DOI 10.47402/ed.ep.c2319910272

CAPÍTULO 10

INTELIGÊNCIA ARTIFICIAL E ETAPAS DO PROCESSO DE APRENDIZAGEM

Marcelo Calixto Soares da Silva
Alex Macedo de Araujo

RESUMO

Esse texto apresenta discussão no desenvolvimento de aprendizado de máquina a nível introdutório utilizado na localização de padrões, com estrutura de treinamento baseado nas práticas de mercado de cientistas de dados e estatísticos. O objetivo é mostrar que o aprendizado de máquina, uma subárea da inteligência artificial pode ser simples na prática, onde conceitos matemáticos complexos podem ser simples. São abordados algoritmos e a sequência de passos em um aprendizado, avaliar e reavaliar o processo até que um valor ideal seja atingido, assim como as etapas do algoritmo.

PALAVRAS-CHAVE: Aprendizado de máquina. Classificação; Regressão. Clusterização. Redes Neurais.

1. INTRODUÇÃO

Muito tem se falado sobre a inteligência artificial com o lançamento de tecnologias como Chat GPT e outras ferramentas que têm a capacidade de interagir e entender o ser humano naquilo que precisa de ajuda ou dúvidas, mas o aprendizado de máquina pode ser muito menos robusto do que o marketing de Big tehcs vendem, mas a forma como a máquina aprende é uma mistura de processos humanos misturados com álgebra e geometria.

Durante muitos anos o que se entendeu como o aprendizado de máquinas se devia a mídia, e ideias de que computadores começariam a se tornar seres autônomos que se pareceriam com humanos, mas a realidade que se tem atualmente são de algoritmos de predição e aprendizado no simples auxílio, o principal uso está em prever o futuro, prever vendas, prever o que o usuário está a procurar, está em auxiliar o ser humano em suas tarefas e trabalhar junto ao homem na otimização de tempo e recursos, nada daquilo de robôs com sentimentos e que buscam se parecer ou serem humanos.

Dado o modo que a ideia do aprendizado de máquina está longe do esperado pela cultura pop, se faz explicar que quase tudo que se refere ao modo que as máquinas aprendem é de conhecimento a muito tempo, algoritmos e fórmulas utilizadas no aprendizado tem origem no teorema de Thomas Bayes, que propôs uma metodologia de dado um acontecimento, prever as chances de acontecer outra situação, como por exemplo, dado que uma pessoa fuma a 50 anos, qual a probabilidade dela ter algum problema de saúde decorrente dessa situação, se utiliza então os algoritmos de probabilidade para descrever as situações futuras.

Um termo que acabou se popularizando academicamente e em áreas é de máquinas preditivas, pois a partir de dados e situações passadas podem aprender a prever situações, geralmente em pequenas escalas, as máquinas preditivas também têm a características de serem máquinas especialistas, focadas em uma tarefa, com baixa capacidade de diversificar dados.

2. FINALIDADE DO APRENDIZADO DE MÁQUINA

Dado que se entende que o aprendizado de máquina é usado como técnicas de fazer previsões e otimização dos recursos, fica-se se dúvidas, de quando os utilizar, muitos gestores se perguntam qual a necessidade e qual a efetividade terão em adotar essas tecnicas para melhorar a performance do negócio, alguns conhecimentos humanos por vezes não necessitam de máquinas treinadas para serem efetivos, um carrinho de cachorro quente em frente a um estádio em dia de jogo terá uma boa venda, pois é uma situação que deverá gerar demanda e que o conhecimento humano pode prever. Como afirma A. Ajay em Máquias Preditivas (2019, p. 13): A predição é o processo de preenchimento de informações ausentes. Ela usa informações disponíveis, geralmente chamadas de “dados”, e as usa para gerar informações que você não tem. Ou seja, o aprendizado trabalha com as incertezas, e tenta reduzir elas através de cálculos, como uma regressão logística abordada mais à frente, que dado as características de um registro, qual a probabilidade de ele pertencer a determinada classe. Com isso se consegue definir as melhores chances de classificar um usuário por exemplo ao que esperar de suas escolhas e necessidades.

