

**Universidade do Minho**

Escola de Engenharia

Departamento de Informática

# **Classificação para Extração de Conhecimento**

Aprendizagem e Extração de Conhecimento

Perfil Sistemas Inteligentes @ MiEI/4º – 1º Semestre

Cesar Analide, Filipe Gonçalves, Paulo Novais

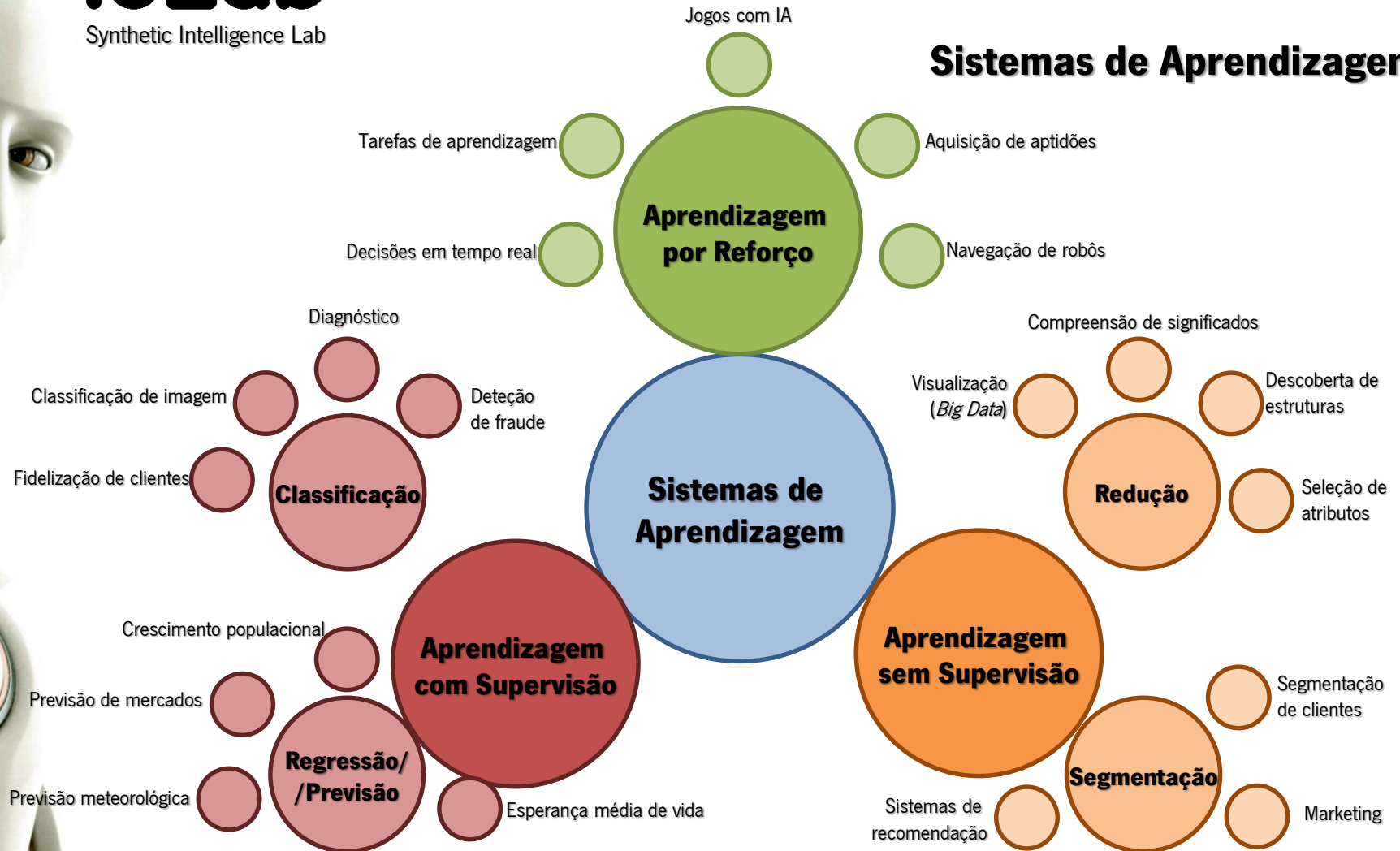


# ISLab

Synthetic Intelligence Lab



## Sistemas de Aprendizagem



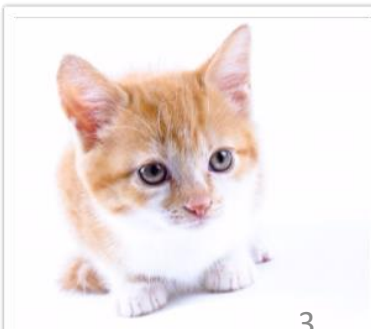
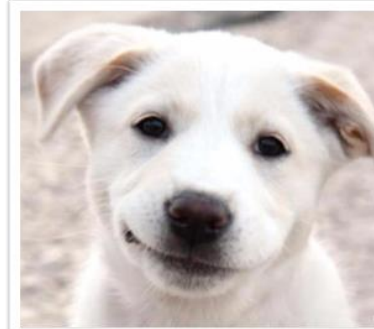


# ISLab

Synthetic Intelligence Lab

## O que é Classificação?

- Um sistema de Classificação **mapeia** cada **objeto** em 1 **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*).





# ISLab

Synthetic Intelligence Lab

