuir** uma **classe** a um **objeto**;
- As classes:
 - são discretas;
 - existem em número finito;
 - não têm ordem;
 - são identificadas por um nome (*label* ou *tag*).



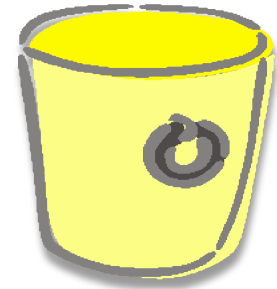
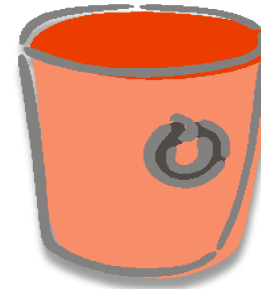
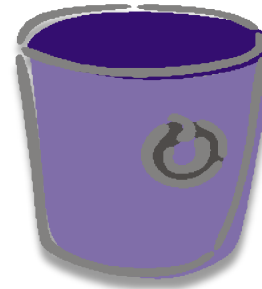
Classificação para Extração de Conhecimento

**O que é
Classificação?**

- O objetivo da Classificação é o de **organizar e categorizar dados** em classes distintas;
- Com a Classificação, pode-se **prever em que balde colocar** as bolas.



?

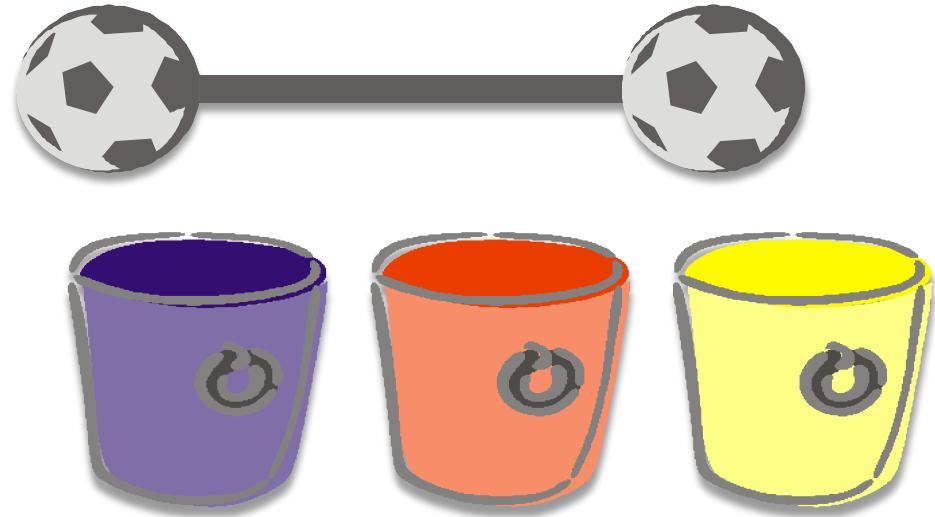


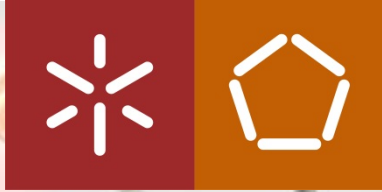


Classificação para Extração de Conhecimento

**O que é
Previsão?**

- O objetivo da Previsão é o de prognosticar ou **deduzir o valor de um atributo, baseado no valor de outros atributos**;
- Com a Previsão pode-se **prever o peso** da bola.

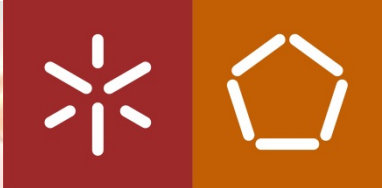




Classificação para Extração de Conhecimento

O que é Previsão?

- Modelação através de técnicas estatísticas:
 - o Regressão Linear: $Y = \alpha + \beta * X + \varepsilon$
 - o Regressão Múltipla: $Y = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \varepsilon$
 - o Regressão Polinomial: $Y = \alpha + \beta_1 * X + \beta_2 * X^2 + \dots$
 - o Regressão de Poisson; (análise de variáveis que exprimem contagens)
 - o Log-linear : $\text{Ln}(Y) = \alpha + \beta * \text{Ln}(X) + \varepsilon$
 - o etc.