2.2 MOTIVAÇÃO DO USO

Dado as carateristas do que é o aprendizado de máquina, é preciso saber quando o usar, como mencionado anteriormente, para ele ser utilizado, é necessário uma complexidade para que faça sentido. Como afirma A. Géron em Mãos a Obras: Aprendizado de Máquina com Scikit-Learn & TensorFlow (2017, p. 7). Problemas para os quais as soluções existentes exigem muita configuração manual ou longas listas de regras, Problemas Complexos para os quais não existe uma boa solução quando utilizado abordagem tradicional. Compreensão de problemas complexos e grandes quantidades de dados. Logo o uso das técnicas vai ser para encontrar padrões bem estabelecidos, sejam lineares ou não lineares, os padrões lineares têm complexidade mais baixa e podem ser demonstrados através de uma separação dimensional onde é possível a segregação por classe ou valores.

2.3 PORQUE PYTHON

Atualmente a grande massa de pessoas que trabalha com a pesquisa de aprendizado de máquina decidiu usar a linguagem de programação Python, atualmente se destacam algumas linguagens como R, Scala, e Julia, as 3 são especialistas na área de dados, Julia ainda está mais em criação e tem grandes chances de ser utilizada no futuro pelo alto desempenho, mas R e Scala estão consolidadas e são utilizadas principalmente academicamente.

Já Python se tornou queridinha pelo seu uso simples, o código é fácil de se entender e rápido de escrever, como existem bibliotecas em demasiado para a linguagem e uma comunidade muito ativa ela acabou se tornando muito popular dentro da comunidade de aprendizado de máquina, Scikit-Learn se tornou a biblioteca mais comum dentro do uso de aprendizado e a mais abrangente em número de algoritmos e tem uma grande biblioteca de exemplos, assim como se tornou base para novas bibliotecas que a utilizam como base para bibliotecas mais visuais.

Um dos pontos baixos do Python para o uso em grande escala é a velocidade de processamento mais baixa que outras linguagens, devido a isso, muitas bibliotecas são feitas em C e integradas ao Python, para unir duas coisas importantes, muita velocidade de C e código simples do Python, então se tem o melhor de dois mundos. Uma ferramenta importante dos testes é o uso de notebooks, pequenos blocos e terminais em web que servem para a testagem e códigos de uso para pesquisa científica, ele permite a livre execução de blocos de códigos sem a necessidade de executar todo código.

O aprendizado de máquina não tem apenas um resultado como saída, existem diversas vertentes e motivações para o uso, como prever valores, classificar situações e agrupar conjuntos semelhantes para entender suas semelhanças, decidir quais produtos podem ser indicados e diversas outras situações que seguem padrões de comportamento, a seguir serão detalhados tipos de aprendizados e suas características. O aprendizado de máquina leva em conta de que para aprender se possa corrigir aquilo que está incorreto, como se faz no aprendizado humano, no aprendizado supervisionado, para que se atinja a melhor performance se age de forma similar ao ser humano, se o algoritmo entende algo como “bola”, mas o objeto é um outro objeto, o rótulo propaga a correção a cada iteração do algoritmo que vai aos poucos reduzindo esse erro na identificação de padrões. No meio não supervisionado é deixado que a máquina encontre padrões que o ser humano não consegue localizar de maneira simples.

3. CLASSIFICAÇÃO

A classificação consiste em através de modelos previamente treinados utilizar os dados anteriores a fim de tentar agrupar os similares por rótulos fornecidos pelo usuário em um algoritmo de aprendizado supervisionado, para a classificação o algoritmo passa pelas variáveis dos registros anteriores e seus rótulos para descobrir em que classe o novo registro tem mais probabilidade de se enquadrar.

4. REGRESSÃO

A regressão tem como objetivo prever valores através de dados anteriores, mas ela tem como principal análise de previsão dados não temporais, embora possa ser utilizado em determinados casos, demanda que os registros sejam individuais, que não tenha dependência de um registro a outro, a regressão mais comum, a linear, tenta de maneira geométrica definir uma reta que possa passar pelos dados e tentar fazer as previsões com a menor quantidade de erro possível, existem diversos algoritmos que se utilizam a regressão, todos tem pôr fim a encontrar o melhor valor com a menor taxa de erro, nascida das pesquisas de Francis Galton que definiu a correlação, as técnicas são muito utilizadas nas áreas estatísticas. Com o tempo mais possibilidades na regressão foram descobertas, como polinomiais que se utilizam de linhas não retas para a classificação de valores, ou mesmo árvores de decisão que são mais comumente utilizadas em classificação também foram adaptadas para análise de regressão, análise de regressão.

