Lista 2
Inteligência Artificial

Iyan Lucas Duarte Marques[1]

[1]Instituto de Ciências Exatas e Informática - Pontifícea Universidade Católica Minas Gerais (PUC-MG)

1. Questão 01

Usando a base de restaurantee os parâmetros default do algoritmo J48, preencha os valores das seguintes métricas para esta base de dados a partir da árvore gerada pelo Weka

	TVP	TNP	TFP	TFN	Recall	Precisão	F1Score
Esperar pelo restaurante	0,500	0,500	0,500	0,500	0,667	0,500	0,571
Não esperar pelo restaurante	0,500	0,500	0,500	0,500	0,400	0,500	0,444

2. Questão 02

uais as principais diferenças entre as funções ID3 e C4.5? Explique.

2.1. ID3

O ID3 foi um dos primeiros algorítmos de classificação e é baseado em um sistema de inferência e em conceitos de sistemas de aprendizagem. Ele constrói árvores de decisão a partir de um dado conjunto de exemplos, sendo a árvore resultante usada para classificar amostras futuras. O ID3 separa um conjunto de treinamento em subconjuntos, de forma que estes contenham exemplos de uma única classe. A divisão é efetuada através de um único atributo, que é selecionado a partir de uma propriedade estatística, denominada ganho de informação, que mede quanto informativo é um atributo. Após a construção de uma árvores de decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados no treinamento. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta a novas situações, podendo, também, se estimar a proporção de erros e acertos ocorridos na construção da árvore.

2.2. C4.5

O algoritmo C4.5 é um aprimoramento do algoritmo ID3, isto devido ao fato de trabalhar com valores indisponíveis, com valores contínuos, podar árvores de decisão e derivar regras.

3. Questão 03

Explique a diferença entre a métrica 'Razão de ganho' adotada pelo algoritmo C4.5 e a métrica 'ganho de informação' adotada pelo ID3.

3.1. ID3

A propriedade ***ganho de informação***, mede como um determinado atributo separa os exemplos de treinamento de acordo com a classificação deles. O ID3 usa o ganho de informação para selecionar, entre os candidatos, os atributos que serão utilizados a cada

passo, enquanto constrói a árvore. Para definir ganho de informação, começa se definindo uma medida, chamada entropia, que caracteriza a impureza de uma coleção arbitrária de exemplos. Dada uma coleção S que contém exemplos positivos e negativos de algum conceito objetivo, a entropia de S relativa a esta classificação lógica é:

$$H = -p_+ + \ln(p_+) - p_- \ln(p_-) \tag{1}$$

[1]

3.2. C4.5

Em cada nó da árvore, o algorítmo escolhe o atributo dos dados que mais efetivamente divide seu conjunto de amostras em subconjuntos enriquecidos, em uma classe ou outra. O critério de divisão é o ganho de informação normalizado (diferença na entropia). O atributo com o maior ganho de informação normalizado é escolhido para tomar a decisão. O algoritmo C4.5 então recorre nas sublistas particionadas.

4. Questão 04
O que significa aprendizado supervisionado? E não supervisionado?

4.1. Supervisionado

O indutor recebe conjunto de exemplos com uma entrada e um rótulo. As técnicas utilizadas são:

- Redes Neurais do tipo Multilayer Perceptron
- Máquinas de Vetores Suporte
- Árvores de Decisão

4.2. Não Supervisionado

O indutor recebe conjunto de exemplos somente com uma entrada e tenta encontrar agrupamentos.

- Redes Neurais do tipo mapas auto-organizáveis
- Algoritmo k-médias

5. Questão 05
Explique e dê exemplos dos tipos de problema: classificação, regressão, agrupamento e associação

- **Classificação:** É um dos tipos de problemas mais utilizados que prevê ou descreve uma classe. O atributo de classificação é nominal.
  Exemplo:

Chuva
Sim
Não
Sim

[1] $[H, \eta]$ é a letra grega eta, símbolo que representa a entropia (somente a letra maiúscula). p é a letra que representa o resultado, positivo ou não de uma coleção S.

- **Regressão:** Tal qual o problema de classificação porém seus atributos de classe são numéricos.
  Exemplo:

Medalhas
5
7
1

- **Agrupamento (clusterização):** O problema busca agrupar as instâncias de acordo com os atributos de entrada. O problema não leva em consideração o atributo de classificação.
  Exemplo: Agrupar pessoas com interesses em comum ao planejar uma festa.
- **Regras de associação:** O problema busca semelhanças e infere associações entre os elementos.
  Exemplo: Se uma pessoa escolheu x, também escolheu y.

