Desafio iFood: Criando um Algoritmo K-Nearest Neighbors do Zero

Introdução

No mundo em constante evolução da ciência de dados e aprendizado de máquina, dominar algoritmos é uma habilidade essencial. O *K-Nearest Neighbors (KNN)* é um dos pilares da aprendizagem supervisionada, amplamente utilizado para classificação e regressão. Neste artigo, enfrentaremos um desafio empolgante: criar nosso próprio algoritmo de *Machine Learning K-Nearest Neighbors* do zero, sem depender de bibliotecas.

Contexto do Desafio

Imagine uma empresa de logística de entrega de alimentos, semelhante ao conhecido *iFood*. Ela mantém um registro detalhado dos clientes, incluindo o número de identificação, os valores das últimas quatro compras e a classificação do *Net Promoter Score (NPS)*. Nosso objetivo? Estimar as classificações NPS para os clientes que ainda não as forneceram, usando o modelo de *Machine Learning K-Nearest Neighbors (KNN)* sem usar nenhuma biblioteca como o *numpy, math* e outras. Para isso precisamos saber mais sobre *Net Promoter Score (NPS)* e o modelo que iremos usar que é o *K-Nearest Neighbors (KNN).* Abaixo explicarei detalhadamente sobre cada um e no final explicarei mais sobre as regras e os dados que serão usados para fazermos a nossa classificação.

Sobre o modelo K-Nearest Neighbors (KNN)

> "*Um bom vizinho é aquele que sorri para você por cima da cerca dos fundos, mas não passa por cima dela.*" -***Arthur Baer***

K-Nearest Neighbors (KNN) é um algoritmo de aprendizado de máquina que se baseia no princípio de que pontos de dados semelhantes tendem a estar próximos no espaço de características. Ele é usado para classificação e regressão, onde a ideia-chave é que objetos com características semelhantes compartilham a mesma classe ou categoria. Mas, o que é isso?

Imagine que você quer prever qual será o meu voto na próxima eleição presidencial se eu morasse no Texas, nos Estados Unidos. Se não souber mais nada sobre mim (e se tiver os dados), uma abordagem sensata seria analisar o voto dos meus vizinhos. Morando no Texas, como eu, meus vizinhos invariavelmente planejam votar em um candidato republicano, logo, "candidato republicano" também é um bom palpite para o meu voto.

Agora, imagine que você sabe mais do que minha localização geográfica (talvez saiba minha idade, minha renda, quantos filhos tenho e assim por diante.). Na medida em que meu comportamento é influenciado (ou caracterizado) por esses fatos, uma análise dos vizinhos mais próximos de mim em todas essas dimensões parece ser um indicador melhor do que uma análise de todos os meus vizinhos. Essa é a ideia central da classificação baseada nos vizinhos mais próximos.

O KNN é usado em diversas aplicações em Ciência de Dados, incluindo:

1. **Classificação de Documentos:** Pode ser usado para classificar documentos com base em seu conteúdo ou tema.
2. **Recomendação de Produtos:** Usado em sistemas de recomendação para encontrar produtos ou conteúdo semelhante aos gostos do usuário.
3. **Detecção de Anomalias:** Pode ajudar a identificar anomalias em conjuntos de dados.
4. **Previsão de Preços e Valores:** Pode ser usado para prever preços de ações, valores imobiliários, entre outros.
5. **Diagnóstico Médico:** Auxilia na classificação de pacientes com base em sintomas e histórico médico.

Características:

- **Não Paramétrico:** O KNN é um algoritmo não paramétrico, o que significa que não faz suposições rígidas sobre a distribuição dos dados. Ele se adapta aos dados disponíveis, tornando-o flexível.
- **Sensibilidade ao Tamanho do Conjunto de Dados:** A eficiência do KNN pode variar dependendo do tamanho do conjunto de dados. A complexidade computacional do KNN aumenta à medida que o conjunto de dados cresce, tornando-o mais lento em grandes conjuntos de dados.
- **Velocidade Ajustável:** A velocidade de cálculo do KNN é ajustável e depende de fatores como o tamanho do conjunto de treinamento e o valor escolhido para K. Um valor menor de K (mais vizinhos) geralmente requer mais cálculos, tornando-o mais lento, enquanto um valor maior de K (menos vizinhos) pode tornar o algoritmo mais rápido, mas potencialmente menos preciso.
- **Simplicidade:** O KNN é um algoritmo relativamente simples de entender e implementar. Sua lógica é direta, tornando-o uma excelente escolha para iniciantes em aprendizado de máquina.
- **Aplicação Versátil:** O KNN pode ser usado tanto para tarefas de classificação quanto de regressão. Na classificação, ele atribui uma classe com base na maioria dos K vizinhos mais próximos, enquanto na regressão, ele calcula a média dos valores de destino dos K vizinhos mais próximos para fazer uma previsão numérica.
- **Amplamente Conhecido e Estudado:** O KNN é um dos algoritmos mais antigos e bem estudados em aprendizado de máquina. Isso significa que há uma riqueza de recursos, tutoriais e informações disponíveis para aprender e aprimorar seu uso.

