escola britânica de artes criativas & tecnologia

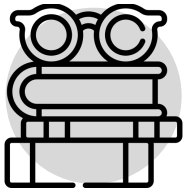
Profissão: Cientista de Dados

escola
britânica de
artes criativas
& tecnologia

GLOSSÁRIO

Árvores II

- **Compreenda a classificação multinominal**
- **Prepare a base**
- **Construa a classificação multinominal**
- **Conheça o cross-validation**
- **Analise os tipos de cross-validation**



Dica: para encontrar rapidamente a palavra que procura aperte o comando CTRL+F e digite o termo que deseja achar.

Compreenda a classificação multinominal

Compreenda a classificação multinominal

Classificação Binária

É um tipo de classificação onde a variável resposta tem apenas duas possíveis categorias. Exemplos incluem sobrevivência ou morte em um naufrágio, cura ou não cura de uma doença, pagamento ou não pagamento de uma dívida.

Classificação Multinomial

É um tipo de classificação onde a variável resposta tem mais de duas possíveis categorias. O exemplo principal dado na aula é classificar um pinguim em uma de três raças, com base em suas características biométricas.

Compreenda a classificação multinominal

Entropia

É uma métrica de impureza usada em árvores de decisão. A entropia é calculada usando logaritmos e é máxima quando não se tem ideia de a que classe o indivíduo pertence, e mínima quando se tem certeza absoluta.

Métrica de Gini

É outra métrica de impureza usada em árvores de decisão. A métrica de Gini é definida como um menos a soma das probabilidades ao quadrado. Assim como a entropia, a impureza é máxima quando não se tem ideia de a que classe o indivíduo pertence, e mínima quando se tem certeza absoluta.

Prepare a base

Prepare a base

- **Árvore de classificação binária**
  É um tipo de algoritmo de aprendizado de máquina que divide os dados em dois grupos com base em uma condição. É usado quando a variável de resposta tem apenas duas categorias possíveis.

- **Árvore de classificação multinomial**
  É um tipo de algoritmo de aprendizado de máquina que divide os dados em mais de dois grupos. É usado quando a variável de resposta tem mais de duas categorias possíveis.

Construa a classificação multinominal

Construa a classificação multinominal

Acurácia

É uma métrica de avaliação de modelos de classificação. Representa a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões.

Custo de Complexidade

É um parâmetro que controla o tamanho da árvore de decisão. Valores maiores de custo de complexidade resultam em árvores menores, o que pode ajudar a evitar o sobreajuste.

Árvore de Classificação Multinomial

É um tipo de modelo de aprendizado de máquina que é usado para classificar instâncias em uma de três ou mais classes.

Medida de Impureza

É uma métrica usada para determinar a melhor divisão em cada etapa da construção da árvore de decisão. Quanto menor a impureza, melhor a divisão.

Construa a classificação multinominal

- **Pacote 'rpart'**
  É uma biblioteca em R usada para construir árvores de decisão.

- **Podar a Árvore**
  É o processo de remover as divisões da árvore de decisão que não contribuem significativamente para a precisão do modelo. Isso pode ajudar a evitar o sobreajuste.

- **Regras de Decisão**
  São as condições usadas para dividir os dados em subconjuntos na árvore de decisão.

- **Sobreajuste**
  É um problema que ocorre quando um modelo de aprendizado de máquina é muito complexo e se ajusta demais aos dados de treinamento, resultando em um desempenho ruim nos dados de teste.

Conheça o cross-validation

Conheça o cross-validation

CCP alfa

É um hiperparâmetro de uma árvore de decisão que controla a complexidade do modelo. O professor usa o conjunto de validação para otimizar este hiperparâmetro.

Hiperparâmetros

São parâmetros de um algoritmo de aprendizado de máquina que são definidos antes do treinamento e não são aprendidos a partir dos dados.

Cross-validation

É uma técnica usada para avaliar a capacidade de um modelo de aprendizado de máquina de generalizar para uma população mais ampla. Envolve a divisão dos dados em conjuntos de treino, validação e teste.

Semente do gerador de números aleatórios

É um número inicial usado para gerar uma sequência de números aleatórios. Mudar a semente pode afetar a acurácia do modelo, pois altera a divisão dos dados em conjuntos de treino, validação e teste.

Analise os tipos de cross-validation

Analise os tipos de cross-validation

- **Exaustivos**

  São métodos de validação cruzada que testam todas as possíveis combinações de observações na base de dados. Incluem o "leave-one-out" (Lino) e o "leave-pair-out" (LPO).

- **Leave-one-out" (Lino)**

  Método exaustivo de validação cruzada onde uma observação é removida da amostra de treino, um modelo é treinado com as demais observações e a observação removida é classificada.

- **Hierárquicos**

  São métodos de validação cruzada usados quando se deseja treinar o modelo e avaliar seu desempenho ao mesmo tempo. Incluem o "k-fold com holdout" e o "nested k-fold".

- **Leave-pair-out" (LPO)**

  Método exaustivo de validação cruzada onde duas observações são removidas, um modelo é treinado com as demais observações e as duas observações removidas são classificadas.

Analise os tipos de cross-validation

Nested k-fold

Método hierárquico de validação cruzada onde um loop interno é usado para treinar o modelo e um loop externo é usado para testar o modelo.

Subamostragem sequencial

Método não exaustivo de validação cruzada onde um grupo é separado, um modelo é treinado com as demais observações e a métrica de avaliação é calculada.

Não exaustivos

São métodos de validação cruzada que não testam todas as possíveis combinações de observações na base de dados. Incluem o "k-fold" e a "subamostragem sequencial".

K-Fold Cross Validation

É uma variação da validação cruzada onde o conjunto de dados é dividido em 'k' subconjuntos de igual tamanho. O modelo é então treinado 'k' vezes, cada vez usando um subconjunto diferente como conjunto de teste e os demais como conjunto de treino.

escola britânica de artes criativas & tecnologia

Bons estudos!