5. CLUSTERIZAÇÃO

A Clusterização é um método de aprendizado não supervisionado, onde o desenvolvedor ou pesquisador apenas fornece os dados previamente tratados e com o uso de algoritmos tenta-se relacionar e aproximar de maneira dimensional os registros mais próximos, existem otimizadores que tentam aproximar quais são as melhores quantidade de agrupamentos e isso ainda depende da avaliação do desenvolvedor. Existem menos algoritmos de clusterização que os demais métodos, o desenvolvedor avalia quais possibilidades e os dados que passará ao algoritmo, mas nessa forma quem vai definir o rótulo de saída é o computador, a clusterização também consegue responder de maneira mais efetiva e rápida quais variáveis podem responder perguntas, quando se usa dendrogramas na avaliação dessas saídas, consegue-se perceber qual variável influencia mais ou menos o agrupamento, se há muita ou não importância para aquele grupo a variável.

6. APRENDIZADO POR REFORÇO

O aprendizado por reforço é o mais específico, ele é realizado em um ambiente controlado, ele depende do ambiente para que possa ser ensinado, o uso mais popular tem sido para jogos que a máquina passa a aprender com o modo de agir do usuário, e carros autônomos, utilizando de treinamentos, a recompensa é dada ao chegar no ponto de destino, ou seja, nesse ambiente pré-determinado e controlado, o algoritmo se adapta as restrições, e tenta calcular a melhor forma de atingir um objetivo, podendo ou não trabalhar com erros e penalizações para que determinadas ações sejam limitadas ou restringidas. Como afirma A. Géron em Mãos a Obras: Aprendizado de Máquina com Scikit-Learn & TensorFlow (2017, p. 458). No Aprendizado por reforço, um agente de software faz observações e realiza ações dentro de um ambiente e, em troca, recebe recompensas. Seu objetivo é aprender a agir de forma a maximizar suas recompensas esperadas de longo prazo.

Series Temporais Séries Temporais são dados que ao contrário dos anteriores tem uma relação um com o outro em forma de uma linha do tempo, entendendo essa forma linear, os dados demonstram determinados períodos de tempo onde se pode prever ações, demandas, questões de clima e doenças, é um ambiente ainda recente, que demanda estudos a longo prazo para a maior coleta de dados, quando mais longo for a pesquisa a respeito do assunto, melhor a qualidade preditiva, como afirma A. Nielsen Análise Prática de Séries Temporais (2016, p.13). A análise e previsão de séries temporais ainda não chegaram à sua Idade de Ouro e, por ora, a análise de séries temporais segue dominada pelos métodos estatísticos tradicionais, bem como pelas técnicas mais simples de aprendizado de máquina, como conjuntos de árvores e ajustes lineares. Ainda se espera um grande salto à frente para predizer o futuro. Ou seja, as análises embora utilizadas a décadas ainda não é tão bem estabelecida como outros tipos de aprendizados.

7. MODELOS ENSEMBLE

Modelos Ensemble são modelos relativamente novos em aplicações e sempre recebendo atualizações como foi com Xgboost lançado em 2014 e altamente utilizado em competições de aprendizado de máquina, esse tipo de modelo utiliza uma média de vários minimodelos treinados para tentar chegar a uma solução melhor que a maioria dos algoritmos clássicos, mas os utilizando de fundo para isso, possuem seus próprios métodos para chegar a um resultado ótimo. Como afirma A. Géron em Mãos a Obras: Aprendizado de Máquina com Scikit-Learn & TensorFlow p458. Suponha-se que se faça uma pergunta complexa a milhares de pessoas

aleatórias e, então, reúna suas respostas. Em muitos casos, verá que esta resposta agregada é melhor do que a resposta de um especialista, o que é chamado de sabedoria das multidões.

Como pode-se ver, o método tenta replicar mais segurança nas respostas, pois na hora de separar os dados de testes, eles podem ser uma fatia ruim de determinada base de dados, e, tentar utilizar a lógica humana de um coletivo de respostas pode ter efeitos de melhora.