6. Questão 06

Utilizando-se a base de dados"Jogar ou não tênis", explique como o algoritmo J48 trata dados contínuos. Para isso, leia o artigo *"A comparative study of decision tree ID3 and C4.5"* que está no CANVAS.

O algoritmo C4.5 ordena todos os valores em ordem crescente e elimina os valores repetidos. Então, começa o processo de particionar para cada um dos valores em 2 classes, um vez que lida com valores numéricos discretizando os mesmos. Reunindo desta forma os elementos com valores menores e iguais, e elementos com valores maiores. Para cada um é calculado o *"gain ratio"* e o valor escolhido para ser o limiar é o de maior *"gain ratio"*.

7. Questão 07

Explique os métodos de amostragem:

7.1. Cross-validaton

O Cross-validation é um procedimento de reamostragem usado para avaliar modelos de aprendizado de máquina em uma amostra de dados limitada.

O procedimento tem um único parâmetro chamado k que se refere ao número de grupos em que uma determinada amostra de dados deve ser dividida. Como tal, o procedimento é frequentemente denominado validação cruzada k-fold. Quando um valor específico para k é escolhido, ele pode ser usado no lugar de k na referência ao modelo, como k = 10 tornando-se uma validação cruzada de 10 vezes.

A validação cruzada é usada principalmente no aprendizado de máquina aplicado para estimar a habilidade de um modelo de aprendizado de máquina em dados invisíveis. Ou seja, usar uma amostra limitada para estimar como o modelo deve funcionar em geral quando usado para fazer previsões sobre dados não usados durante o treinamento do modelo.

É um método popular porque é simples de entender e porque geralmente resulta em uma estimativa menos tendenciosa ou menos otimista da habilidade do modelo do que outros métodos, como uma divisão simples de treinamento / teste.

O procedimento do Cross-validation é:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
   (a) Take the group as a hold out or test data set
   (b) Take the remaining groups as a training data set
   (c) Fit a model on the training set and evaluate it on the test set
   (d) Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

7.2. Holdout

Este estimador divide os exemplos em uma porcentagem fixa de exemplos p para treinamento e $(1 - p)$ para teste, considerando normalmente $p > 1/2$. Valores típicos são $p = 2/3$ e $(1 - p) = 1/3$, embora não existam fundamentos teóricos sobre estes valores.

Uma vez que uma hipótese construída utilizando todos os exemplos, em média, apresenta desempenho melhor que uma hipótese construída utilizando apenas uma parte dos exemplos, este método tem a tendência de super estimar o erro verdadeiro. Para pequenos conjuntos, nem sempre é possível separar uma parte dos exemplos.

De forma a tornar o resultado menos dependente da forma de divisão dos exemplos, pode-se calcular a média de vários resultados de holdout através da construção de várias partições obtendo-se, assim, uma estimativa média do holdout.

8. Questão 08

Explique em que consiste o problema de desbalanceamento de classe no contexto de aprendizado de máquina. Cite e explique as principais abordagens para balanceamento de classes. Veja em Etapas de pré-processamento /Parte 1 -Processamento -Balanceamento

A classe desbalanceada ocorre quando temos um dataset que possui muitos exemplos de uma classe e poucos exemplos da outra classe. Nessa situação se você tem um modelo de classificação, por exemplo, o resultado desse modelo será enviesado, ou seja ele tende a classificar os novos dados como sendo da classe que possui mais exemplos.

- **Undersampling** – Consiste em reduzir de forma aleatória os exemplos da classe majoritária.
- **NearMiss** – É um algoritmo de undersampling que consiste em reduzir de forma aleatória os exemplos da classe majoritária, porém ele seleciona os exemplos com base na distância. Quando instâncias das duas classes estão muito próximas umas das outras, removemos as instâncias da classe majoritária para termos uma maior separabilidade entre as classes.
- **Oversampling** – Consiste em replicar dados aleatórios da classe minoritária. Como estamos duplicando os dados já existentes este método está propício a dar overfitting.

- **Smote** – Consiste em gerar dados sintéticos (não duplicados) da classe minoritária a partir de vizinhos. Ele calcula quais são os vizinhos mais próximos e as características desses vizinhos para criar novos dados. Se o número de dados gerados for muito grande podemos ter um overfiting.
- **Engenharia de features** – Consiste em gerar novas features dando maior robustez ao modelo e diminuindo o viés da classe majoritária.
- **Usar diferentes custos para as diferentes classes** – É definido custos de classificação para cada classe.
- **Induzir um modelo para uma classe** – A classe minoritária ou a classe majoritária são aprendidas separadamente.