TP555 - AI/ML

Lista de Exercícios #7

Árvores de Decisão

1. Considere o conjunto de treinamento dado na tabela abaixo. Ele é composto por 3 atributos de entrada binários (A1, A2 e A3) e uma saída binária, y. Usando o método ID3, encontre manualmente uma árvore de decisão para este conjunto de dados. Apresente os cálculos feitos para se determinar cada um dos nós.

Exemplo	A1	A2	A3	Output *y*
x1	1	0	0	0
x2	1	0	1	0
x3	0	1	0	0
x4	1	1	1	1
x5	1	1	0	1

2. Considere o conjunto de treinamento dado na tabela abaixo. Ele é composto por 2 atributos de entrada binários (x1 e x2) e uma saída binária, y. Usando o método ID3, encontre manualmente uma árvore de decisão para este conjunto de dados. Apresente os cálculos feitos para se determinar cada um dos nós. Qual o valor do ***Remainder*** para os atributos x1 e x2 durante a escolha do primeiro nó? Qual dos dois atributos é escolhido como primeiro nó? Baseado nesses valores de ***Remainder***, é possível termos uma outra versão da árvore que também classifique corretamente todos os dados do conjunto de treinamento?

XOR		
x1	**x2**	**y**
0	0	0
0	1	1
1	0	1
1	1	0

3. **Exercício sobre árvores de decisão utilizando a métrica ID3**: Neste exercício você criar uma árvore de decisões para prever se o senhor Jair pagará o empréstimo que ele está solicitando junto a um banco para montar uma indústria farmacêutica especializada na produção de hidroxicloroquina. Jair possui os seguintes atributos: **Possui casa própria? Não - Estado civil: Casado - Experiência de trabalho: 3**. Portanto, dado estes três

atributos sobre o senhor Jair, e a árvore montada acima, deve-se emprestar ou não o dinheiro a ele?

OBS.: Todos os atributos são discretos, ou seja, assumem valores de um conjunto finito de valores. Por exemplo, o atributo experiência de trabalho assume apenas os seguintes valores: 0, 1, 2, 3, 4 e 5.

Possui casa própria?	Estado civil	Experiência de trabalho (0-5)	Pagou?
Sim	Solteiro	3	Sim
Não	Casado	4	Sim
Não	Solteiro	5	Sim
Sim	Casado	4	Sim
Não	Divorciado	2	Não
Não	Casado	4	Sim
Sim	Divorciado	2	Sim
Não	Casado	3	Não
Não	Casado	4	Sim
Não	Casado	2	Não
Sim	Casado	2	Sim
Não	Solteiro	2	Sim
Não	Divorciado	3	Não
Não	Solteiro	3	Sim
Sim	Divorciado	3	Sim
Sim	Solteiro	2	Não
Sim	Casado	3	Sim

4. **Exercício sobre árvores de decisão utilizando a métrica ID3**: Considere o conjunto de treinamento para classificação de mamíferos dado na tabela abaixo.

Name	Body Temperature	Gives Birth?	Four-legged?	Hibernates?	Mammal?
porcupine	warm-blooded	yes	yes	yes	yes
cat	warm-blooded	yes	yes	no	yes
bat	warm-blooded	yes	no	yes	yes
whale	warm-blooded	yes	no	no	yes
salamander	cold-blooded	no	yes	yes	no
komodo dragon	cold-blooded	no	yes	no	no

python	cold-blooded	no	no	yes	no
salmon	cold-blooded	no	no	no	no
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

Ele é composto por 4 atributos de entrada (**Body Temperature**, **Gives Birth?**, **Four-legged?** e **Hibernates?**) e uma saída binária, **Mammal?**. Faça o seguinte:

1. Usando o método ID3, encontre **manualmente** uma árvore de decisão para este conjunto de dados. **Apresente todos os cálculos** feitos para se determinar cada um dos nós da árvore. **OBS**.: A coluna **Name** da tabela não deve ser considerada para encontrar a árvore.
2. Observando a árvore obtida, existem atributos que podem ser descartados da base de treinamento, ou seja, existem atributos que não são importantes para a classificação? Se sim, quais são eles?
3. Baseado na sua resposta do item anterior, as árvores de decisão podem ser utilizadas para que tipo de tarefa além, claro, de classificação?
4. Em seguida, de posse da árvore de decisão, classifique os exemplos de teste da tabela abaixo. A classificação feita pela árvore de decisão coincide com as classes da tabela?