8. REDES NEURAIS

Redes Neurais Redes neurais ou Deep learning como é conhecido em alguns meios, é uma subárea do aprendizado de máquina, muitas vezes mais complexo, pois tem uma estrutura variável que pode ser ajustada a várias possibilidades, a rede neural em si é um tipo de estrutura de grafo direcionado, onde existem camadas de entradas e saídas de informações, entre essas camadas existem outras, as camadas ocultas, nelas é onde é feito o cálculo e aplicadas as fórmulas do aprendizado. As redes neurais nada mais são do que uma simulação do cérebro humano, é comum entre pessoas leigas acharem que um computador é uma grande máquina de processamento, mas a realidade é que um ser humano é uma máquina com capacidade de processamento de informações enorme, máquinas são excepcionais em calcular matemática e algoritmos que levem em consideração lógica, achar padrões e resolver problemas, mas cérebros humanos são muito superiores em criatividade, em inovar e entender problemas nas áreas humanas, logo a rede neural tem se focado justamente em localizar padrões não lineares e complexos que algoritmos mais simples não conseguem localizar de forma tão eficiente. O Back propagation, algoritmo desenvolvido na década de 80 foi o principal responsável pelo aumento do uso das redes neurais, com ele, a correção das camadas se tornou algo mais simples, onde se passou a conseguir mais bem resultados nas predições, com ele os valores de erro podem ser corrigidos e reanalisados em uma um looping de iterações até que se tenha a melhor resposta possível. Outro fator muito importante para as redes neurais justamente em grande escala, é o fator de processamento paralelizado, a Nvidia, grande marca de placas gráficas tem tornado a inteligência artificial como o carro-chefe de serviços, o processamento paralelizado permite um processamento muito mais eficiente para as redes neurais, quanto maior for a complexidade melhor o processamento em placas gráficas é em relação ao uso de processadores. Existem diversas arquiteturas diferentes e serão citadas as mais utilizadas e comuns.

8.1 Redes neurais multilayer perceptron

São as mais comuns e descritas acima, elas geralmente são usadas para resolver pequenos problemas e mais comuns no meio de pesquisa e aprendizado, utilizadas para regressão e classificação. Servem para resolver os principais problemas humanos e comerciais, elas se adéquam as necessidades na previsão, alguns algoritmos tradicionais utilizam técnicas de redes neurais para definir classificações como o algoritmo SVM.

8.2 Redes neurais convolucionais

As Redes Neurais Convolucionais são redes neurais artificiais profundas que podem ser usadas para classificar imagens, agrupá-las por similaridade (busca de fotos) e realizar reconhecimento de objetos dentro de cenas. São algoritmos que podem identificar rostos, indivíduos, sinais de rua, e demais objetos.

As redes recorrentes são um poderoso conjunto de algoritmos de redes neurais artificiais especialmente úteis para o processamento de dados sequenciais, como som, dados de séries temporais ou linguagem natural. Elas tentam prever qual será a próxima palavra, ou mesmo qual a próxima venda.

Redes Neurais Generativas São redes capazes de construir e criar coisas a partir de modelos anteriores, e trabalham com duas redes adversárias que tentam melhorar competindo uma com a outra, são responsáveis por aumentar o dimensionamento de imagens para jogos, assim como de manutenção em fotos e vídeos gerando pixels faltantes, ou mesmo fazer a simulação de rostos e demais montagens.

Modelos semi supervisionados de aprendizado utilizam registros rotulados e não rotulados para a aprendizagem, um pouco mais complexos mas amplamente estudados academicamente atualmente por conseguirem aproveitar diversos recursos de registros mas que não possuem um rótulo, ele utiliza rótulos de alguns registros e analisa através deles quais as possibilidades mais próximas dos registros não rotulados serem, geralmente usam grafos ou cálculo dos vizinhos mais próximos para chegarem a esse resultado, sendo assim conseguem com poucos dados chegar a um sistema de previsão razoável e aplicável.

8.3 Redes neurais recorrentes

As redes recorrentes são um poderoso conjunto de algoritmos de redes neurais artificiais especialmente úteis para o processamento de dados sequenciais, como som, dados de séries

temporais ou linguagem natural. Elas tentam prever qual será a próxima palavra, ou mesmo qual a próxima venda.