## O que é Classificação?

- Um sistema de Classificação **mapeia** cada **objeto** em 1 **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*).

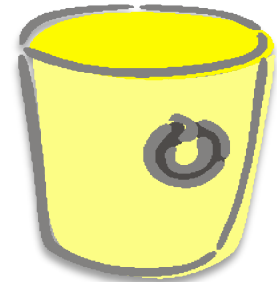
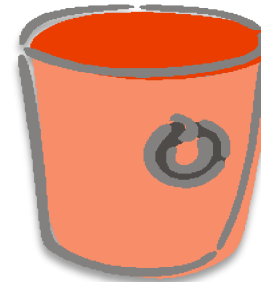
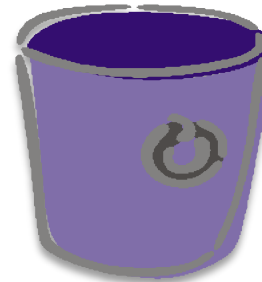






## O que é Classificação?

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





# ISLab

Synthetic Intelligence Lab

## 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.





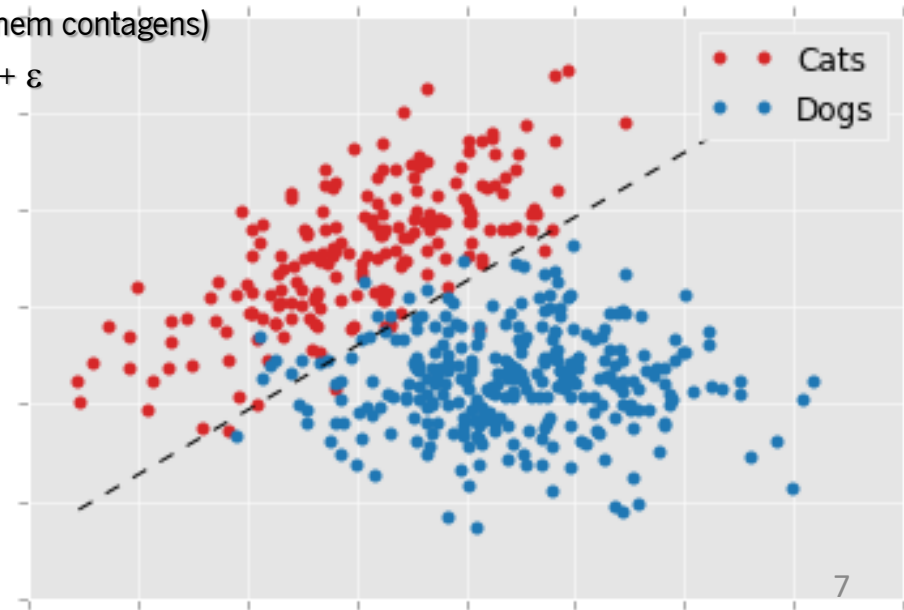
## O que é Previsão?

### ■ Modelação através de técnicas estatísticas:

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

### ■ Sendo:

- Y a variável a prever;
- X a(s) variável(is) preditiva(s);
- $\alpha$  o cruzamento no eixo YY';
- $\varepsilon$  a representação dos erros.