Name	Body Temperature	Gives Birth?	Four-legged?	Hibernates?	Mammal?
human	warm-blooded	yes	no	no	yes
pigeon	warm-blooded	no	no	no	no
elephant	warm-blooded	yes	yes	no	yes
turtle	cold-blooded	no	yes	no	no
penguin	warm-blooded	no	no	no	no
dolphin	warm-blooded	yes	no	no	yes
platypus	warm-blooded	no	yes	no	yes
spiny anteater	warm-blooded	no	yes	yes	yes

5. Houve algum erro de classificação no conjunto de teste? Se sim, o que poderia ser feito para melhorar a acurácia do classificador.
6. Baseado na sua resposta do item anterior, apresente a nova árvore de decisão (ou seja, apresenta o desenho da nova árvore) e **todos os cálculos** feitos para encontrá-la.
7. A nova árvore de decisão classifica os exemplos de validação com 100% de acurácia? Apresente os resultados de classificação feitos pela nova árvore.

5. **Exercício sobre árvores de decisão utilizando o método ID3**. Considere o conjunto de treinamento dado pela tabela abaixo.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0

2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	M	Luxury	Large	C1

Ele é composto por 3 atributos de entrada e uma saída binária. Usando o método ID3, encontre **manualmente** uma árvore de decisão para este conjunto de dados. Apresente todos os cálculos feitos para se determinar cada um dos nós da árvore e um desenho da árvore obtida. **OBS**.: A coluna **Customer ID** da tabela não deve ser considerada para encontrar a árvore.

6. **Exercício sobre árvores de decisão utilizando a métrica ID3**: Considere o conjunto de treinamento para classificação de doenças dado pela tabela abaixo.

Example Number	Fever	Vomiting	Diarrhea	Shivering	y (class)
1	No	No	No	No	Healthy (H)
2	Average	No	No	No	Influenza (I)
3	High	No	No	Yes	Influenza (I)
4	High	Yes	Yes	No	Salmonella Poisoning (S)
5	Average	No	Yes	No	Salmonella Poisoning (S)
6	No	Yes	Yes	No	Bowel Inflammation (B)
7	Average	Yes	Yes	No	Bowel Inflammation (B)

Ele é composto por 4 atributos de entrada (**Fever, Vomiting, Diarrhea, Shivering**) e uma saída, **y**. Faça o seguinte:

1. Usando o método ID3, encontre **manualmente** uma árvore de decisão para este conjunto de dados. **Apresente todos os cálculos** feitos para se determinar cada um dos nós da árvore. **OBS.:** A coluna "**Example Number**" da tabela não deve ser considerada para encontrar a árvore. (**Dica**: Leia o documento do link a seguir para um exemplo de como calcular a entropia para casos onde tem-se mais de duas classes, como é o caso deste exercício. Cálculo da entropia para mais de duas classes.)
2. Observando a árvore obtida, quais atributos poderiam ser descartados da base de treinamento, ou seja, quais atributos não são importantes para a classificação?
3. De posse da árvore de decisão, classifique os exemplos de validação da tabela abaixo. A classificação feita pela árvore de decisão coincide com as classes da tabela de validação, ou seja, o classificador atinge acurácia de 100%?

