Lista 2 - Árvore de decisão

Vitor Lucio

Questão 01

c) Iris_Versicolor, íris_Setosa, Iris_Versicolor, Iris_Virgínica

Questão 02

c) I e II, apenas.

Questão 03

	Precisão	Recall	F1Score	TVP	TFN	TFP	TVN
A	0.58	0.58	0.58	0.58	0.41	0.06	0.94
B	0.65	0.83	0.72	0.83	0.16	0.07	0.93
C	0.76	0.66	0.71	0.66	0.33	0.06	0.94
D	0.89	0.87	0.88	0.87	0.12	0.09	0,91

Questão 04

O algoritmo CART utiliza essa métrica (Gini) para decidir como dividir os dados em subgrupos em cada nó da árvore. A ideia é escolher a divisão que minimize a impureza nos subgrupos resultantes.

A fórmula é: **$Gini(D)=1-\sum_{i=1}^{c}(pi)^2$**

Em que: D é o conjunto de dados em um nó da árvore. c é o número de classes no conjunto de dados. pi é a proporção de registros da classe i no conjunto de dados D. Um valor de Gini próximo de 0 indica uma pureza alta, enquanto um valor de Gini próximo de 1 indica uma impureza alta.

Questão 05

Parte 1 - Processsamento - Balanceamento

Aprendizado de Máquina têm seu desempenho prejudicado na presença de dados desbalanceados. Algoritmos tendem a favorecer a classificação de novos dados na classe majoritária.

As principais técnicas para resolver este problema são:

- Redefinir o tamanho do conjunto de dados
- Utilizar diferentes custos de classificação para as diferentes classes
- Induzir um modelo para uma classe

Parte 2 - Processamento - Dados ausentes

Base de dados podem conter dados ausentes e isto apresenta
dificuldades relacionadas à qualidade dos dados. Dados ausentes podem ser causadas por problemas nos equipamentos que realizam a coleta, ou só a falta mesmo dos dados.

Formas de resolver:

- Eliminar as instâncias com dados ausentes
- Definir e preencher manualmente valores para os atributos com valores ausentes
- Utilizar algum método para automaticamente definir valores para atributos com valores ausentes
- Empregar algoritmos de AM que lidam internamente com valores ausentes

Parte 3 - Processamento - Dados inconsistentes e redundantes

- Dados **inconsistentes** são aqueles que possuem valores conflitantes em seus atributos.
- Dados **redundantes** podem se referir tanto a instâncias quanto a atributos, no qual instancias estão repetidas ou atributos repetem o mesmo valor.
- **O fato é que bases reais têm muita redundância e inconsistência!**
- Existem filtros para eliminar tais problemas, como RemoveMisclassified.

Parte 4 - Processamento - Conversão simbólica-numérica

- Quando o atributo é do tipo **simbólico** e assume apenas dois valores, **utilizar um dígito binário**.
- Quando o atributo é do tipo **simbólico,** assume mais de dois valores e **ordinal, utilizar valores numéricos em ordem** (1,2,3,4).
- Quando o atributo é do tipo **simbólico**, assume mais de dois valores e **nominal**, os atributos devem manter inexistência de uma relação de ordem (binarizar o atributo).

Parte 5 - Processamento - Conversão numérico-simbólica

Algumas técnicas de AM foram desenvolvidas para trabalhar com valores qualitativos. Alguns destes algoritmos têm o seu desempenho reduzido quando o fazem. Nestas situações, a recomendação é discretizar o atributo.

Atributo quantitativo discretizado: o conjunto de possíveis valores é dividido em intervalos, e cada intervalo de valores quantitativos é convertido em um valor qualitativo.

Os métodos de discretização podem ser **supervisionados** e não **supervisionados**. As técnicas supervisionadas geralmente tem melhores resultados.

Estratégias utilizadas pelos diferentes métodos :

- **Larguras iguais**: Divide o intervalo original de valores em subintervalos em mesma largura.
- **Frequências iguais:** Atribui o mesmo número de objetos a cada subintervalo.
- Uso de um algoritmo de agrupamento de dados
- Inspeção visual

Parte 6 - Processamento - transformação de atributos numéricos

Quando os menor e maior de valor dos atributos são muito diferentes ou se eles estão em escalas diferentes, o valor numérico de um atributo precisa ser transformado em outro valor numérico.

A normalização por reescala: define uma nova escala de valores, limites mínimo e máximo, para todos os atributos.

$$v_{Novo} = \min + \frac{v_{Atual} - \text{menor}}{\text{maior} - \text{menor}}(\max - \min)$$

A normalização por padronização: método usado para ajustar os dados de forma que tenham média zero e desvio padrão unitário. Esse processo é especialmente útil para comparar dados que estão em diferentes escalas

$$v_{Novo} = \frac{v_{Atual} - \mu}{\sigma}$$

Parte 7 - Processamento - Redução de dimensionalidade

Maldição de dimensionalidade: Problemas ocorrem quando há muitos atributos. Para resolver é feito a combinação ou eliminação dos atributos irrelevantes.

Agregação:

- Substituem os atributos originais por novos atributos formados pela combinação de grupos de atributos
- Levam à perda dos valores originais dos atributos, o que pode ser importante dependendo do contexto (finanças, saúde, etc)

Seleção de atributos:

- Mantem uma parte dos atributos originais e descartam os demais atributos

Para avaliar a qualidade desses métodos é usado **Embutida**, **Baseada em Filtro**, **Baseada em Wrapper**.

Embutida

- A seleção do subconjunto é embutida no próprio algoritmo de aprendizado (ex: Árvore de decisão)

Baseada em Filtro

- Em uma etapa de pré-processamento, é utilizado um filtro sobre o conjunto de atributos original que filtra um subconjunto de atributos do conjunto original, sem levar em consideração o algoritmo de aprendizado que utilizará esse subconjunto (Ex: correlação)

Baseada em Wrapper

- Utiliza o próprio algoritmo de aprendizado como uma caixa-preta para a seleção;
- Para cada possível subconjunto, o algoritmo é consultado e o subconjunto que apresentar a melhor combinação entre redução da taxa de erro e redução do número de atributos é em geral selecionado