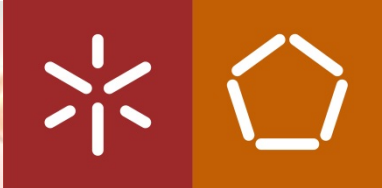


Classificação para Extração de Conhecimento

Classificação versus Previsão?

- Classificação:
 - o Cria-se um modelo, baseado na distribuição dos dados;
 - o O modelo é usado para classificar novos dados;
 - o Dado um modelo, pode-se prever uma nova classe para mapear novos dados;
 - Utiliza-se a Classificação para prognosticar valores discretos.
- Previsão:
 - o Cria-se um modelo, baseado na distribuição dos dados;
 - o O modelo é usado para prever valores futuros dos dados ou, até, valores desconhecidos;
 - Utiliza-se a Previsão para prognosticar valores contínuos.

Classificação é Previsão de valores discretos ou nominais.



Classificação para Extração de Conhecimento

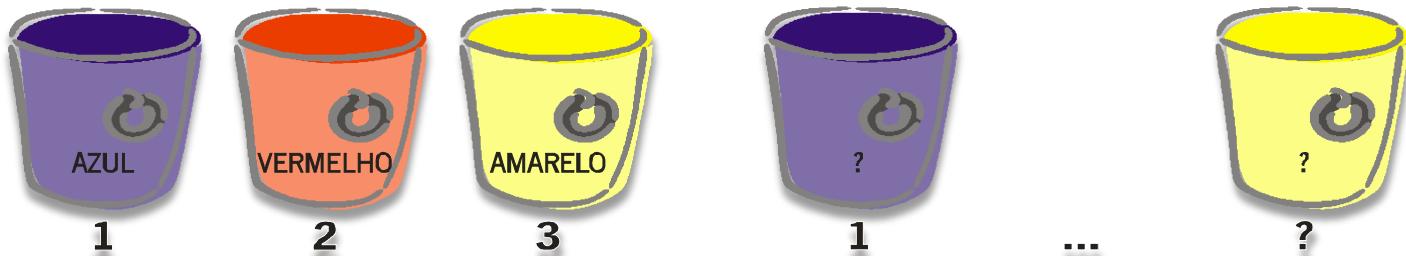
Classificação versus Segmentação?

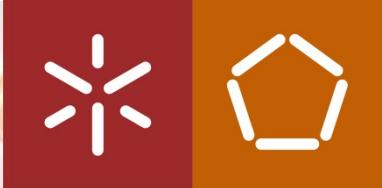
- A **Classificação** é um processo Supervisionado:

○ Conhecem-se as classes (nome/ *label*) e a sua quantidade.

- A **Segmentação** é um processo de Classificação não-Supervisionado:

○ Nem se conhecem as classes nem a quantidade de classes existentes.

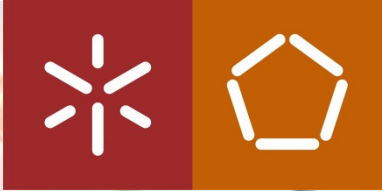




Classificação para Extração de Conhecimento

Exemplos de Classificação

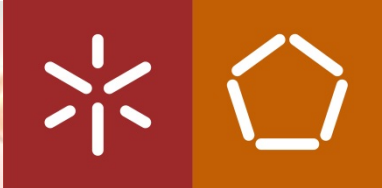
- Determinar se uma aplicação financeira apresenta risco **baixo, médio** ou **alto**; (mas não determinar o montante do investimento)
- Determinar se um cliente é um **bom** candidato à concessão de um empréstimo **ou não**; (mas não determinar o valor do empréstimo)
- Determinar se um produto **é ou não** cancerígeno;
- Determinar se um indivíduo **é ou não** portador de doença grave. (importante a medida do erro e o custo do erro!)



