MACHINE LEARNING COM O GOOGLE COLAB

Armando Handaya[1]
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP
Campus Guarulhos

Resumo
Este minicurso abordará o tema do Machine Learning, ou Aprendizado de Máquina, e a sua implementação será feita na plataforma do Google Colab. O termo aprendizado de máquina significa fazer o computador aprender alguma coisa com os dados que lhe forem inseridos. Ao nosso redor não faltam dados e isso nos gera informações muito úteis. A partir dessas informações podemos tirar algumas conclusões, como por exemplo, fazer uma estimativa de venda para o próximo ano ou descobrir a existência de uma correlação entre dia de semana e número de atendimentos em um determinado Pronto Socorro. Classificar um e-mail como spam com base em presença de algumas palavras ou concluir que um aluno está para evadir com base em determinadas informações pessoais e familiares. Neste minicurso vamos fazer um sobrevôo da Inteligência Artificial, que é um tema "guarda-chuva" do Aprendizado de Máquina. Abordaremos os variados tipos de aprendizado, fazendo a implementação na fase inicial de Pré-Processamento. Faremos o processo de Limpeza de dados usando a linguagem Python. Em seguida aprenderemos a parte teórica e prática de um dos Métodos de Processamento de Aprendizado de Máquina: Algoritmo do Naive-Bayes.
Palavras-chave: Aprendizado de Máquina; Inteligência Artificial; Algoritmo Naive-Bayes; Python.

1. INTRODUÇÃO

A ascensão exponencial da tecnologia transformou radicalmente a maneira como interagimos com o mundo ao nosso redor, e o campo do Machine Learning emerge como uma força propulsora nesse cenário dinâmico. Em seu âmago, o Machine Learning representa uma revolução na capacidade das máquinas de aprender e evoluir a partir de dados, proporcionando insights, previsões e automação de tarefas de maneiras antes inimagináveis. Este campo interdisciplinar, situado na interseção da ciência da computação, estatística e inteligência artificial, desvenda novas possibilidades ao capacitar sistemas a aprenderem padrões, tomar decisões e se adaptar continuamente.

À medida que nos aprofundamos nesse fascinante domínio, explorar-se-á desde os fundamentos do aprendizado de máquina até suas aplicações práticas em diversas

[1] Doutor em Engenharia Elétrica pela Universidade de São Paulo (USP). Professor do EBTT do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Guarulhos, SP, Brasil. E-mail: ahand@ifsp.edu.br.

áreas, como saúde, finanças, marketing e muito mais. Descobriremos como algoritmos de aprendizado de máquina processam informações, identificam padrões e aprimoram seu desempenho ao longo do tempo. Ao desbravar esse universo inovador, mergulharemos nos desafios éticos, nas promissoras oportunidades e na constante evolução que define o Machine Learning como uma das forças motrizes da era digital. No minicurso proposto faremos uma introdução desse tema e desvendamos as bases matemáticas do algoritmo Naive-Bayes que faz a tarefa de classificação. Esse algoritmo é um dos primeiros apresentados em Ciência de Dados para classificação de objetos.

2. APRESENTAÇÃO HISTÓRICA

A história da aprendizagem de máquina é uma narrativa empolgante que se desenrola ao longo das décadas, marcando a evolução de uma disciplina que transformou radicalmente a forma como interagimos com a tecnologia. Nos primórdios do século XX, os alicerces conceituais foram lançados, dando origem a um campo que hoje permeia todos os aspectos de nossa vida digital.

A semente inicial foi plantada na década de 1940, quando o matemático e lógico Alan Turing propôs a ideia de máquinas capazes de aprender sem intervenção humana. Seu trabalho seminal, "*Computing Machinery and Intelligence*" delineou o conceito do "*Jogo da Imitação*", que mais tarde se tornaria conhecido como o Teste de Turing. Embora sua visão tenha sido revolucionária, as limitações computacionais da época impediram a implementação prática dessas ideias.