8.4 Redes neurais generativas

São redes capazes de construir e criar coisas a partir de modelos anteriores, e trabalham com duas redes adversárias que tentam melhorar competindo uma com a outra, são responsáveis por aumentar o dimensionamento de imagens para jogos, assim como de manutenção em fotos e vídeos gerando pixels faltantes, ou mesmo fazer a simulação de rostos e demais montagens.

8.5 Modelo semi supervisionado

Modelos semi supervisionados de aprendizado utilizam registros rotulados e não rotulados para a aprendizagem, um pouco mais complexos mas amplamente estudados academicamente atualmente por conseguirem aproveitar diversos recursos de registros mas que não possuem um rótulo, ele utiliza rótulos de alguns registros e analisa através deles quais as possibilidades mais próximas dos registros não rotulados serem, geralmente usam grafos ou cálculo dos vizinhos mais próximos para chegarem a esse resultado, sendo assim conseguem com poucos dados chegar a um sistema de previsão razoável e aplicável.

9. SISTEMAS DE RECOMENDAÇÃO

Sistemas de recomendação são um dos mais importantes para o mercado corporativo, pois ele ajuda diretamente a impulsionar vendas e direcionamento a conteúdos, o varejo online o utiliza a muito tempo, pois consegue resolver a analisar vendas que um ser humano talvez não pudesse resolver, um exemplo muito grande do mercado se refere a venda das fraldas versus cerveja, se calculou que muitos pais acabavam indo comprar fraldas em mercado e levavam cerveja pela conveniência de já estarem no mercado, logo alguns mercados passaram a colocar os itens perto uns dos outros, ao contrário por exemplo de itens que estão diretamente ligados, como um sabão em pó e um amaciante, que servem ao mesmo propósito, o algoritmo de recomendação consegue localizar relações entre produtos que não são relacionados diretamente, ele pode requerer um processamento computacional alto, e ajustes manuais, é um sistema complexo de implementação, mas ajuda o marketing a reduzir custos, pois indicar exatamente o que o cliente precisa é algo que reduz a propaganda paga em larga escala e otimiza agrupamentos de interesses, na ciência social se chama isso de rede social, através de interesses em comuns, redes como o facebook acerta e muito no direcionamento de gostos do usuário, mas tem criado bolhas ideológicas e de conteúdos onde a diversidade de ideias acaba sendo

reduzida, existem diversos debates de como os sistemas de recomendação podem ser positivos ou negativos no âmbito da interação humana.

10. ALGORITMOS GENÉTICOS

Algoritmos genéticos são utilizados dentro do aprendizado de máquina para tentar fazer otimizações de algoritmos e sistemas evolutivos que tentam melhorar predições através da mudança e escolha de melhores adaptações, se utiliza simulações e escolha daqueles que se adaptam melhor ao problema que tenta resolver, tentando assim otimizar o processo.

11. COLETA DE DADOS E TREINAMENTO

Principal tarefa quando se inicia um treinamento ou qualquer pesquisa, é a de conseguir os dados, os dados por diversas vezes podem vir de locais diferentes, desde arquivos de texto, bancos de dados ou data lakes, a quantidade e a qualidade dos dados conseguidos geralmente dizem muito sobre o resultado final de qualquer análise dos dados realizada, dados podem estar desatualizados facilmente, exemplo de censo do IBGE que com mais de 12 anos de diferença entre um e outro ocasiona um déficit de projeções e correções de bases sociais.

A coleta de dados pode também se categorizar pelo momento em que as metas, expectativas e qual é o motivo e aonde se quer chegar no projeto também é definido, geralmente se trabalha com um variável, conhecido como target para parte do mercado, o target é geralmente uma nomenclatura de “y”, a saída esperada de uma função, as variáveis que dirão qual a saída.

11.2 Preparação dos dados

A preparação dos dados é a parte que em geral leva a maior parte do tempo de qualquer manipulação na criação de um modelo de aprendizado, alguns cientistas de dados ou estatísticos podem confirmar que até 90% do tempo é somente nessa parte, são etapas primordiais para se criar modelos bem-sucedidos, onde a entrada será realmente condizente com o uso de modelos.

