Aprendizagem automática (P02)

Inteligência Artificial, 2022-23

Nuno Mendes (2727), Rosário Silva (21138) Tiago Azevedo (21153) Francisco Pereira (21156)

Introdução

O objetivo deste projeto é implementar e analisar diferentes abordagens e métodos de Machine Learning (ML) para resolver um problema específico usando um conjunto de dados aberto/público.

Podemos encontrar muitos conjuntos de dados públicos em https://www.kaggle.com/datasets.

O Dataset escolhido para este projeto foi "Previsão de Clientes de Cartão de Crédito", uma vez que é um bom exemplo para colocar em prática a matéria abordada na unidade curricular.

O projeto deve usar um único conjunto de dados para regras de classificação, agrupamento e associação.

Função ML A

A abordagem que nós optamos foi a Regra de Classificação, uma vez que vis a atribuição de classes/categorias, atribui uma classe aos novos dados, o atributo de classe é discreto, o que tem poucos valores distintos, e o modelo é baseado nas relações existentes entre os vários atributos e o atributo de classe.

Os objetivos definidos são: um mapeamento dos diferentes níveis de educação, género, estado civil, entre outros…

Os critérios de seleção de dados são: selecionou-se todos os dados.

Função ML B

A abordagem que nós optamos foi a Regra de Agrupamento que pretende agrupar objetos semelhantes de acordo com as semelhanças encontradas entre os atributos, é usado como uma funcionalidade primária de mineração de dados , para organizar clientes em segmentos, pode ser usado como uma técnica de pré-processamento para outros algoritmos, discretizar atributos contínuos na indução de árvores de classificação.

Os objetivos definidos são: divisão de 2 grupos.

Os critérios de seleção de dados são: 16154 amostras no primeiro cluster e 7555 amostras no segundo cluster, com base no número de meses de inatividade por um período de um ano.

Função ML C

A abordagem que nós optamos foi a Regra de Associação que tem como objetivo básico encontrar elementos que implicam na presença de outros elementos em uma mesma transação, encontrar relacionamento ou padrões frequentes entre conjuntos de dados.

Os objetivos definidos são: efetuar um novo output de um dataframe com as regras de associação minadas fazendo o uso do algoritmo Apriori. Cada linha deverá representar uma regra diferente.

Os critérios de seleção de dados são: Não existe nenhum critério de seleção.

Função ML D

A abordagem que nós optamos foi a SVM que tem como objetivo básico um algoritmo poderoso e versátil que pode ser usado para resolver problemas de classificação e regressão lineares e não lineares, e é particularmente útil ao trabalhar com dados de alta dimensão, o objetivo é escolher um hiperplano com a maior margem possível, pois este terá o menor erro de generalização, os pontos mais próximos do hiperplano, chamados vetores de suporte, determinam a posição do hiperplano.

Os objetivos definidos são: Efetuar um output relativo ao treino de um modelo de ML e cross validation.

Os critérios de seleção de dados são: Não existe nenhum critério de seleção.

Análise de Resultados

Link do repositório: https://github.com/tiagoazevedo22/LESI3_SmartCampus

Função ML A:

Este resultado é o resultado de um modelo de classificação usando o algoritmo KNN.

A tabela mostra o mapeamento dos diferentes níveis de educação, bandeiras de rotatividade, gênero, estado civil, categoria de renda e categoria de cartão.

A matriz de confusão mostra o número de verdadeiros positivos (2086), falsos positivos (48), falsos negativos (202) e verdadeiros negativos (196) para as previsões do modelo.

O relatório de classificação mostra a precisão, recall, f1-score e suporte para as duas classes, "Existing Customer" e "Attrited Customer".

A precisão é o número de verdadeiros positivos dividido pela soma de verdadeiros positivos e falsos positivos, e mede quantas das previsões positivas foram corretas.

O recall é o número de verdadeiros positivos dividido pela soma de verdadeiros positivos e falsos negativos, e mede quantos dos casos positivos reais foram corretamente identificados pelo modelo.

O f1-score é a média harmônica de precisão e recall, e leva em conta tanto os valores de precisão e recall.

O suporte é o número de observações em cada classe.

As estatísticas finais mostram o número total de previsões corretas e incorretas e o percentual de previsões corretas.

Em resumo, a saída mostra o desempenho do modelo KNN na classificação das classes "Existing Customer" e "Attrited Customer".

O modelo tem uma alta precisão de 90% e uma precisão de 91% para a classe "Existing Customer", mas um recall e f1-score menores de 49% e 61% respetivamente para a classe "Attrited Customer".

Isso significa que o modelo é bom em identificar clientes existentes, mas não tão bom em identificar clientes desistentes.