As décadas seguintes testemunharam avanços modestos, mas fundamentais. Na década de 1950, Arthur Samuel, pioneiro na área, cunhou o termo "*machine learning*" e desenvolveu um programa capaz de jogar damas, melhorando suas habilidades com a prática. Esse marco definiu a trajetória futura, abrindo caminho para abordagens mais sofisticadas.

Nos anos 60 e 70, o campo experimentou uma explosão de teorias e experimentações, impulsionadas pelo otimismo gerado pelos avanços computacionais. No entanto, os recursos limitados de hardware e a ausência de grandes conjuntos de dados prejudicaram o progresso substancial.

O renascimento da aprendizagem de máquina ocorreu nos anos 80 e 90, com o surgimento de técnicas mais avançadas. Algoritmos de redes neurais, inspirados no

funcionamento do cérebro humano, começaram a ganhar destaque. A descoberta do algoritmo de retropropagação pelo trio de cientistas David Rumelhart, Geoffrey Hinton e Ronald Williams proporcionou um impulso crucial à viabilidade prática das redes neurais.

A virada do milênio testemunhou avanços extraordinários, impulsionados pela disponibilidade de grandes conjuntos de dados e poder computacional exponencialmente maior. A ascensão da internet e a interconectividade global abriramas comportas para a era moderna da aprendizagem de máquina. Empresas e pesquisadores começaram a explorar aplicações em campos tão diversos como reconhecimento de voz, visão computacional, recomendação de conteúdo e muito mais.

A década de 2010 marcou a consolidação da aprendizagem de máquina como uma força dominante. Algoritmos de aprendizado profundo, como as redes neurais convolucionais e as redes neurais recorrentes, lideraram avanços impressionantes em tarefas complexas. Empresas líderes de tecnologia, como Google, Facebook e Amazon, integraram algoritmos de aprendizagem de máquina em seus produtos e serviços, moldando a experiência digital diária de milhões de pessoas.

À medida que entramos na próxima década, a história da aprendizagem de máquina está longe de ser concluída. O campo continua a evoluir, impulsionado por avanços em hardware, algoritmos mais sofisticados e a crescente compreensão das nuances da inteligência artificial. À medida que nos aprofundamos neste emocionante capítulo, somos testemunhas da revolução que a aprendizagem de máquina proporciona, transformando radicalmente a maneira como enfrentamos os desafios tecnológicos e abraçamos o futuro digital.

3. TIPOS DE APRENDIZAGEM DE MÁQUINA

A aprendizagem de máquina (ML) é uma disciplina vasta e dinâmica, composta por diversos métodos que se adaptam a diferentes contextos e objetivos. Essa diversidade é essencial para enfrentar a complexidade das tarefas que a ML podeabordar. Vamos explorar os principais tipos de aprendizagem de máquina, destacando suas características distintas e aplicações específicas.

3.1 Aprendizagem Supervisionada

A aprendizagem supervisionada é um dos paradigmas mais comuns em ML. Nesse modelo, o algoritmo é treinado com um conjunto de dados rotulados, onde as entradas estão associadas a saídas desejadas. O objetivo é fazer com que o modelo generalize bem para novos dados, prevendo corretamente as saídas. Aplicações incluem reconhecimento de imagem, classificação de texto e previsão de séries temporais.

3.2 Aprendizagem Não Supervisionada

Ao contrário da supervisionada, a aprendizagem não supervisionada lida com conjuntos de dados não rotulados. O algoritmo explora padrões e relações intrínsecas nos dados sem orientação explícita. Agrupamento (clustering) e redução de dimensionalidade são exemplos de técnicas não supervisionadas. Essa abordagem é crucial em situações em que as categorias ou relações não são conhecidas a priori.

3.3 Aprendizagem Por Reforço

Na aprendizagem por reforço, um agente interage com um ambiente, tomando decisões sequenciais para maximizar uma recompensa acumulativa ao longo do tempo. O agente aprende a associar ações a resultados positivos, explorando e refinando suas estratégias com base no feedback recebido. Jogos, robótica e otimização de recursos são domínios onde a aprendizagem por reforço brilha.