Classificação para Extração de Conhecimento

Construção de um Classificador

- A construção de um modelo de Classificação desenvolve-se em três etapas:
 - o Construção do modelo;
(aprendizagem)
 - o Avaliação do modelo;
(correção ou precisão)
 - o Utilização do modelo;
(classificação ou previsão)



Classificação para Extração de Conhecimento

Construção do Modelo

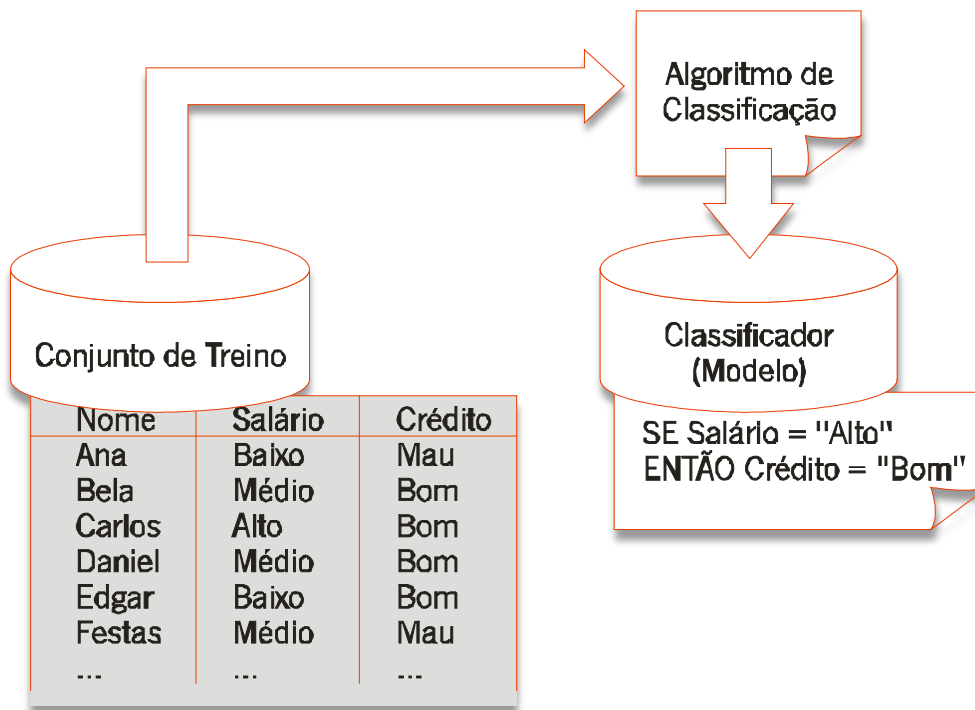
- Aprendizagem:
 - Cada objeto é atribuído a uma classe pré-definida, de acordo com o determinado por um dos atributos, designado o **identificador da classe**;
 - O conjunto de todos os objetos usados para a construção do modelo denomina-se **conjunto de treino**;
 - O modelo pode assumir representações baseadas em:
 - Regras de Classificação (regras de produção na forma Se-Então);
 - Árvores de Decisão;
 - Fórmulas matemáticas.



Classificação para Extração de Conhecimento

Construção do Modelo

- Aprendizagem:

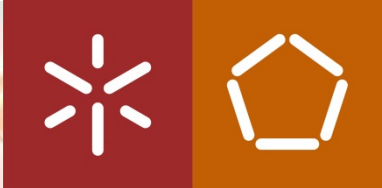




Classificação para Extração de Conhecimento

Avaliação do Modelo

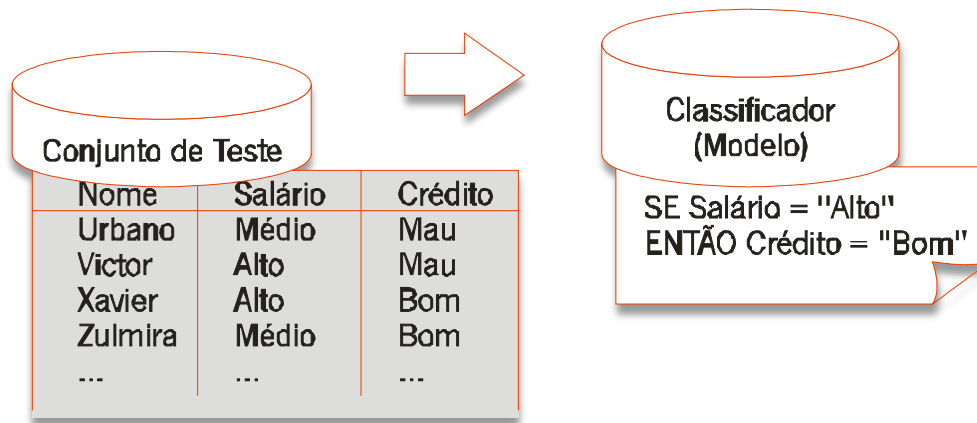
- Correção ou previsão:
 - Estimar a taxa de correção do modelo, baseada num conjunto de objetos de teste;
 - Comparar as classes do conjunto de teste com os resultados da classificação feita pelo modelo;
 - A taxa de correção é a percentagem dos casos de teste corretamente classificados pelo modelo;
 - O conjunto de objetos de teste deve ser independente do conjunto dos objetos de treino, para evitar o sobreajustamento do modelo.