11.3 Padronização, variáveis binárias e variáveis categóricas

Para o bom entendimento do algoritmo, as variáveis devem estar em escalas parecidas, pois os algoritmos podem entender que valores discrepantes têm importâncias diferentes, e não é a verdade em alguns casos, existem vários tipos de algoritmos que criam escalas que podem ser utilizadas para que o dado tenha um tipo de escala similar, e é importante que seja feito em todas colunas com valores contínuos, onde um valor 3 realmente maior que o valor 1, e não um tipo de classe, o método da normalização de dados somente pode ser aplicada quando essa

relevância é realmente ligada a continuidade de uma escala. Variáveis categóricas podem ser representadas de maneiras diferentes, as que usam algum tipo de escala, como alto médio e baixo, elas levam em consideração que as classes são desiguais ou hierárquicas, em alguns casos ou algoritmos somente são aceitos valores numéricos, nessa situação pode se usar escala de acordo com a hierarquia dos dados substituindo os valores por escalas inteiras. Em caso de variáveis categóricas onde não há hierarquia, é um pouco mais detalhado, pois o tratamento utilizado pode variar de acordo com a dispersão, as seleções utilizam a percepção do analista ou programador.

11.4 Ausentes

Existe um caso muito específico de situação que acontece com dados quando se trabalha com estruturas diversas e fontes, por vezes os dados podem simplesmente não existir, se o dado inexiste, o que fazer? Não há um consenso sobre isso, cada situação pode ser tratada de maneira diferente, a variável que falta, é uma informação imprescindível? Se for pode se tratar com inclusões de medidas de distribuição, é algo comum que se insira dados de mediana para se preencher alguns dados, outros dados as vezes podem ser excluídos, ou mesmo repensar a coleta, dependendo do algoritmo, e do modo que se quer trabalhar, por vezes a ausência não é um problema, um registro também pode ser parte primordial de uma abordagem, bibliotecas atuais em linguagens como python, não são favoráveis a trabalhar com valores ausentes, situações que cientistas de dados e estatísticos por vezes diferenciam em seus pensamentos no modo de trabalhar, como a abordagem aqui está enviesava ao modo de cientistas de dados por se tratar de predição com uso computacional, é mais interessante em preencher esses dados através de métodos de inputs, mas quem vai fazer a decisão da melhor forma é o analista, por vezes diversas formas podem acabar sendo satisfatórias para se adequar aos dados ausentes.

11.5 Seleção de variáveis

Uma parte muito importante de um bom treinamento é ter as melhores caracteristicas para fazer a predição, como comentado anteriormente, isso possibilita um menor uso computacional, mas como saber quais variáveis são importantes? Existem alguns métodos considerados dentro de algoritmos de melhores variaveis, são eles:

Filtrar, um conjunto de variáveis é submetido a um teste estatístico que maximiza um critério pré-definido e entrega como saída, um número menor de variáveis. Não há uma resposta do modelo de aprendizado para o processo de filtragem. O modelo recebe as variáveis escolhidas e tem que trabalhar com elas.

Embrulhar, um conjunto de variáveis é submetido a um "embrulho" onde as variáveis são filtradas e em seguida submetidas a um modelo de aprendizado de máquina. A resposta do modelo é utilizada para novamente filtrar as variáveis até que se chegue às variáveis mais relevantes. A resposta do modelo de aprendizado de máquina é utilizada como forma de melhorar a escolha das variáveis.

Embutir se diferencia dos demais métodos na forma como a seleção de variáveis e o processo de aprendizagem ocorre. Métodos de embutir não separam o aprendizado da seleção das variáveis como é feito nos métodos de embrulhar

Dentro da maioria dos algoritmos se tem modelos prontos de seleção de atributos, que mostra o número de vezes que uma variável foi utilizada, quanto mais vezes mais o modelo a entende como importância na prática da predição.

A colinearidade grosso modo, se trata de duas estruturas de dados serem tão parecidas, que uma acompanha a outra a modo que uma aumenta, a outra aumenta de maneira similar, com uma proporção que po ser igual ou até determinada faixa que pode variar de acordo com quem a observa, por vezes a faixa de entendimento pode depender do analista, mas uma coisa fica clara, com um alto custo de processamento, porque utilizar dados que dizem a mesma coisa? Não parece fazer muito sentido pensando no custo, nesse caso, quando se tem variáveis que possuem esse tipo de caracteristica, opta-se por excluir alguma a grosso modo, pois ambas já estão ali dizendo a mesma coisa, ou, técnicas mais avançadas como o PCA, que é analise de componentes principais, algoritmo muito famoso por excluir esse custo computacional, ele relaciona as colunas que tem essa ligação de colinearidade e as transforma em uma com uma média ponderada desses ítens, sendo assim a informação que poderia ser consumida repetidademente, se torma uma mais concisa e estruturada.