3.4 Aprendizagem Semi-supervisionada

Este tipo de aprendizagem combina elementos da supervisão e da não supervisão. Uma parcela dos dados é rotulada, enquanto o restante permanece não rotulado. O modelo usa tanto a informação fornecida quanto a inferência a partir dos dados não rotulados para realizar tarefas específicas. A aprendizagem semi-supervisionada é útil quando rotular grandes conjuntos de dados é oneroso ou difícil.

Essa variedade de abordagens na aprendizagem de máquina destaca aflexibilidade e a adaptabilidade dessa disciplina. À medida que novos desafios surgem eos conjuntos de dados se tornam mais complexos, a diversidade de técnicas permite que os praticantes escolham a abordagem mais adequada para atingir seus objetivos, impulsionando assim a constante evolução do campo da aprendizagem de máquina.

4. ETAPAS NA FASE DE PRÉ-PROCESSAMENTO

O pré-processamento de dados é uma fase crucial no ciclo de vida de qualquer projeto de aprendizado de máquina. Ele representa a fundação sobre a qual modelos preditivos são construídos, influenciando diretamente a qualidade e eficácia das análises subsequentes. Vamos explorar detalhadamente as diversas etapas do pré-processamento, destacando a importância de cada uma delas.

4.1 Coleta de Dados

A primeira etapa do pré-processamento é a coleta de dados. Isso pode envolver a aquisição de dados brutos de diferentes fontes, como bancos de dados, APIs, arquivos CSV, ou até mesmo dados gerados por sensores. Garantir a integridade e relevância dos dados nesta fase é crucial para o sucesso do modelo.

4.2 Limpeza de Dados

Uma vez coletados, os dados frequentemente contêm imperfeições, como valores ausentes, outliers ou erros de medição. A limpeza de dados é a etapa em que essas imperfeições são identificadas e corrigidas. Isso pode incluir preenchimento de valores faltantes, remoção de outliers e normalização de dados.

4.3 Transformação de Dados

A transformação de dados visa ajustar a estrutura dos dados para atender melhor aos requisitos do modelo. Isso pode incluir a codificação de variáveis categóricas, a criação de variáveis dummy, a normalização de escalas e a redução de dimensionalidade. Essas transformações são essenciais para garantir que o modelo seja capaz de interpretar e extrair padrões significativos dos dados.

4.4 Seleção de Características

A seleção de características é um processo de escolha das variáveis mais relevantes para a tarefa em questão. Reduzir a dimensionalidade do conjunto de dados pode melhorar a eficiência computacional e evitar o overfitting. Métodos como análise de importância de características e algoritmos de seleção automática podem ser empregados aqui.

4.5 Tratamento de Desbalanceamento

Em tarefas de classificação, é comum encontrar conjuntos de dados desbalanceados, onde as classes têm quantidades significativamente diferentes de instâncias. O tratamento adequado desse desequilíbrio é vital para garantir que o modelo não seja tendencioso em relação à classe majoritária.

4.6 Divisão em Conjuntos de Treinamento e Teste

Para avaliar a capacidade de generalização do modelo, os dados são divididos em conjuntos de treinamento e teste. O modelo é treinado nos dados de treinamento e avaliado nos dados de teste, permitindo uma estimativa realista de seu desempenho em novos dados.

4.7 Normalização e Padronização

Normalizar e padronizar os dados garantem que todas as variáveis estejam na mesma escala, o que é crucial para algoritmos sensíveis à escala, como redes neurais e métodos baseados em distância. A normalização ajusta os dados para uma escala específica, enquanto a padronização os converte para uma escala com média zero e desvio padrão um.

Em conjunto, essas etapas formam um processo meticuloso e iterativo de preparação de dados para a modelagem de aprendizado de máquina. A qualidade do pré-processamento não apenas influencia diretamente a eficácia do modelo, mas também desempenha um papel crucial na interpretabilidade e confiabilidade dos resultados obtidos. Portanto, uma atenção cuidadosa a cada uma dessas etapas é essencial para o sucesso de projetos de aprendizado de máquina.