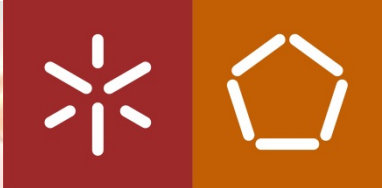


Classificação para Extração de Conhecimento

Avaliação do Modelo

- Correção ou previsão:
 - o Qual a correção deste modelo?

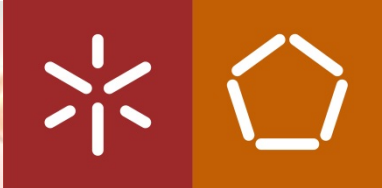




Classificação para Extração de Conhecimento

Utilização do Modelo

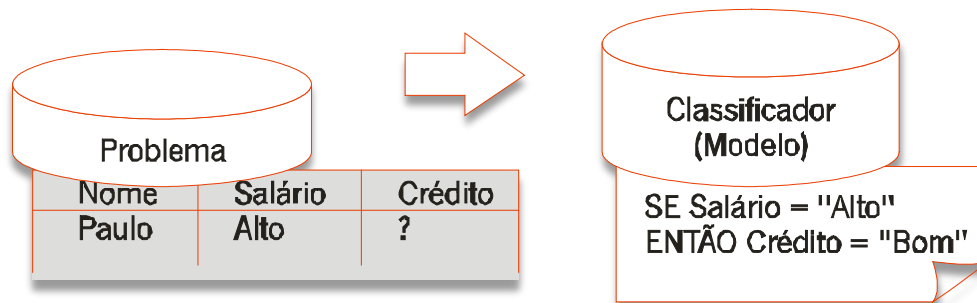
- Classificação:
 - o O modelo deve ser utilizado para classificar objetos desconhecidos (ausentes do conjunto de objetos de treino):
 - Atribuir o nome de uma classe a um objeto novo;
 - Prognosticar o valor de um determinado atributo.

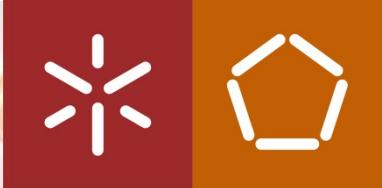


Classificação para Extração de Conhecimento

Utilização do Modelo

- Classificação:
 - o Qual o nível de crédito do Paulo?
 - o Resposta: "Bom".