11.6 Escolha do modelo

O primeiro passo na escolha do modelo, é entender qual saída, qual a finalidade do que se quer como saída, a saída no geral é de um valor que se tem a resposta entre agrupamento, classificação ou regressão.

Na Classificação, se entende como um registro é algo, pode ser algum objeto, pode ser sim ou não, pode ser a probabilidade entre várias formas, mas basicamente se trata com a saída de probabilidade do estado do registro, exemplos práticos, dados características de uma determinada pessoa, levando em conta, renda, sexo, idade, qual a probabilidade de ela comprar o produto de tal loja? Terá uma saída entre 0 e 1, dependendo do corte utilizado da intenção de

efetividade, ele será classificado como propenso cliente, ou não. Pensando na forma da escolha do modelo, a separação entre duas tabelas, a tabela que representa x e uma que representa y, x é as características e y a classe de clientes já determinados, é primordial que para a classificação a saída dos registros de treinamento já estejam claras, como o algoritmo de classificação está sendo supervisionado, o analista vai indicar a máquina a que classe aquele registro é, e através de cálculo de erro o modelo se ajusta a tentar errar o mínimo possível.

Na regressão a estrutura da tabela é similar, a diferença é que o retorno é sempre único, o valor esperado de saída de acordo com as entradas, não existem classes e não há multivariação de possibilidades, apenas um valor previsto, e uma margem para mais ou menos de acordo com a média dos erros

O Agrupamento já tenta solucionar outro problema, a coluna de saída que o analista quer, ainda não existe, ou seja, será passado apenas para o algoritmo o x, o y será retornado por ele, ele usará a distância euclediana para tentar achar a similaridade não observável através de visualização simples e indicar onde estão os agrupamentos, mas dessa forma, o treinamento também pode ser mais complexo, pelo fato que geralmente os agrupamentos podem ser em vários tamanhos diferentes, o agrupamento pode ser 2 grupos, pode ter 3, 4, ou infinitas quantidades, no treinamento o usuário faz a definição e a máquina vai se adequar as características pedidas pelo analista.

11.7 Avaliação

Durante o processo de treinamento é bem comum que o processo seja feito repetidas vezes, pois nem sempre as expectativas de saída são realmente atingidas, por vezes é preciso modificar, voltar ao início e novamente treinar o modelo, é repassado e alterado parâmetros a fim de melhorar

Dentro da regressão quando se avalia o erro, geralmente o erro é definido pelo erro quadrado médio, onde o erro de cada classificação é elevado ao quadrado e revertido novamente usando a raiz quadrada e feito a média de todos registros, o motivo de ser realizado esse tipo de tarefa, é que de que erros podem ser negativos e positivos, quando se eleva um valor e depois reverte, é feita com que todos fiquem positivos, pois com médias negativas os erros não poderiam ser somados, tendo os valores médios são somados e divididos pelo total de registros, nisso se tem uma ideia da faixa real da predição da regressão, esse não é o único método de avaliação, mas o comum de ser utilizado desde estatística quando aprendizado de máquina.

A avaliação na classificação é um pouco mais abrangente, existem mais possibilidades ao invés do erro médio das classes, se uma matriz de confusão, a matriz mostra qual foi a classificação errada feita nas classes erradas e corretas, com uma matriz de relações, dessa forma é possível ter ideia se existe algum viés dentro da classificação, ou se alguma está tão próxima a outra que a classificação fica mais complicada, é um método de avaliação bem eficiente a humanos, assim como a acurácia, uma forma simplificada de se entender qual os resultados das saídas da classificação.

Ainda na classificação uma das métricas mais utilizadas é a curva ROC, que capta o treinamento de maneira eficiente, mostrando a relação entre falsos positivos e positivos ao londo de cada iteração do treinamento, essa medida é eficiente para descrever a efetividade do treinamento, quanto mais rápido é a subida da curva melhor está sendo o treinamento, quando mais perto do centro, menos o aprendizado está se adequando a necessidade.

