Engenharia de Computação

INTELIGÊNCIA ARTIFICIAL

Algoritmos de Classificação

Guilherme Baccarin

23 de Junho de 2022

Proposta **2**

Desenvolvimento **3**
Base de Dados 3
Dicionário 3

Algoritmos **4**
Árvores de Decisão (Decision Tree) 4
Naïve Bayes 4
Máquina de Vetores de Suporte — SVM 4
Aperfeiçoamento dos dados 5
Exclusão de alguns atributos 5
Remoção de outliers 5

Resultados **6**

Conclusão **7**

Proposta

Neste trabalho foi proposto o desenvolvimento de um algoritmo que usasse três diferentes modelos de aprendizado de máquina para resolver um problema de classificação. Foi usada uma base de dados retirada da internet, de tema e formato livre, desde que representasse um problema de classificação. Um banco de dados de classificação é definido por ter um desfecho que seja composto por dois ou mais conjuntos de possibilidades.

Os algoritmos deveriam ser treinados e analisados, primeiramente, com o banco de dados sem modificações, para que pudessem ter sua acurácia comparada posteriormente com os resultados do treino feito após o pré-processamento dos dados. A acurácia é medida com a porcentagem de acerto que o algoritmo desempenhou após ser testado com 30% dos dados da mesma base que foi usada para treiná-lo(com os outros 70% de dados).

Os algoritmos foram desenvolvidos em python e os resultados foram obtidos com as médias de 3 execuções cada, antes e depois do pré-processamento e foram rodados na máquina local. Foi implementado, além dos resultados de acurácia, um timer para gerar resultados de tempo de treino.

Desenvolvimento

Base de Dados

Este conjunto de dados contém informações cardiovasculares de pacientes e foi retirado do link.

Dicionário

age	Idade do paciente
sex	Sexo do paciente
cp	Tipo de dor torácica ~ 0 = Angina típica, 1 = Angina atípica, 2 = Dor não anginosa, 3 = Assintomática
trtbps	Pressão arterial em repouso (em mm Hg)
chol	Coletoral em mg/dl obtido através do sensor de IMC
fbs	(fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False
restecg	(glicemia em jejum > 120 mg/dl) ~ 1 = Verdadeiro, 0 = Falso
thalachh	Frequência cardíaca máxima atingida
oldpeak	Pico anterior
slp	Declive
caa	Número de vasos principais
thall	Resultado do teste Thalium~ (0,3)
exng	Angina induzida por exercício ~ 1 = Yes, 0 = No
output	Target variable

Desfecho: Possibilidade de ataque do coração.

Algoritmos

Árvores de Decisão (Decision Tree)

Quando vamos construir uma árvore de decisão, o algoritmo tenta encontrar a melhor variável para começar a construção do nó, depois ele escolhe o melhor ponto para separar os dados: uma parte para um lado, e outra para outro. E a partir disso, vai construindo a árvore, sempre procurando a melhor condição possível.

Naïve Bayes

O classificador multinomial Naïve Bayes é um dos modelos mais populares no aprendizado de máquina. Tomando como premissa a suposição de independência entre as variáveis do problema, o modelo de Naïve Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes pré-definidas.
Sendo um modelo adequado para classificação de atributos discretos, o Naïve Bayes tem aplicações na análise de crédito, diagnósticos médicos ou busca por falhas em sistemas mecânicos.
É interessante saber que o Naïve Bayes é um dos modelos mais conhecidos a aplicar o conceito de probabilidade. Esse modelo, como o nome indica, faz uso do teorema de Bayes como princípio fundamental.

Máquina de Vetores de Suporte — SVM

Máquina de Vetores de Suporte (Support Vectors Machine — SVM, do inglês) é um algoritmo de aprendizagem de máquina utilizado tanto para classificação quanto para regressão.
SVM é responsável por encontrar a melhor fronteira de separação entre classes/rótulos possível para um dado conjunto de dados que sejam linearmente separáveis. Para o SVM, as diversas fronteiras de separação possíveis que são capazes de separar completamente as classes são chamadas de hiperplanos. Dessa forma, o SVM busca encontrar o melhor hiperplano para um dado data set cujas classes são linearmente separáveis.

Aperfeiçoamento dos dados

Exclusão de alguns atributos

Este procedimento se resume basicamente em remover algum atributo ou coluna desnecessárias para o aprendizado do algoritmo. Para o desenvolvimento deste trabalho, foram removidas as informações sobre o colesterol obtido através do sensor de IMC, o de glicemia em jejum e a informação se existia a possibilidade de ataque previamente conhecido.

Remoção de outliers

Um outlier é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise.

Resultados

Todo o código utilizado para o desenvolvimento deste trabalho está disponível no gitHub.

Sem tratamento de dados

Execução	Árvore de busca	Naive Bayes	SVM
1°	79.120%	84.615%	69.207%
2°	83.516%	83.813%	67.039%
3°	84.714%	83.256%	70.304%

Com tratamento de dados

Execução	Árvore de busca	Naive Bayes	SVM
1°	82.365%	85.265%	72.956%
2°	81.695%	87.582%	72.585%
3°	88.569%	84.623%	74.845%

Conclusão

Tendo os resultados em mãos, foi possível observar que o tratamento dos dados, em todos os casos foram eficientes, mostrando um ganho em acurácia, sendo o SVM o que nestes testes teve o maior aumento percentual, evoluindo em quase 5 pontos percentuais em referência sem o tratamento de dados, mesmo sendo o algoritmo que em valores absolutos, apresente a menor acurácia.
Com os resultados obtidos, fica evidente que modelos diferentes podem ser aplicados para problemas diferentes e tipos de dados diferentes, pois cada um tem a sua particularidade e seus pontos positivos e negativos.