Classificação para Extração de Conhecimento

Métodos de Classificação

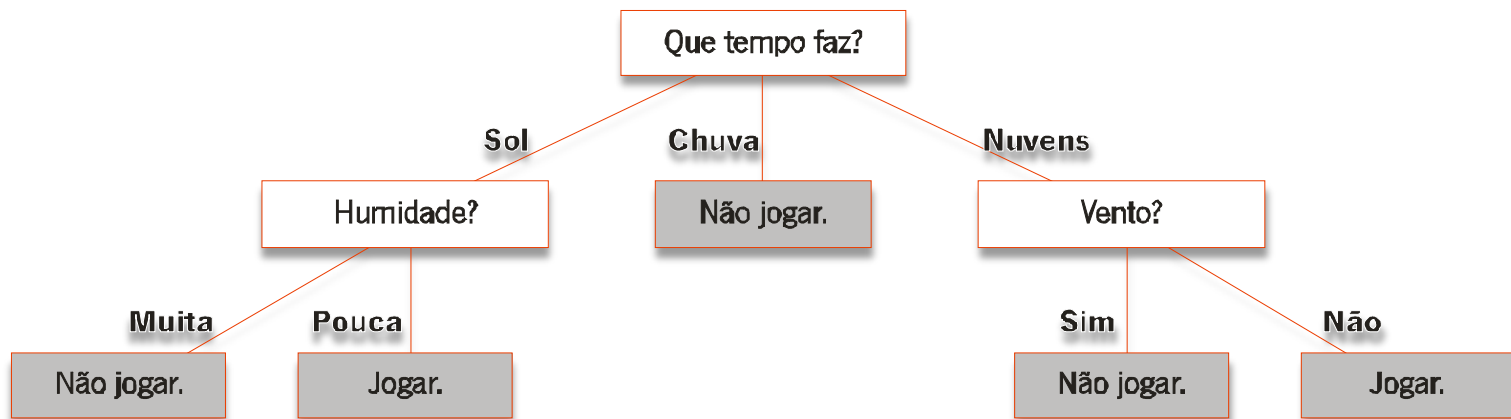
- Árvores de Decisão;
- Redes Neurais Artificiais;
- Classificação Bayesiana;
- Vizinho mais próximo (*k-Nearest Neighbour*);
- Raciocínio Baseado em Casos;
- Algoritmos Genéticos;
- Conjuntos Difusos (*Fuzzy Sets*);
- etc.

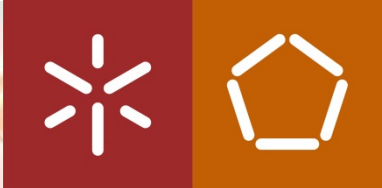


Classificação para Extração de Conhecimento

Árvores de Decisão

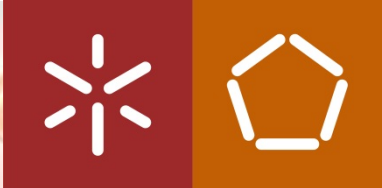
- Cada **nó da árvore** implementa um teste a um atributo;
- Cada **ramo da árvore** respeita a um valor para o atributo testado;
- Cada **folha da árvore** atribui uma classificação.





Classificação para Extração de Conhecimento

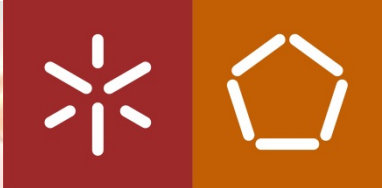
- Árvores de Decisão** ■ A geração de **Árvores de Decisão** segue, normalmente, uma abordagem *top-down* em duas fases:
- o Construção da Árvore:
 - No início, todos os exemplos de treino fazem parte da raiz;
 - Os exemplos são divididos recursivamente, em função do atributo selecionado.
 - o Poda da Árvore (*prunning*):
 - O objetivo é remover (alguns) ramos para evitar o uso de dados que representem lixo ou ruído, de modo a aumentar a correção e precisão da classificação.



Classificação para Extração de Conhecimento

Árvores de Decisão ■ Geração de Árvores de Decisão:

- A Árvore começa com **um nó representando todos os dados**;
- Se a **amostra pertence toda à mesma classe**, então esse **nó** torna-se uma **folha** à qual é atribuído o nome da classe;
- Caso contrário, seleccionar o atributo que **melhor divide** a amostra em classes individuais (como seleccionar o atributo?);
- A recursividade termina quando:
 - A amostra no nó pertence toda à mesma classe;
 - Não há mais atributos para dividir.



Classificação para Extração de Conhecimento

Árvores de Decisão

▪ Ideias importantes na geração de Árvores de Decisão:

o Critério de divisão:

- Como selecionar o atributo pelo qual dividir a árvore;
- Diferentes funções:

ganho de informação;

gini index (medida estatística de desigualdade: $[0;1]$, 0 = igual, 1 = diferente);

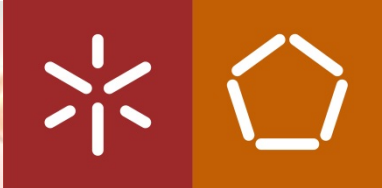
etc.

o Critério de paragem:

- Quando terminar a divisão de nodos (medidas de impureza, ...);

o Regras de identificação:

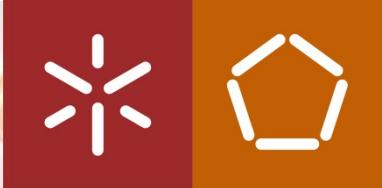
- O nome a dar ao nodo será o da classe à qual pertencem a maior parte dos elementos da amostra.