Validation Examples	Fever	Vomiting	Diarrhea	Shivering	y (class)
1	High	No	No	Yes	Influenza (I)
2	High	No	No	No	Bowel Inflammation (B)
3	High	Yes	Yes	Yes	Salmonella Poisoning (S)
4	No	No	Yes	Yes	Bowel Inflammation (B)
5	Average	Yes	Yes	Yes	Bowel Inflammation (B)
6	Average	No	Yes	Yes	Salmonella Poisoning (S)

4. Houve algum erro de classificação no conjunto de validação? Se sim, o que poderia ser feito para melhorar a acurácia do classificador.
5. Apresente a nova árvore de decisão e **todos os cálculos** feitos para encontrá-la.
6. A nova árvore de decisão classifica os exemplos de validação com 100% de precisão?

7. Treine e ajuste uma ***árvore de decisão*** para o conjunto de dados das luas (*moons dataset*).
   a. Gere um conjunto de dados das luas usando: make_moons(n_samples = 10000, noise = 0.4, random_state=42).
   b. Divida-o em um conjunto de treinamento e um conjunto de testes usando: train_test_split (X, y, test_size=0.25, random_state=42).
   c. Plote os dados do conjunto de treinamento em relação às classes a que pertencem. Ou seja, defina marcadores diferentes para identificar cada um das classes na figura. Por exemplo, use círculos para denotar exemplos que pertencem à classe 0 e quadrados para denotar exemplos que pertencem à classe 1.

d. Use o ***Grid Search*** com validação cruzada (com a ajuda da classe ***GridSearchCV***) para encontrar bons valores de hiperparâmetro para um DecisionTreeClassifier. (**Dica**: tente vários valores para max_leaf_nodes.)
e. Treine o modelo com o conjunto de treinamento usando os valores do hiperparâmetro e meça o desempenho do modelo no conjunto de teste. Você deve obter aproximadamente 85% a 87% de precisão.
f. Plote as seguintes informações
    - A árvore de decisão encontrada com o valor ótimo do hiperparâmetro.
    - A matriz de confusão.
    - A fronteira de decisão.
    - A curva ROC.

8. Neste exercício você irá continuar o exercício anterior e criar uma floresta de árvores de decisão.
    a. Continuando o exercício anterior, gere 1000 subconjuntos a partir do ***conjunto de treinamento***, com cada um contendo 100 exemplos selecionados aleatoriamente. (**Dica**: use a classe ***ShuffleSplit*** do ScikitLearn para isso. O ***ShuffleSplit*** fornece índices para subconjuntos de treinamento e teste, porém, neste exercício você irá apenas utilizar os índices criados para o subconjunto de treinamento, podendo ignorar os índices do subconjunto de testes. O conjunto de testes que será utilizado é o criado no exercício anterior com a função ***train_test_split***. A documentação da classe ***ShuffleSplit*** pode ser acessada através do seguinte link: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ShuffleSplit.html)
    b. Treine uma ***árvore de decisão*** em cada um dos 1000 subconjuntos de treinamento, usando os melhores valores de hiperparâmetros encontrados no exercício 3 ou execute o ***Grid Search*** novamente. Avalie cada uma das 1000 ***árvores de decisão*** no conjunto de teste original, ou seja o conjunto criado no exercício 3 (lembre-se, não é o subconjunto de testes gerado pelo ***ShuffleSplit***). Como foram treinadas em conjuntos menores, essas ***árvores de decisão*** provavelmente terão desempenho pior que a ***árvore de decisão*** do exercício 3, atingindo provavelmente cerca de 80% de precisão.
    c. Agora vem a mágica das ***florestas aleatórias***. Para o conjunto de teste original, gere predições com as 1000 árvores de decisão e mantenha apenas a predição mais frequente (**Dica**: você pode usar a função mode() da biblioteca SciPy para isso: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mode.html). Essa abordagem fornece previsões por maioria de votos a partir do conjunto de teste original.
    d. Meça a precisão das predições obtidas com conjunto de teste original (**Dica**: utilize a função ***accuracy_score*** para medir a precisão). Você deve obter uma precisão um pouco maior que o modelo do exercício 3 (cerca de 0,5 a 1,5%

maior). Ao final deste exercício, você terá treinado o que é conhecido como um classificador baseado em ***florestas aleatórias***.