**Implementação de Árvores de Decisão**

FCUP

Inteligência Artificial (CC2006) 2018/2019

Trabalho de Grupo AK

Eduardo Morgado – 201706894 – MIERSI

Simão Cardoso – 201604595 – MIERSI

Sónia Rocha – 201704679 – MIERSI

1. **Introdução:**
   1. **Uma Aprendizagem Supervisionada:**

Para um qualquer problema/ambiente, existem variáveis possíveis de estudar () e, a partir de uma dada função , podemos concluir/chegar a uma solução para o problema inicial, no entanto, encontrar , a função que mapeia as variáveis para as suas imagens/soluções, é uma tarefa difícil, uma vez que, para um problema do dia-a-dia como por exemplo, a escolha do meio de transporte baseado no tempo e trânsito, essa função não é diretamente observável. O objetivo da aprendizagem máquina é então, encontrar essa função para um qualquer problema.

Existem inúmeros métodos/algoritmos de aprendizagem máquina, tal como a *Figura 1* mostra, os algoritmos que vamos estudar são algoritmos para árvores de decisão.

Uma imagem com texto

Descrição gerada automaticamente

Figura - Mapa geral de algoritmos de aprendizagem máquina (imagem retirada de **[1]**)

Tal como referido anteriormente, um problema é composto por um conjunto de observações/variáveis e pela sua solução/*outcome*, um problema pode ser facilmente resolvido caso a função que mapeie as suas variáveis, , seja conhecida, no entanto, para a maior parte dos problemas essa função é desconhecida, apenas podemos observar as variáveis e o seu resultado, com isso, surgem diferentes técnicas de aprendizagem, baseadas na forma como tratam os dados. As técnicas de aprendizagem máquina podem sem divididas em 3 grandes grupos (*Figura 2*):

* Aprendizagem supervisionada: os dados, as sequências de observações estão rotuladas;
* Aprendizagem semi-supervisionada(*reinforcement*): a sequência de observações possuí dados rotulados e dados não rotulados;
* Aprendizagem não supervisionada: a sequência de observações não tem rótulos

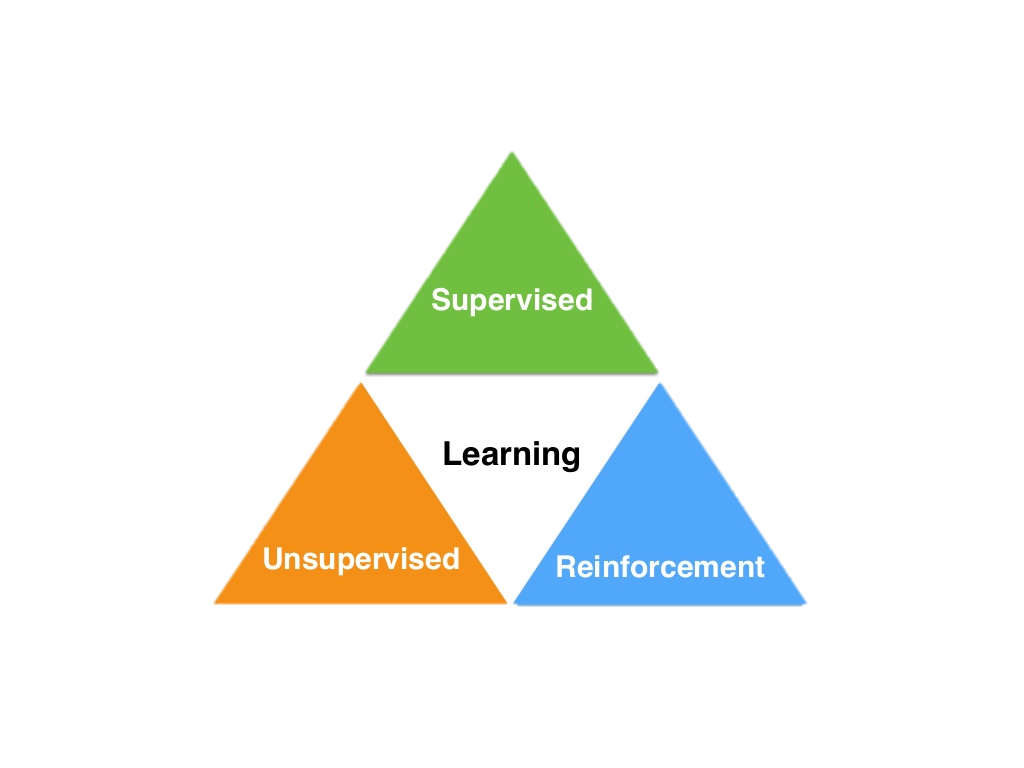


Figura - Distribuição geral de algoritmos de aprendizagem

Sendo assim, na aprendizagem supervisionada, possuímos um conjunto de observações/variáveis (*dataset*), , e um conjunto de classificações corretas (*outcome*), , e os algoritmos vão concluir uma aproximação da função que mapeia os dados, , de forma a, para uma nova observação para o problema possamos concluir o resultado de forma correta, utilizando essa nova função **[1]**.

Os problemas de aprendizagem supervisionada podem ser de dois tipos **[1]**:

* *Classification*: o resultado do problema é uma variável categórica[[1]](#footnote-1);
* *Regression*: o resultado do problema é uma variável contínua[[2]](#footnote-2)
  1. **O que é uma Árvore de Decisão:**

As árvores de decisão enquadram-se na aprendizagem máquina supervisionada. As árvores de decisão podem ser utilizadas para resolver problemas de *classification* e *regression* **[7]**. São muito utilizadas, não só pelo facto de mostrarem como resolver o problema (a deslocação na árvore através de regras inferidas a partir do *dataset* fornecido **[7]** permite ver a lógica da solução) mas também pelo facto de refletirem o pensamento humano tornando-as fáceis de perceber e produzindo interpretações boas.