Classificação para Extração de Conhecimento

Árvores de Decisão

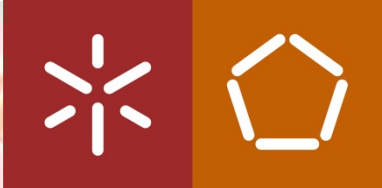
- Exemplos de algoritmos de Árvores de Decisão:
 - o ID3, ID4, ID6;
 - o C4.5, C5;
 - o J48;
 - o CHAID, CART.
- **Árvores de Decisão são algoritmos de classificação especialmente adequados para problemas com muitas dimensões.**



Classificação para Extração de Conhecimento

Redes Neurais Artificiais

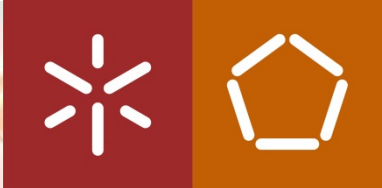
- Redes Neurais Artificiais – Vantagens:
 - o Problemas sem solução algorítmica conhecida ou de difícil implementação;
 - o Problemas com dados incompletos;
 - o Problemas sem soluções conhecidas (influencia a aprendizagem);
 - o Implementação simples;
 - o A correção dos resultados é “normalmente” alta;
 - o O resultado produzido pode ser representado por valores discretos, por valores contínuos ou por vetores de valores.



Classificação para Extração de Conhecimento

Redes Neurais Artificiais

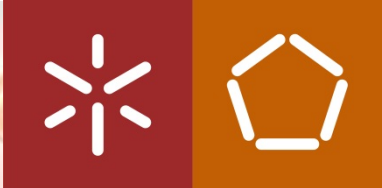
- Redes Neurais Artificiais – Desvantagens:
 - o Dificuldade de classificação quando o problema é caracterizado por um vasto conjunto de padrões;
 - o O treino pode ser um processo muito demorado;
 - o O treino é um processo computacionalmente pesado;
 - o A estratégia de treino pode não ser de fácil identificação/implementação;
 - o Dificil compreensão do significado da função de aprendizagem (métodos de alteração dos pesos das ligações).



Classificação para Extração de Conhecimento

Redes Neurais Artificiais

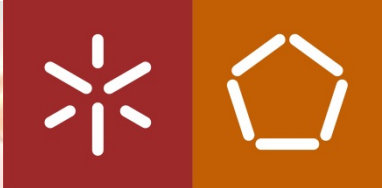
- Exemplos de algoritmos de Redes Neurais Artificiais:
 - *Back propagation*:
 - estratégia de aprendizagem baseada na propagação do erro;
 - *Kohonen*:
 - organização dos neurónios com vista a manter vizinhos aqueles em que predomina a mesma capacidade (variação idêntica da ativação);
 - *Hopfield*:
 - arquitetura totalmente ligada, onde todos os neurónios são de *input* e de *output* e cuja ativação é assíncrona (de valores +1/-1 ou 0/1);
 - etc.



Classificação para Extração de Conhecimento

Redes Neurais Artificiais

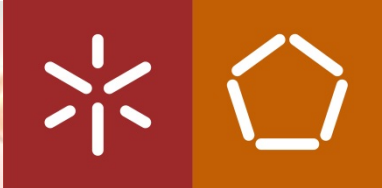
- Construção da RNA:
 - o O número de **nodos de entrada** corresponde ao número de dimensões (atributos) do problema;
 - o O número de nodos nas **camadas intermédias** é determinado durante as fases de treino;
 - o O número de **nodos de saída** corresponde ao número de classes.



Classificação para Extração de Conhecimento

Redes Neurais Artificiais

- Treino da RNA:
 - o O **objetivo** final do **treino** de uma RNA é o de **calibrar os pesos** das ligações, de modo a obter-se a manifestação do comportamento pretendido;
 - o Adoção de regras de aprendizagem para atualização dos pesos das ligações.