5. ALGORITMOS DE CLASSIFICAÇÃO

Na fase do Processamento iremos abordar, neste minicurso, apenas um dos algoritmos existentes de Classificação: o Naive-Bayes. A lista mais completa, embora não esgote todo o assunto, pode ser vista abaixo.

Naive Bayes. Baseado no Teorema de Bayes, assume independência condicionalentre os recursos. É eficaz em tarefas de classificação, especialmente em problemas de processamento de linguagem natural.

Regressão Logística. Utilizado principalmente para problemas de classificação binária, a regressão logística modela a probabilidade de uma instância pertencer a uma classe específica. Ele utiliza a função logística para transformar a saída em valores entre 0 e 1.

Máquinas de Vetores de Suporte (SVM - Support Vector Machines). Eficiente em problemas de classificação binária e múltipla, as SVMs procuram encontrar um hiperplano que maximize a margem entre as classes no espaço de características. Podemlidar com dados não lineares usando truques de kernel.

Árvores de Decisão. Estruturas hierárquicas de decisão que dividem o conjunto de dados com base em condições nos atributos. São fáceis de interpretar, mas podem serpropensas a overfitting, especialmente em árvores profundas.

K-Vizinhos Mais Próximos (K-NN - k-Nearest Neighbors). Classifica uma instância com base nas classes das instâncias vizinhas mais próximas no espaço de características. A escolha do número 'k' de vizinhos influencia a sensibilidade do modelo.

Random Forest. Uma coleção de árvores de decisão que trabalham em conjuntopara reduzir overfitting e melhorar a robustez. Cada árvore vota na classe e a classe commais votos é escolhida como a previsão final.

Gradient Boosting. Constrói modelos sequencialmente, corrigindo os erros do modelo anterior. Algoritmos populares incluem XGBoost e LightGBM, conhecidos por seu desempenho em competições de ciência de dados.

Redes Neurais. Modelos complexos inspirados na estrutura do cérebro humano. Com múltiplas camadas, as redes neurais aprendem representações hierárquicas e são adequadas para problemas complexos, como processamento de imagem e linguagem natural.

Mostraremos um caso específico de classificação de e-mails como normal ou spam com base na presença de algumas palavras previamente escolhidas. Exemplificaremos os fundamentos do algoritmo de Naive Bayes a partir desse caso e montaremos o Modelo do classificador a partir da Tabela de probabilidades. Em seguida, os mesmos passos da criação desse modelo podem ser aplicados em outros casos apresentados como exercícios.

REFERÊNCIAS

BURKOV, A.; SILVA, R. A.; ANDRADE, M. M. **The Hundred-Page Machine Learning Book em português**. 1° Ed, Editora Andly Burkov, 2019

ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise**. 1º Ed, Editora Casa do Código, 2020

FELTRIN, F. B. **Ciência de Dados e Aprendizado de Máquina: Uma abordagem prática as redes neurais artificiais**. 1° Ed, Editora Amazon Serviços de Varejo do Brasil Ltda, 2020

GÉRON, D. **Machine Learning: This Book Includes: Machine Learning for Beginners, Machine Learning with Python**. 1º Ed., Sem editor (Publicação Independente), 2019

IZBICKI, R.; MENDONÇA, T. **Aprendizado de máquina : uma abordagem estatística**. [livro eletrônico] São Carlos, SP : Rafael Izbicki, 2020. Disponível em: https://www.google.com.br/books/edition/Aprendizado_de_m%C3%A1quina_uma_abordagem_es/6O8OEAAAQBAJ?hl=pt-BR&gbpv=1 Acesso em: 13 mar 2024.

MULLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. 1ºEd, Editora O'Reilly, 2016

MUELLE, J. P.; MASSARON, L. **Machine Learning for Dummies**. 1º Ed, Editora John Wiley & Sons, 2016