Fórmula do Algoritmo K-Nearest Neighbors (KNN)

A fórmula matemática para o algoritmo KNN envolve o cálculo da distância entre pontos no espaço de características para determinar a classe de um novo ponto. Uma das métricas de distância mais comuns usadas é a Distância Euclidiana.

A Distância Euclidiana entre dois pontos, x e y, com coordenadas xi e yi em n dimensões, é calculada da seguinte maneira:

$$d(x, y) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (xn - yn)^2)}$$

Nesta fórmula:

- d(x, y) é a distância entre os pontos x e y.
- n é o número de dimensões (ou características) do espaço.
- xn e yn são as coordenadas do ponto x e y na dimensão i.

O algoritmo KNN é um algoritmo de classificação e regressão baseado na ideia de que os pontos em um espaço de características semelhantes devem ter rótulos semelhantes. Para classificar um novo ponto, o KNN calcula a distância entre esse ponto e todos os pontos de treinamento conhecidos. Em seguida, ele seleciona os k pontos mais próximos (os "vizinhos mais próximos") com base na métrica de distância, que é frequentemente a Distância Euclidiana.

Após encontrar os k vizinhos mais próximos, o KNN determina a classe do novo ponto (ou seu valor, no caso de regressão) com base na classe predominante dos vizinhos.

Esta fórmula matemática ajuda a calcular as distâncias entre os pontos, permitindo que o KNN identifique os vizinhos mais próximos e tome decisões de classificação ou regressão com base em sua proximidade.

Lembre-se de que o KNN é um algoritmo simples e eficaz, amplamente utilizado em aprendizado de máquina para tarefas de classificação e regressão.

Escolha de K

Uma decisão crítica ao usar o KNN é a escolha do valor de K. O valor de K representa o número de vizinhos próximos que o algoritmo considera ao fazer uma previsão. Escolher um valor apropriado de K é fundamental, pois ele afeta o viés e a variância do modelo. Um valor pequeno de K pode levar a previsões instáveis e sensíveis ao ruído nos dados, enquanto um valor grande de K pode resultar em previsões mais suaves, mas potencialmente enviesadas.

Determinar o valor de K ideal geralmente envolve técnicas de validação cruzada, como a validação cruzada *k-fold*, que ajuda a avaliar o desempenho do modelo com diferentes valores de K e escolher aquele que oferece o melhor equilíbrio entre viés e variância.

Métricas de Distância

O *K-Nearest Neighbors (KNN)* pode usar várias métricas de distância, dependendo do problema e das necessidades específicas. A métrica de distância mais comum e padrão usada pelo KNN é a Distância Euclidiana. No entanto, o KNN é altamente flexível e permite a escolha de diferentes métricas de distância, como:

1. **Distância Euclidiana:** Essa é a métrica de distância padrão, como discutida anteriormente, e é a mais comumente usada.
2. **Distância de *Manhattan*:** Também conhecida como norma L1, é a soma dos valores absolutos das diferenças entre as coordenadas.

3. **Distância de *Minkowski*:** É uma métrica de distância generalizada que engloba tanto a Distância Euclidiana quanto a Distância de Manhattan. A Distância de *Minkowski* inclui um parâmetro "p" que permite ajustar a métrica de acordo com a necessidade, onde "p" é um valor real positivo.
4. **Distância Ponderada:** Além das métricas mencionadas acima, é possível aplicar pesos diferentes às diferentes características ao calcular a distância. Isso é conhecido como Distância Ponderada e pode ser útil quando algumas características são mais importantes que outras.

A escolha da métrica de distância depende da natureza do problema e dos dados. Por exemplo, a Distância de Manhattan é mais apropriada quando as características têm unidades diferentes, enquanto a Distância Euclidiana é eficaz quando as unidades são semelhantes. A seleção da métrica e a escolha do valor de "p" (para Distância de *Minkowski*) são etapas importantes ao aplicar o KNN a um problema específico.

Vantagens do K-Nearest Neighbors (KNN)

O algoritmo *K-Nearest Neighbors (KNN)* apresenta várias vantagens notáveis. Primeiramente, ele é amplamente elogiado por sua simplicidade. Sua lógica é direta e fácil de entender, tornando-o uma excelente escolha para introdução a algoritmos de aprendizado de máquina. Além disso, o KNN é não paramétrico, o que significa que ele não faz suposições rígidas sobre a distribuição dos dados. Isso torna o KNN altamente flexível e adequado para uma ampla gama de aplicações.