Função ML B:

Este resultado mostra que o algoritmo KMeans foi utilizado para agrupar os dados em 2 clusters com base na coluna "Months_Inactive_12_mon".

Os dados foram divididos em 2 grupos, com 16154 amostras no primeiro cluster e 7555 amostras no segundo cluster.

A precisão do algoritmo é de 0,0028636318751851485.

Função ML C:

Esse código executa a mineração de regra de associação em um conjunto de dados chamado "BankChurners.csv" usando o algoritmo Apriori.

O conjunto de dados é carregado primeiro em um Pandas DataFrame e, em seguida, uma lista de transações é criada a partir dos dados.

A função a priori é então usada para minerar as regras de associação da lista de transações, com suporte mínimo, confiança mínima, elevação mínima, comprimento mínimo e máximo das regras especificadas como parâmetros de entrada. As regras extraídas são então convertidas em uma lista e passadas para a função inspect() que formata em um dataframe pandas com as colunas 'Left_Hand_Side', 'Right_Hand_Side', 'Support', 'Confidence', 'Lift'.

A saída final são as 10 principais regras com maior elevação usando nlargest(n = 10, colunas = 'Lift').

A saída final é um DataFrame pandas contendo as regras de associação, onde cada linha representa uma regra, com as colunas representando o lado esquerdo, lado direito, suporte, confiança e aumento de cada regra.

A coluna 'Suporte' mostra a percentagem de transações em que a regra ocorre, a coluna 'Confiança' indica a percentagem do tempo que a regra é verdadeira e a coluna 'Lift' representa o aumento na proporção da ocorrência da regra, regra comparada com sua ocorrência esperada se os itens fossem independentes.

Este output é um DataFrame contendo as regras de associação mineradas a partir do conjunto de dados usando o algoritmo Apriori. Cada linha representa uma regra diferente e as colunas representam as seguintes informações:

- Left_Hand_Side: A parte esquerda da regra, que é a antecedente ou a parte "se" da regra.
- Right_Hand_Side: A parte direita da regra, que é a consequente ou a parte "então" da regra.
- Suporte: A percentagem de transações em que a regra ocorre.
- Confiança: A percentagem de vezes que a regra é verdadeira, ou seja, a percentagem de transações em que a parte esquerda também contém a parte direita.
- Elevação: Aumento na razão de ocorrência da regra em relação à sua ocorrência esperada se os itens fossem independentes.

Por exemplo, a primeira linha do output mostra que 0,364% das transações contêm a parte esquerda (antecedente) e a parte direita (consequente) da regra, a regra é verdadeira em 66,6667%

das vezes, a elevação desta regra é 6,535948, o que significa que esta regra é 6,535948 vezes mais provável de ocorrer do que se a parte esquerda e a parte direita fossem independentes.

Vale a pena notar que os valores em Left_Hand_Side e Right_Hand_Side estão codificados de uma forma que pode não ser imediatamente interpretável, parece que os dados foram codificados de alguma forma, então é difícil dizer o que esses valores representam sem mais contexto.

Função ML D:

Este output parece ser o resultado de um processo de treinamento de modelo de aprendizado de máquina e validação cruzada.

A primeira linha indica que o output é um DataFrame com 10127 linhas e 9 colunas.

As próximas linhas dão o nome de cada coluna e o número de valores não-nulos para cada coluna, bem como o tipo de dados de cada coluna.

A última linha mostra o uso de memória.

As próximas linhas mostram o processo de ajuste de 10 dobras para cada um dos 18 candidatos, com um total de 180 ajustes.

O modelo usa o algoritmo SVM com um núcleo linear e rbf, e o parâmetro C com diferentes valores.

O tempo gasto para ajustar cada candidato também é mostrado. O "CV" significa validação cruzada.

Parece que a saída está mostrando os resultados de um processo de validação cruzada de busca em grade, onde uma grade de todas as combinações possíveis dos parâmetros especificados é treinada e avaliada usando a validação cruzada, a fim de encontrar a melhor combinação de parâmetros para o modelo.

Este é o processo de ajuste fino dos parâmetros do modelo para otimizar seu desempenho.

Conclusão

A realização deste projeto permitiu à nossa equipa efetuar uma pesquisa pormenorizada dos diferentes algoritmos especialmente concebidos para a previsão de um determinado resultado mediante o input de um dataset.

A nossa equipa reconhece a importância e usabilidade destes algoritmos para a previsão de vários cenários do dia a dia. Isto poderá permitir as empresas de um âmbito comercial puderem analisar as suas listagens de clientes e implementar processos para maximizar vendas ou retenção de clientes.

Além disto, é de frisar que uma máquina opera estes dados de uma forma mais simples, mais organizada e muito mais rápido do que um humano, pelo que poderá contribuir para a eficiência da empresa no geral.