## Classificação *versus* Previsão?

### ■ Classificação:

- Cria-se um modelo, baseado na distribuição dos dados;
- O modelo é usado para classificar novos dados;
- Dado um modelo, pode-se prever uma nova classe para mapear novos dados;

### ■ Utiliza-se a Classificação para prognosticar valores discretos.

### ■ Previsão:

- Cria-se um modelo, baseado na distribuição dos dados;
- 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 *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.





# ISLab

Synthetic Intelligence Lab

## 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!)





## Construção de um Classificador

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





# ISLab

Synthetic Intelligence Lab

## 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 associação, regras de produção na forma Se-Então);
  - Árvores de Decisão;
  - Fórmulas matemáticas.

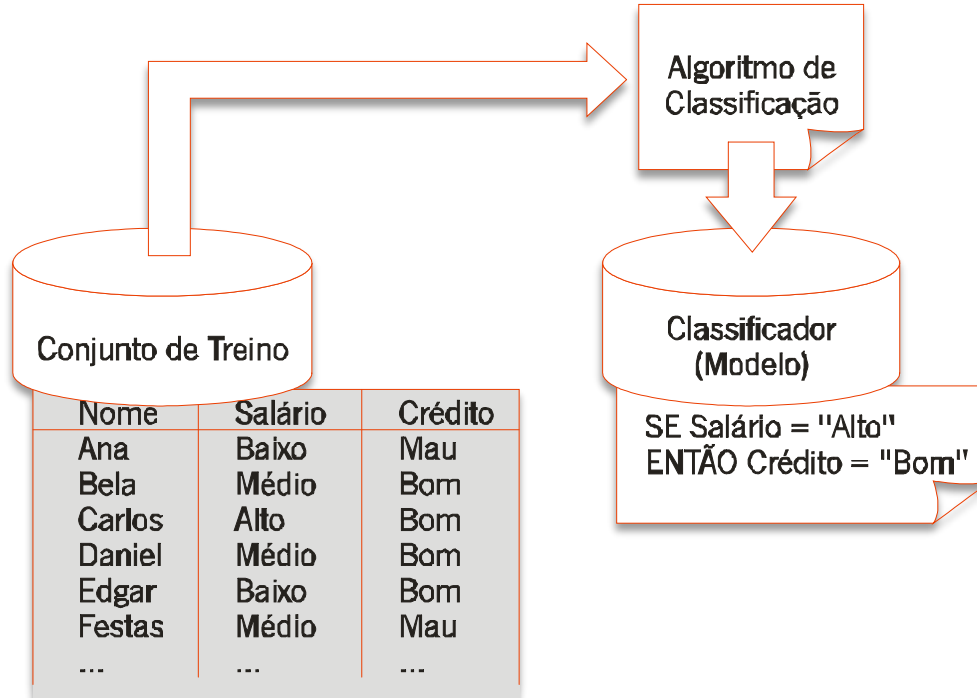






## Construção do Modelo

- Aprendizagem:



- Correção ou avaliação:

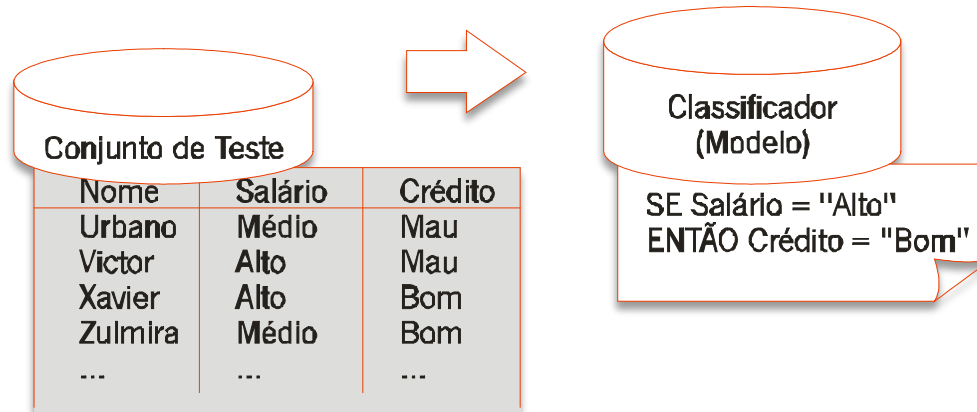
- independente do conjunto dos objetos de treino, para evitar o sobre ajustamento do modelo; representativa da população.





