escola britânica de artes criativas & tecnologia

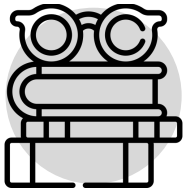
Profissão: Cientista de Dados

escola
britânica de
artes criativas
& tecnologia

GLOSSÁRIO

Hierárquicos/ aglomerativos

- **Diferencie previsão de explicação**
- **Faça inferência sobre os parâmetros**
- **Diferencie qualidade de complexidade**
- **Selecione modelos**
- **Regularização**



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

Conheça o Scikit-learn

Conheça o Scikit-learn

- **Algoritmos não supervisionados**

  São algoritmos de aprendizado de máquina que não precisam de um conjunto de dados de treinamento com respostas corretas. Eles são usados para encontrar padrões e estruturas ocultas nos dados.

- **Agrupamento hierárquico**

  É um método de agrupamento que busca construir uma hierarquia de grupos. As estratégias para agrupamento hierárquico geralmente são de dois tipos: aglomerativas (de baixo para cima) ou divisivas (de cima para baixo).

- **Agrupamento aglomerativo**

  É um tipo de algoritmo de agrupamento hierárquico que começa com cada observação em seu próprio grupo e, em seguida, combina os grupos com base em alguma medida de similaridade.

- **Função 'transform'**

  É uma função em Python usada para padronizar os dados, ou seja, para transformar os dados de modo que tenham média zero e desvio padrão um.

Conheça o Scikit-learn

Ligação

É o critério usado para determinar a distância entre conjuntos de observações em um algoritmo de agrupamento aglomerativo. Os tipos comuns de ligação incluem ligação única, ligação completa, ligação média e ligação de Ward.

Padronização

É o processo de transformar os dados para que tenham média zero e desvio padrão um. Isso é feito para garantir que todas as variáveis tenham a mesma escala e, portanto, a mesma importância no modelo.

Crie algoritmo

Crie algoritmo

- **Critério de Ligação Completa**
  É um método usado para calcular a distância entre dois clusters em um algoritmo de agrupamento hierárquico. No critério de ligação completa, a distância entre dois clusters é definida como a maior distância entre um objeto em um cluster e um objeto no outro cluster.

- **Pacote 'distance' do módulo 'spal'**
  É um pacote Python usado para calcular distâncias entre pontos ou clusters em um algoritmo de agrupamento hierárquico.

- **Matriz de Distâncias**
  É uma tabela que contém as distâncias entre todos os possíveis pares de pontos em um conjunto de dados. A matriz é simétrica, com a diagonal principal sendo sempre zero, pois a distância de um ponto a ele mesmo é zero.

Visualize pelo Dendograma

Visualize pelo Dendograma

Dendrograma

É um diagrama de árvore que visualiza a hierarquia de clusters formada por agrupamento hierárquico. O eixo X representa o índice ou rótulo de cada ponto, enquanto o eixo Y representa a distância de agrupamento.

Rótulo de Ponto

É o identificador único de cada ponto de dados no dendrograma, representado no eixo X.

Ponto de Quebra

É a distância no dendrograma onde a formação de um novo cluster ocorre. A análise do ponto de quebra pode ajudar a determinar o número ideal de clusters.

Conheça os tipos de ligação

Conheça os tipos de ligação

Ligação Completa

Define a distância de um ponto a um conjunto como a distância do ponto ao elemento mais distante do conjunto.

Ligação Simples

Define a distância de um ponto a um conjunto como a distância do ponto ao elemento mais próximo do conjunto.

Ligação Média

Calcula todas as distâncias do ponto a cada ponto do grupo e tira a média entre todas essas distâncias.

Ligação de Ward

Baseada no conceito da soma de quadrados das distâncias e é um critério mais estatístico.

Aplique distância

Aplique distância

- **Distância Dice**

  É uma medida de distância que leva em consideração a presença ou ausência de características específicas em dois indivíduos. É uma alternativa à distância euclidiana para dados categorizados.

- **Distância Gower**

  É uma combinação da distância Dice para dados categorizados e a distância Manhattan para dados quantitativos. É a medida de distância mais recomendada para trabalhar com dados mistos.

- **Distância Euclidiana**

  É uma medida de distância comum que é calculada usando o teorema de Pitágoras.

- **Distância Manhattan (City block)**

  Esta medida de distância é baseada na ideia de que só se fazem deslocamentos horizontais e verticais. Independentemente do caminho escolhido, a distância total percorrida será a mesma.

Classifique

Classifique

Centroide

Em clusterização, o centroide de um cluster é um ponto central que é uma média aritmética de todos os pontos no cluster.

Random Forest

É um algoritmo de aprendizado supervisionado. Como o nome sugere, este algoritmo cria a floresta com um número de árvores.

Variáveis Explicativas (x) e a Variável Resposta (o grupo)

Em modelagem estatística e machine learning, as variáveis explicativas são as características ou fatores que influenciam uma variável resposta. A variável resposta é a característica ou fator que é influenciado ou previsto.

escola britânica de artes criativas & tecnologia

Bons estudos!