Outra vantagem importante do KNN é sua capacidade de se adaptar a dados com estruturas de *cluster* visíveis. Quando os pontos de dados compartilham características semelhantes, eles tendem a estar próximos uns dos outros no espaço de características. O KNN aproveita essa proximidade para fazer previsões precisas. Portanto, se os seus dados possuem agrupamentos bem definidos, o KNN pode ser uma escolha particularmente eficaz.

Desvantagens do K-Nearest Neighbors (KNN)

Apesar das vantagens, o KNN não é isento de limitações. Uma desvantagem notável é sua sensibilidade a valores atípicos (*outliers*). Pontos de dados extremos podem afetar adversamente as previsões do KNN, tornando-o menos adequado para dados com valores discrepantes. Além disso, o KNN requer armazenamento dos dados de treinamento para realizar previsões, o que pode ser desafiador quando o conjunto de treinamento é extenso.

O KNN também sofre da chamada "maldição da dimensionalidade." À medida que a dimensionalidade dos dados aumenta, o espaço de características se torna cada vez mais esparsamente povoado, o que pode levar a previsões menos precisas. Essa é uma desvantagem particularmente relevante ao lidar com dados de alta dimensionalidade.[1]

1 A visulaização do modelo está disponível em https://ml-playground.com/

O que é o Net Promoter Score (NPS)?

Agora vamos parar um pouco de falar sobre nosso modelo porque sua cabeça já deve estar a mil, vamos falar de uma coisa mais simples que é o *Net Promoter Score (NPS)* que é uma métrica de satisfação do cliente que foi desenvolvida por Fred Reichheld em 2003. É uma métrica simples, mas poderosa, que se concentra em uma pergunta-chave:

> *"Em uma escala de 0 a 10, o quanto você recomendaria nossa empresa/produto/serviço a um amigo ou colega?"*

Com base nas respostas a essa pergunta, os clientes são classificados em três categorias:

- **Promotores:** Clientes que deram notas de 9 ou 10 são considerados promotores. Eles estão altamente satisfeitos e propensos a recomendar a empresa.
- **Neutros:** Clientes que deram notas de 7 ou 8 são considerados neutros. Eles estão satisfeitos, mas não entusiásticos, e podem ou não recomendar a empresa.
- **Detratores:** Clientes que deram notas de 0 a 6 são considerados detratores. Eles não estão satisfeitos e são menos propensos a recomendar a empresa.[2]

Como Calcular o NPS?

O cálculo do NPS é simples. Após coletar as respostas dos clientes à pergunta-chave, você calcula a porcentagem de promotores e detratores em relação ao total de respostas. A fórmula é a seguinte:

NPS = (% de Promotores) - (% de Detratores)

O resultado é um número que pode variar de -100 a +100. Quanto maior o NPS, mais promotores a empresa tem em relação aos detratores.

O que o NPS Revela?

O NPS é uma métrica valiosa porque vai além da simples avaliação de satisfação do cliente. Ele fornece informações sobre o quão leais e entusiastas os clientes são em relação à empresa. Aqui estão algumas das informações que o NPS pode revelar:

1. **Lealdade do Cliente:** Empresas com NPS mais alto geralmente têm clientes mais leais e propensos a fazer compras repetidas.
2. **Potencial de Crescimento:** Um NPS positivo geralmente indica um potencial de crescimento, pois os promotores são mais propensos a trazer novos clientes.
3. **Identificação de Problemas:** O NPS permite identificar áreas de insatisfação e problemas que precisam ser resolvidos para melhorar a experiência do cliente.
4. ***Benchmarking:*** O NPS também pode ser usado para comparar o desempenho de uma empresa com o de concorrentes, ajudando a identificar áreas em que a empresa pode melhorar.

---

2 Neste link conseguimos ver uma imagem de como a métrica é simples: https://assets-global.website-files.com/633ec8b202a7496fb9c9dda9/64020daae50d701ac16ebd7a_Net%20Promoter%20Score.jpeg

Utilização do NPS

Empresas de todos os setores usam o NPS para medir o sucesso e a satisfação do cliente. Ele é frequentemente aplicado em pesquisas de satisfação, e a pontuação é acompanhada regularmente para avaliar as tendências ao longo do tempo. Com base nos resultados do NPS, as empresas podem tomar medidas para melhorar a satisfação do cliente, aprimorar produtos e serviços e fortalecer o relacionamento com os clientes.

Conclusão

O Net Promoter Score (NPS) é uma métrica fundamental para medir a satisfação do cliente e a lealdade à marca. Sua simplicidade e eficácia o tornam uma ferramenta valiosa para empresas que desejam melhorar o atendimento ao cliente e alcançar o crescimento dos negócios. Incorporar o NPS ao seu artigo pode ajudar a destacar a importância da satisfação do cliente e como o KNN pode ser usado para prever e melhorar o NPS.