## Avaliação do Modelo

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





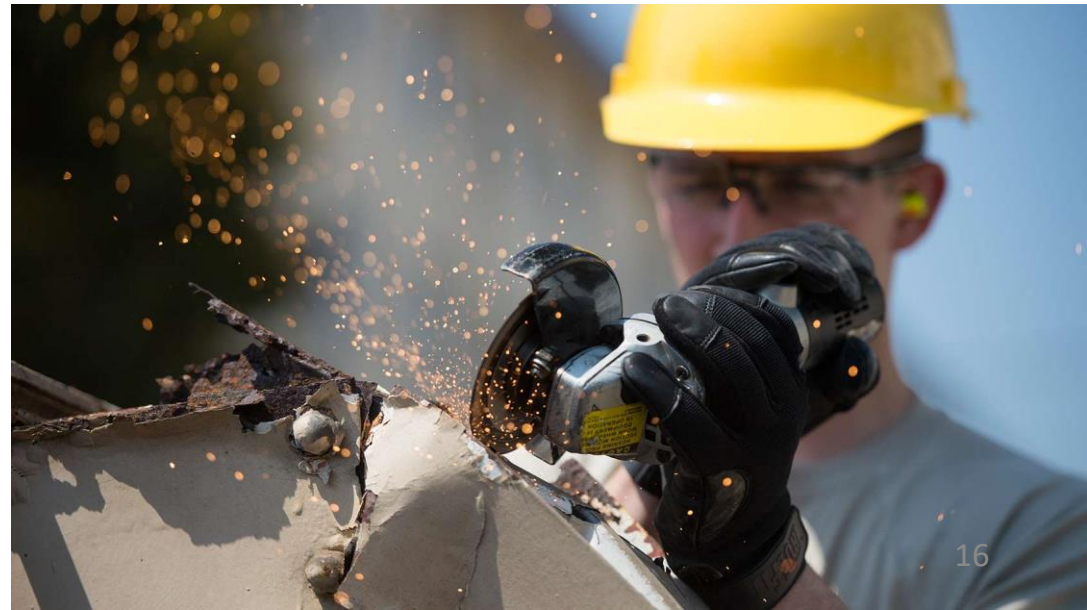
# ISLab

Synthetic Intelligence Lab

## Utilização do Modelo

### ■ Classificação:

- 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.

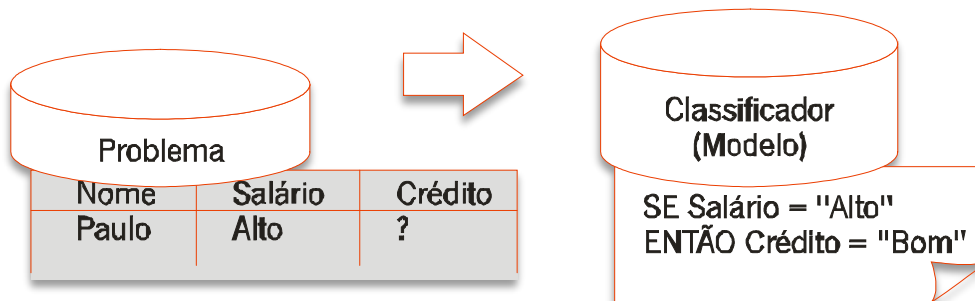




## Utilização do Modelo

- Classificação:

