UNIVERSIDADE DE ÉVORA

Aprendizagem Automática - Trabalho Prático - Relatório

Trabalho realizado por:

Rui Roque nº 42720

Tomás Dias nº 42784

Estruturas de dados e funções utilizadas

Foram utilizadas as seguintes estruturas de dados e funções auxiliares:

- ***class* Valor_Classe()**: Esta classe auxilia o cálculo da impureza de um atributo realizado pela função **calcula_impureza()** descrita de seguida. Recebe como argumento uma **classe** (tipo **String**) de um determinado conjunto de exemplos e tem um atributo **count** (tipo **int**) que guarda o número de ocorrências dessa classe no conjunto de exemplos.

- ***def* calcula_impureza()**: Esta função calcula a **impureza** (tipo **int**) de um atributo de um conjunto de exemplos, retornando-a. Recebe como argumentos a lista de valores do atributo no conjunto de exemplos (**coluna_atributo**), a lista das classes de cada exemplo no conjunto de exemplos (**coluna_classe**), as classes do conjunto (**classes**) e o critério a ser utilizado para o cálculo da impureza (**criterio**).

- ***def* determina_melhor_particao()**: Esta função determina a partição, de entre as partições correspondentes aos valores dos atributos do conjunto de exemplos, com menor impureza, devolvendo uma lista (**melhor_particao**) em que nela estão contidos a lista de valores do atributo com menor impureza (**melhor_coluna**), a localização do atributo no conjunto de exemplos (**melhor_coluna_indice**) e o nome do atributo (**atributo**). Recebe como argumentos o conjunto de exemplos (**x**), a lista das classes de cada exemplo no conjunto de exemplos (**y**), a lista dos atributos (**atributos**), as classes do conjunto (**classes**) e o critério a ser utilizado para o cálculo da impureza (**criterio**).

- ***class* No()**: Esta classe guarda a informação acerca de um nó da árvore de decisão. Recebe como argumentos um atributo do conjunto de exemplos (**atributo**), os valores do atributo que irão ser os ramos do nó (**ramos**), o atributo do nó que o antecede (**no_pai**) e o ramo pelo qual o nó é proveniente (**ramo_pai**).

- ***class* Folha()**: Esta classe guarda a informação acerca de uma folha da árvore de decisão. Recebe como argumentos a classe dos exemplos classificados (**classe**), o número de exemplos classificados (**n_exemplos**), o atributo do nó que o antecede (**no_pai**) e o ramo pelo qual a folha é proveniente (**ramo_pai**).

- ***def* cresce_arvore()**: Esta função é recursiva e tem como objetivo guardar toda a informação necessária para a construção da árvore (nós, folhas e ramos). Recebe como argumentos o conjunto de exemplos (**x**), a lista das classes de cada exemplo no conjunto de exemplos (**y**), a lista dos atributos (**atributos**), a partição com menor impureza (**melhor_particao**), o critério a ser utilizado para o cálculo da impureza (**criterio**) e a lista que irá guardar a informação da árvore (**arvore**).

- ***def* output_ramo()**: Esta função é recursiva e auxilia na apresentação da árvore. Recebe como argumentos o atributo do nó (**atributo**), um dos ramos do nó (**ramo**) e a lista com a informação da árvore (**arvore**).

- ***def* output_arvore()**: Esta função tem como objetivo apresentar a árvore de decisão. Recebe como argumentos a raiz da árvore (**raiz**), os ramos da raiz (**raiz_ramos**) e a lista com a informação da árvore (**arvore**).

- **corre_arvore()**: Esta função auxilia o cálculo da exatidão de um exemplo realizado pela método **score()**. Recebe como argumento a lista com a informação da árvore (**arvore**), um exemplo (**exemplo**), e os atributos do exemplo (**atributos**). Identifica a raiz, o ramo e a identificação do ramo que o exemplo corre passando estes valores como argumentos da função **corre_subarvore()** devolvendo a classe vinda da mesma.

- **corre_subarvore()**: Esta classe auxilia a classe **corre_arvore()** recursivamente tendo como objetivo a identificação do nó, ramo, identificação do ramo até chegar a uma folha devolvendo a classe do exemplo percorrido na árvore.