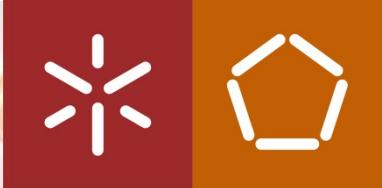


Classificação para Extração de Conhecimento

Outros Métodos de Classificação

▪ Classificação Bayesiana:

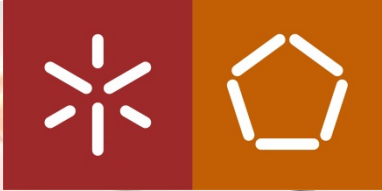
- o São classificadores estatísticos;
- o Conseguem prognosticar a pertença de um objeto a uma classe em termos de valores probabilísticos;
- o Um classificador *Naïve Bayesian* assume total independência entre os atributos;
- o Apresenta bom desempenho em grandes conjuntos de dados e demonstra boa precisão;
- o Aprendizagem incremental.



Classificação para Extração de Conhecimento

Outros Métodos de Classificação

- Vizinho mais próximo (*k-Nearest Neighbour*):
 - o Algoritmo de classificação baseado na aprendizagem por analogia;
 - o Os objetos de treino são descritos por atributos numéricos de 'n' dimensões;
 - o Cada objeto representa um ponto no espaço de 'n' dimensões;
 - o A resolução de um problema procura o vizinho mais próximo de entre os objetos de treino no espaço 'n'-dimensional.

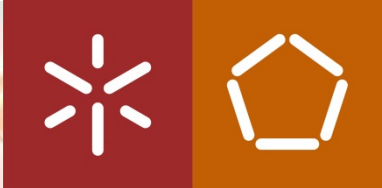


Classificação para Extração de Conhecimento

Outros Métodos de Classificação

▪ Raciocínio Baseado em Casos:

- o Os objetos são representados através de descrições simbólicas complexas;
- o Baseia-se no cálculo de **similaridades** para fundar a resolução do problema na adaptação da solução anterior.

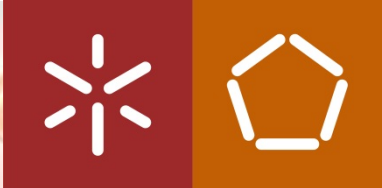


Classificação para Extração de Conhecimento

Outros Métodos de Classificação

▪ Algoritmos Genéticos:

- o Incorporam ideias relacionadas com a evolução natural das espécies;
- o Existam dois valores lógicos A e B, e duas classes X e Y:
 - A regra “se A e não B então X” pode ser representada pelo cromossoma 100;
 - A regra “se não A e B então Y” pode ser representada pelo cromossoma 011;
- o São utilizados, por exemplo, para avaliar a adequação (*fitness*) de outros algoritmos.

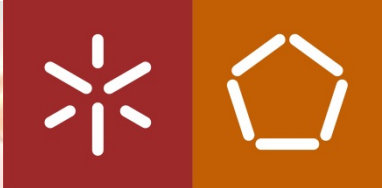


Classificação para Extração de Conhecimento

Outros Métodos de Classificação

▪ Conjuntos Difusos (*Fuzzy Sets*):

- o Lógica multivalor, para lidar com esquemas de raciocínio aproximado (não preciso);
- o A manipulação de regras tem a desvantagem de estabelecer limites muito rígidos:
 - se $SALÁRIO > 50K$ então $CRÉDITO = \text{“Aprovar”}$
- o Pode ser usada a lógica difusa (*fuzzy logic*) para “tornear” esta desvantagem, atribuindo valores de verdade entre 0 e 1 à noção de pertença de um objeto a uma classe.



Classificação para Extração de Conhecimento

Referências bibliográficas

- Data Mining: Concepts and Techniques
Jiawei Han, Micheline Kamber
- Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations
Ian Witten, Eibe Frank



Intelligent Systems Lab

Contactos

- Universidade do Minho
- Escola de Engenharia
- Departamento de Informática
- <http://islab.di.uminho.pt>
- DI-3.22