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





# ISLab

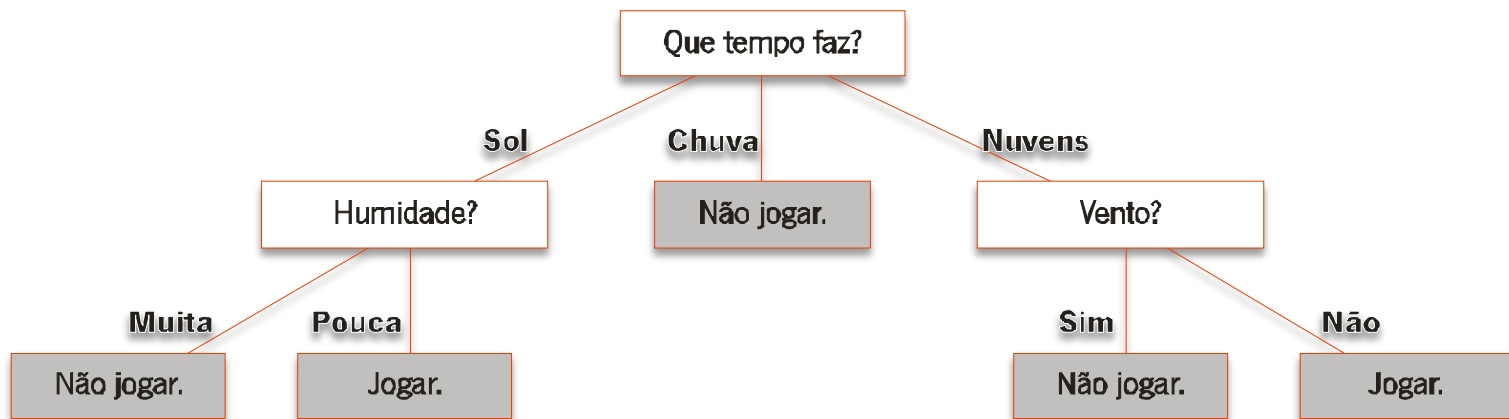
Synthetic Intelligence Lab

## 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.

## 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.





## Árvores de Decisão

- A geração de **Árvores de Decisão** segue, normalmente, uma abordagem *top-down* em duas fases:
  - Construção da Árvore:
    - No início, todos os exemplos de treino fazem parte da raiz;
    - Os exemplos são divididos recursivamente, em função do atributo selecionado.
  - 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.





## 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, selecionar o atributo que **melhor divide** a amostra em classes individuais (como selecionar o atributo?);
- A recursividade termina quando:
  - A amostra no nó pertence toda à mesma classe;
  - Não há mais atributos para dividir.

- Decisões importantes na geração de Árvores de Decisão:
  - 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.
  - Critério de paragem:
    - Quando terminar a divisão de nodos (medidas de impureza, ...);
  - Regras de identificação:
    - O nome a dar ao nodo será o da classe à qual pertencem a maior parte dos elementos da amostra.

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



# ISLab

Synthetic Intelligence Lab

## Redes Neurais Artificiais

### ■ Redes Neurais Artificiais – Vantagens:

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





## Redes Neuronais Artificiais

### ■ Redes Neuronais Artificiais – Desvantagens:

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



## 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.



# ISLab

Synthetic Intelligence Lab

## Redes Neuronais Artificiais

### ■ Construção da RNA:

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



# ISLab

Synthetic Intelligence Lab

## Redes Neuronais Artificiais

- Treino da RNA:

- 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;
- Adoção de regras de aprendizagem para atualização dos pesos das ligações.



## Outros Métodos de Classificação

### ▪ Classificação Bayesiana:

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





# ISLab

Synthetic Intelligence Lab

## Outros Métodos de Classificação

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



# ISLab

Synthetic Intelligence Lab

## Outros Métodos de Classificação

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



## Outros Métodos de Classificação

### ■ Algoritmos Genéticos:

- Incorporam conceitos relacionados com a evolução natural das espécies;
- 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;
- São utilizados, por exemplo, para avaliar a adequação (*fitness*) de outros algoritmos.



## Outros Métodos de Classificação

- Conjuntos Difusos (*Fuzzy Sets*):
  - Lógica multivalor, para lidar com esquemas de raciocínio aproximado (não preciso);
  - A manipulação de regras tem a desvantagem de estabelecer limites muito rígidos:
    - se  $SALÁRIO > 50K$  então  $CRÉDITO = \text{“Aprovar”}$
  - 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.



# ISLab

Synthetic Intelligence Lab

## 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





**Universidade do Minho**

Escola de Engenharia

Departamento de Informática

## Aprendizagem e Extração de Conhecimento

Perfil Sistemas Inteligentes @ MiEI/4º – 1º Semestre

Cesar Analide, Filipe Gonçalves, Paulo Novais