Classe DecisionTreeREPrune

Após a implementação das funções acima descritas, passou-se para a criação da classe geradora de árvores de decisão.

A classe recebe como argumentos os atributos do conjunto de dados a partir do qual a árvore será gerada (**atributos**), o critério para calcular a impureza (**criterion**) e a indicação se terá ou não pruning (**prune**). No entanto, não foi implementado o pruning sendo o valor predefinido de ***prune=False***.

Aquando da criação do objeto da classe, é criada uma lista **arvore** que ira guardar toda a informação acerca da árvore de decisão (nós, ramos e folhas). Os elementos desta lista serão objetos da classe **No** ou da classe **Folha**.

Como descrito no enunciado do trabalho, implementou-se os seguintes métodos:

- ***def* fit()**
- ***def* score()**

Testes e análise de desempenho sobre os conjuntos de dados

Ficheiro: ***weather.nominal.csv***

```
Árvore de Decisão resultante (utilizando o critério de impureza gini):


outlook
  --> overcast (ramo de: outlook - RAIZ):
  --> { yes : 2 } (folha proveniente do ramo: overcast)
  --> rainy (ramo de: outlook - RAIZ):
  --> windy (nó proveniente do ramo: rainy):
  --> FALSE (ramo de: windy):
  --> { yes : 2 } (folha proveniente do ramo: FALSE)
  --> TRUE (ramo de: windy):
  --> { no : 2 } (folha proveniente do ramo: TRUE)
  --> sunny (ramo de: outlook - RAIZ):
  --> windy (nó proveniente do ramo: sunny):
  --> high (ramo de: windy):
  --> { no : 3 } (folha proveniente do ramo: high)
  --> normal (ramo de: windy):
  --> { yes : 1 } (folha proveniente do ramo: normal)
```

```
Árvore de Decisão resultante (utilizando o critério de impureza entropia):


outlook
  --> overcast (ramo de: outlook - RAIZ):
  --> { yes : 2 } (folha proveniente do ramo: overcast)
  --> rainy (ramo de: outlook - RAIZ):
  --> windy (nó proveniente do ramo: rainy):
  --> FALSE (ramo de: windy):
  --> { yes : 2 } (folha proveniente do ramo: FALSE)
  --> TRUE (ramo de: windy):
  --> { no : 2 } (folha proveniente do ramo: TRUE)
  --> sunny (ramo de: outlook - RAIZ):
  --> windy (nó proveniente do ramo: sunny):
  --> high (ramo de: windy):
  --> { no : 3 } (folha proveniente do ramo: high)
  --> normal (ramo de: windy):
  --> { yes : 1 } (folha proveniente do ramo: normal)
```

```
Árvore de Decisão resultante (utilizando o critério de impureza erro):


humidity
  --> high (ramo de: humidity - RAIZ):
  --> outlook (nó proveniente do ramo: high):
  --> overcast (ramo de: outlook):
  --> { yes : 1 } (folha proveniente do ramo: overcast)
  --> rainy (ramo de: outlook):
  --> windy (nó proveniente do ramo: rainy):
  --> FALSE (ramo de: windy):
  --> { yes : 1 } (folha proveniente do ramo: FALSE)
  --> TRUE (ramo de: windy):
  --> { no : 1 } (folha proveniente do ramo: TRUE)
  --> sunny (ramo de: outlook):
  --> { no : 3 } (folha proveniente do ramo: sunny)
  --> normal (ramo de: humidity - RAIZ):
  --> temperature (nó proveniente do ramo: normal):
  --> cool (ramo de: temperature):
  --> { no : 1 } (folha proveniente do ramo: cool)
  --> hot (ramo de: temperature):
  --> { yes : 1 } (folha proveniente do ramo: hot)
  --> mild (ramo de: temperature):
  --> { yes : 2 } (folha proveniente do ramo: mild)


Percentagem de casos corretamente classificados 25.00%
```