Numa árvore de decisão, os nós representam atributos[[3]](#footnote-3) cada ramo representa uma regra de decisão[[4]](#footnote-4) e cada folha um rótulo/*label[[5]](#footnote-5).* Um pseudocódigo possível para uma árvore de decisão está representado na *Figura 3*. A *Figura 5* e *4* apresentam uma árvore de decisão para um dado *dataset*, respetivamente.

Figura -Pseudoçõdigo de árvores de decisão (pseudocódigo retirado de **[7]**)

**DecisionTree(dataset):**

1 Se não existem mais atributos então

2 retorna(valorMaisComum(dataset))

3 root <- bestAtribute(dataset)

4 left <- splitData(root.atribute).left

5 right <- splitData(root.atribute).right

6 node\_left <- DecisionTree(left)

7 node-right <- DecisionTree(right)

8 root <- root U left

9 root <- root U right

10 retorna(root)

Existem vários algoritmos para árvores de decisão, tal como a *Figura 1* mostra, no entanto, iremos focar-nos em 3:

* **ID3**: algoritmo baseado em pesquisa gulosa, utiliza *information gain* e *entropy* para construir a árvore, algoritmo para dados categóricos;[[6]](#footnote-6)
* **C4.5**: algoritmo muito semelhante ao ID3, mas para dados contínuos;
* **CART**: utiliza o *gini index* para calcular a impureza dos dados

Uma imagem com texto

Descrição gerada automaticamente

Uma imagem com texto, mapa

Descrição gerada automaticamente

Figura - Dataset (imagem retirada de **[2]**)

Figura -Árvore de decisão para a Figura 4 (imagem retirada de **[2]**)

* + 1. **Pesquisa Não Informada**

1. **Algoritmos de Indução em Árvores de decisão:**

class Table{ //classe para armazenar uma configuração da tabela

String str\_table; //string com a configuração da tabela

int table[]; //matrix representando a tabela

int row,col; //valores das posições, linha coluna, da célula vazia na tabela

int last\_move; //movimento para gerar esta comfiguracão (1->Up/-1->Down/-2->Left/2->Right)

Table(int[][]table,int last\_move); //construtor da tabela a partir de matriz

Table(int[]table,int last\_moves,String s); //construtor de tabela a partir de array

boolean isValid(int r,int c); //retorna true caso valores estejam dentro da tabela

int getBlankRow(); //retorna a linha da célula vazia

int getBlanckCol(); //retorna a coluna da célula vazia

int getLastMove(); //retorna o operador utilizado para gerar esta tabela

String getString(); //retorna a string da configuração da tabela

int[][] getTable(); //retorna uma matriz cópia da tabela

LinkedList<Table> getDescendents(); //retorna a lista das tabelas produzidas a partir desta

}

**Nota Adicional:** Este relatório, bem como o programa utilizado para obtenção dos dados, pode ser acedido no repositório da equipa, <https://github.com/thejoblessducks/the-15-that-was-the-puzzle.git>, ou em <https://github.com/eamorgado/ProjetoIA-15Puzzle.git>

**7. Referências:**

**[1]** Machine Learning Mastery n.d., *Supervised and Unsupervised Machine Learning Algorithms*, Jason Brownlee, consultado pela última vez a 27 de abril de 2019, [<https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>](https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/)

**[2]** S.Russell, P. Norving, “Artificial Intelligence A Modern Approach”, 3rd Ed. , Chapter 18. Learning from Examples 698-700, 2009

**[3]** QUINLAN, J. R. (1986). Induction of decision trees. Machine Learning, 1(1):81-106.

**[4]** QUINLAN, J. R. (1993). C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

**[5]** BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. (1984). Classification and Regression Trees. Wadsworth.

**[6]** Wikipédia 16 de fevereiro de 2019, *C4.5 algorithm*, visitado pela última vez a 20 de Abril de 2019, [<https://en.wikipedia.org/wiki/C4.5\_algorithm>](https://en.wikipedia.org/wiki/C4.5_algorithm)

**[7]** Dataaspirant 30 de janeiro de 2017, *How Decision Tree Algorithm Works*, Rahul Saxena, visitado pela última vez a 27 de abril de 2019, [<http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/>](http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/)

1. Variável não numérica, por exemplo “yes”/”no”. [↑](#footnote-ref-1)
2. Variável numérica, por exemplo idade, pode tomar um qualquer valor real. [↑](#footnote-ref-2)
3. Um atributo é o tipo de observação feita, por exemplo, para a escolha de um transporte público podemos fazer observações baseados em critérios como a existência de trânsito, as condições meteorológicas, o tempo de viagem, esses critérios são atributos. [↑](#footnote-ref-3)
4. Uma regra de decisão é um valor de um critério, por exemplo, para o problema do transporte público, se considerar como um critério a existência de trânsito e se, perante todas as observações, os seus dados possíveis são “sim” e “não” então para o nó/atributo <existência de trânsito> iremos ter dois ramos/regras na nossa árvore: =sim e =não. [↑](#footnote-ref-4)
5. Um rótulo é o *outcome* do caminho que está a ser explorado na árvore, para o problema do transporte público, podemos considerar como rótulos/*outcomes* possíveis: metro,autocarro,elétrico,comboio e avião, por exemplo. [↑](#footnote-ref-5)
6. Na implementação que iremos fazer, o ID3 irá poder classificar dados categóricos e dados contínuos. [↑](#footnote-ref-6)