11.7 Generalização

Quando pessoas que estão iniciando no entendimento do aprendizado de máquina, é comum que se fique focando em tentar chegar ao valor de 100% em acerto, afinal de contas, 100% parece algo que imponente e que trará a perfeição, mas a perfeição no caso do aprendizado não existe, no caso de um treinamento que chegue a 100%, a realidade é que o modelo de treinamento se ajustou tanto aos dados, que ele sempre vai acertar quando ver aqueles dados, mas na hora da recepção a novos dados, será impactado pois será considerado ineficiente.

A generalização é o quão abrangente um modelo pode ser, quando ele recebe registros novos, qual a capacidade dele prever com exatidão ou próximo a isso novos registros, por vezes é necessário reavaliar as predições e ver se modelos continuam funcionando bem, mas um algoritmo capaz de ser genérico a ponto de captar ainda oscilações nos dados, terá uma longevidade maior, um grande exemplo é uma vacina com eficácia de 50%, pode parecer ruim, mas a vacina da gripe costuma ter esse valor, mas ela consegue se adequar melhor a mudanças no vírus que uma vacina com eficacia de 95%, pois ela será tão especialista em trabalhar com uma versão do vírus, que será inútil em pouco tempo em caso de uma nova variação, e algoritmos de aprendizagem funcionam da mesma forma, a taxa de erro e grau de generalização trabalham juntas para serem abrangentes e eficientes.

12. CONSIDERAÇÕES FINAIS

Como abordado nesse texto, é possível verificar que a aplicação dos métodos mais simples de aprendizado de máquina tem uma curva de aprendizado relativamente pequeno. Alguns casos muito simples, esse método não precisa ser aplicado, por vezes dentro do banco de dados, pode-se ter uma variável que é calculada a posteriori com a mudança da linha que determina a regressão, fazendo com que não haja necessidade de novo treinamento, utilizado, por exemplo, em bancos e instituições financeiras para designar um tipo de score para um cliente, ou um limite pré aprovado, quando se precisa localizar os padrões ou escolher entre modelos mais sofisticados, a sequência de passos do aprendizado de máquina pode passar a ser mais interessante, sistemas de recomendação podem usar estruturas mais complexas como grafos, modelos robustos que demandam mais poder computacional, treinamentos que envolvem paralelismo como aprendizado baseado em leituras de imagens grandes possuem seus próprios segmentos de trilha, esse modelo de processo de aprendizado aqui abordado representa modelos simples que podem ser feitos individualmente por analistas que utilizam um computador pessoal, pode-se aplicar em produção de pequenos sistemas com um valor agregado maior, é comum que o aprendizado seja tratado de maneira fora do escopo de programadores e uma área diferenciada da pesquisa, mas é possível aplicar nos desenvolvimentos de sistemas transacionais de maneira mais tranquilas possíveis, compreendendo as características e a aplicação pragmática dos modelos matemáticos.

REFERÊNCIAS

AGRAWAL, A. **Máquinas Preditivas: A Simples Economia da Inteligência Artificial**. Rio de Janeiro: Alta Books, 2019.

ALCOFORADO, L. **Utilizando a linguagem R: Conceitos, manipulação, visualização, modelagem e elaboração de relatórios**. Rio de Janeiro: Alta Books, 2021.

AMARAL, F. **Introdução à Ciência de dados: Mineração de dados e Big Data**. Rio de Janeiro: Alta Books, 2016

BASTOS, P. **Redes Bayesianas: Aplicações em confiabilidade e no diagnóstico de perdas não técnicas**. Curitiba: Appris, 2017.

BRUCE, P. **Estatística Prática para Cientistas de Dados**. Rio de Janeiro: Alta Books, 2019.

FOURIER, J. **Ensaio filosófico sobre as probabilidades**. Rio de Janeiro: PUC Rio, 2010.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro: Alta Books, 2019.

GRUS, J. **Data Science do Zero: Primeiras regras com o Python**. Rio de Janeiro: Alta Books, 2016.

HARRISON, M. **Machine Learning: Guia de Referência Rápida**. São Paulo: Novatec, 2020.

MCKINNEY, W. **Python para Análise de Dados.** São Paulo: Novatec, 2018.

NEGRI, R. **Reconhecimento de Padrões: Um estudo dirigido**. São Paulo: Blucher, 2021.