Ficheiro: ***contact-lenses.csv***

```
Árvore de Decisão resultante (utilizando o critério de impureza gini):


tear-prod-rate
  --> normal (ramo de: tear-prod-rate - RAIZ):
  --> astigmatism (nó proveniente do ramo: normal):
  --> no (ramo de: astigmatism):
  --> age (nó proveniente do ramo: no):
  --> pre-presbyopic (ramo de: age):
  --> { soft : 2 } (folha proveniente do ramo: pre-presbyopic)
  --> presbyopic (ramo de: age):
  --> spectacle-prescrip (nó proveniente do ramo: presbyopic):
  --> hypermetrope (ramo de: spectacle-prescrip):
  --> { soft : 1 } (folha proveniente do ramo: hypermetrope)
  --> myope (ramo de: spectacle-prescrip):
  --> { none : 1 } (folha proveniente do ramo: myope)
  --> young (ramo de: age):
  --> { soft : 1 } (folha proveniente do ramo: young)
  --> yes (ramo de: astigmatism):
  --> age (nó proveniente do ramo: yes):
  --> pre-presbyopic (ramo de: age):
  --> { none : 1 } (folha proveniente do ramo: pre-presbyopic)
  --> presbyopic (ramo de: age):
  --> spectacle-prescrip (nó proveniente do ramo: presbyopic):
  --> hypermetrope (ramo de: spectacle-prescrip):
  --> { none : 1 } (folha proveniente do ramo: hypermetrope)
  --> myope (ramo de: spectacle-prescrip):
  --> { hard : 1 } (folha proveniente do ramo: myope)
  --> young (ramo de: age):
  --> { hard : 2 } (folha proveniente do ramo: young)
  --> reduced (ramo de: tear-prod-rate - RAIZ):
  --> { none : 8 } (folha proveniente do ramo: reduced)


Percentagem de casos corretamente classificados 83.33%
```

```
Árvore de Decisão resultante (utilizando o critério de impureza entropia):


tear-prod-rate
  --> normal (ramo de: tear-prod-rate - RAIZ):
  --> astigmatism (nó proveniente do ramo: normal):
  --> no (ramo de: astigmatism):
  --> age (nó proveniente do ramo: no):
  --> pre-presbyopic (ramo de: age):
  --> { soft : 2 } (folha proveniente do ramo: pre-presbyopic)
  --> presbyopic (ramo de: age):
  --> spectacle-prescrip (nó proveniente do ramo: presbyopic):
  --> hypermetrope (ramo de: spectacle-prescrip):
  --> { soft : 1 } (folha proveniente do ramo: hypermetrope)
  --> myope (ramo de: spectacle-prescrip):
  --> { none : 1 } (folha proveniente do ramo: myope)
  --> young (ramo de: age):
  --> { soft : 1 } (folha proveniente do ramo: young)
  --> yes (ramo de: astigmatism):
  --> age (nó proveniente do ramo: yes):
  --> pre-presbyopic (ramo de: age):
  --> { none : 1 } (folha proveniente do ramo: pre-presbyopic)
  --> presbyopic (ramo de: age):
  --> spectacle-prescrip (nó proveniente do ramo: presbyopic):
  --> hypermetrope (ramo de: spectacle-prescrip):
  --> { none : 1 } (folha proveniente do ramo: hypermetrope)
  --> myope (ramo de: spectacle-prescrip):
  --> { hard : 1 } (folha proveniente do ramo: myope)
  --> young (ramo de: age):
  --> { hard : 2 } (folha proveniente do ramo: young)
  --> reduced (ramo de: tear-prod-rate - RAIZ):
  --> { none : 8 } (folha proveniente do ramo: reduced)


Percentagem de casos corretamente classificados 83.33%
```

```
Árvore de Decisão resultante (utilizando o critério de impureza erro):


astigmatism
  --> no (ramo de: astigmatism - RAIZ):
  --> { soft : 10 } (folha proveniente do ramo: no)
  --> yes (ramo de: astigmatism - RAIZ):
  --> { none : 8 } (folha proveniente do ramo: yes)


Percentagem de casos corretamente classificados 66.67%
```

Ficheiro: ***vote.csv***

```
Árvore de Decisão resultante (utilizando o critério de impureza gini):


physician-fee-freeze
  --> n (ramo de: physician-fee-freeze - RAIZ):
  --> { democrat : 84 } (folha proveniente do ramo: n)
  --> y (ramo de: physician-fee-freeze - RAIZ):
  --> synfuels-corporation-cutback (nó proveniente do ramo: y):
  --> n (ramo de: synfuels-corporation-cutback):
  --> { republican : 75 } (folha proveniente do ramo: n)
  --> y (ramo de: synfuels-corporation-cutback):
  --> adoption-of-the-budget-resolution (nó proveniente do ramo: y):
  --> n (ramo de: adoption-of-the-budget-resolution):
  --> water-project-cost-sharing (nó proveniente do ramo: n):
  --> n (ramo de: water-project-cost-sharing):
  --> export-administration-act-south-africa (nó proveniente do ramo: n):
  --> n (ramo de: export-administration-act-south-africa):
  --> { democrat : 1 } (folha proveniente do ramo: n)
  --> y (ramo de: export-administration-act-south-africa):
  --> { republican : 1 } (folha proveniente do ramo: y)
  --> y (ramo de: water-project-cost-sharing):
  --> { republican : 7 } (folha proveniente do ramo: y)
  --> y (ramo de: adoption-of-the-budget-resolution):
  --> aid-to-nicaraguan-contras (nó proveniente do ramo: y):
  --> n (ramo de: aid-to-nicaraguan-contras):
  --> { democrat : 3 } (folha proveniente do ramo: n)
  --> y (ramo de: aid-to-nicaraguan-contras):
  --> { republican : 3 } (folha proveniente do ramo: y)


Percentagem de casos corretamente classificados 93.10%
```

```
Árvore de Decisão resultante (utilizando o critério de impureza entropia):


physician-fee-freeze
  --> n (ramo de: physician-fee-freeze - RAIZ):
  --> { democrat : 84 } (folha proveniente do ramo: n)
  --> y (ramo de: physician-fee-freeze - RAIZ):
  --> synfuels-corporation-cutback (nó proveniente do ramo: y):
  --> n (ramo de: synfuels-corporation-cutback):
  --> { republican : 75 } (folha proveniente do ramo: n)
  --> y (ramo de: synfuels-corporation-cutback):
  --> anti-satellite-test-ban (nó proveniente do ramo: y):
  --> n (ramo de: anti-satellite-test-ban):
  --> adoption-of-the-budget-resolution (nó proveniente do ramo: n):
  --> n (ramo de: adoption-of-the-budget-resolution):
  --> water-project-cost-sharing (nó proveniente do ramo: n):
  --> n (ramo de: water-project-cost-sharing):
  --> export-administration-act-south-africa (nó proveniente do ramo: n):
  --> n (ramo de: export-administration-act-south-africa):
  --> { democrat : 1 } (folha proveniente do ramo: n)
  --> y (ramo de: export-administration-act-south-africa):
  --> { republican : 1 } (folha proveniente do ramo: y)
  --> y (ramo de: water-project-cost-sharing):
  --> { republican : 5 } (folha proveniente do ramo: y)
  --> y (ramo de: adoption-of-the-budget-resolution):
  --> { democrat : 3 } (folha proveniente do ramo: y)
  --> y (ramo de: anti-satellite-test-ban):
  --> { republican : 5 } (folha proveniente do ramo: y)


Percentagem de casos corretamente classificados 93.10%
```

Para o critério de impureza “**erro**” ocorre um erro durante a apresentação da árvore. O valor de exatidão da árvore é de **93.10%**.

Ficheiro: ***soybean.csv***

```
Árvore de Decisão resultante (utilizando o critério de impureza erro):


plant-stand
  --> lt-normal (ramo de: plant-stand - RAIZ):
  --> { brown-spot : 170 } (folha proveniente do ramo: lt-normal)
  --> normal (ramo de: plant-stand - RAIZ):
  --> { alternarialeaf-spot : 251 } (folha proveniente do ramo: normal)


Percentagem de casos corretamente classificados 13.48%
```

Ocorrem erros na construção da árvore para os critérios de impureza “**gini**” e “**entropia**”.