Fundamentos Matemáticos dos Algoritmos de Machine Learning: K-Nearest Neighbors (KNN)

Seu Nome

21 de junho de 2025

Resumo

Este documento explora os fundamentos do K-Nearest Neighbors (KNN), um algoritmo de aprendizado supervisionado, não-paramétrico e baseado em instâncias. A análise foca nos conceitos matemáticos centrais do algoritmo: as métricas de distância e o processo de votação majoritária para classificação ou média para regressão.

1 Introdução ao K-Nearest Neighbors (KNN)

O K-Nearest Neighbors (KNN), ou K-Vizinhos Mais Próximos, é um dos algoritmos mais intuitivos do Machine Learning. Sua premissa é simples: **"Diga-me quem são seus vizinhos e eu direi quem você é."** Ele classifica um novo ponto de dados com base na classe da maioria de seus vizinhos mais próximos no espaço de atributos.

Diferentemente de algoritmos como a Regressão Logística, o KNN é **não-paramétrico**, o que significa que ele não faz suposições sobre a distribuição dos dados. Além disso, é um algoritmo **baseado em instâncias** (ou *lazy learner*), pois não "aprende"um modelo explícito a partir dos dados de treinamento. Em vez disso, ele memoriza todo o conjunto de treinamento e realiza os cálculos apenas no momento da previsão.

2 O Componente Central: Métricas de Distância

A matemática do KNN reside na definição de "proximidade". Para determinar quais são os "vizinhos mais próximos"de um novo ponto, precisamos de uma métrica para calcular a distância entre dois pontos no espaço de atributos n-dimensional.

Sejam dois pontos $P = (p_1, p_2, \ldots, p_n)$ e $Q = (q_1, q_2, \ldots, q_n)$ em um espaço de n dimensões (onde cada dimensão é um atributo). As métricas de distância mais comuns são:

2.1 Distância Euclidiana

É a distância em linha reta entre dois pontos, e a métrica mais utilizada no KNN. É calculada como a raiz quadrada da soma das diferenças quadráticas entre as coordenadas de cada ponto.

$$d(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

2.2 Distância de Manhattan

Também conhecida como "distância do táxi"ou "distância L1", ela calcula a soma das diferenças absolutas entre as coordenadas. Imagine se mover entre dois pontos em uma cidade, restrito a se mover ao longo dos quarteirões (apenas na vertical e horizontal).

$$d(P,Q) = |p_1 - q_1| + |p_2 - q_2| + \cdots + |p_n - q_n| = \sum_{i=1}^{n} |p_i - q_i|$$

2.3 Distância de Minkowski

É uma generalização das distâncias Euclidiana e de Manhattan. Ela é definida por um parâmetro $p \geq 1$.

$$d(P,Q) = \left(\sum_{i=1}^{n} |p_i - q_i|^p \right)^{1/p}$$

Casos especiais da Distância de Minkowski:

- Se $p = 1$, ela se torna a **Distância de Manhattan**.
- Se $p = 2$, ela se torna a **Distância Euclidiana**.

A escolha da métrica de distância pode impactar significativamente o desempenho do modelo, dependendo da natureza dos dados.

3 O Algoritmo em Ação

O funcionamento do KNN para classificar um novo ponto de dados X_{novo} pode ser resumido nos seguintes passos:

1. **Escolha do Hiperparâmetro K:** O usuário define o número de vizinhos (K) a serem considerados. Este é um parâmetro crucial do modelo.
2. **Cálculo das Distâncias:** O algoritmo calcula a distância entre o novo ponto X_{novo} e **todos** os pontos do conjunto de treinamento, utilizando a métrica de distância escolhida (e.g., Euclidiana).
3. **Identificação dos K Vizinhos Mais Próximos:** As distâncias calculadas são ordenadas da menor para a maior. O algoritmo seleciona os K pontos de treinamento que possuem as menores distâncias em relação a X_{novo}.
4. **Votação Majoritária (para Classificação):** Para problemas de classificação, o algoritmo verifica a classe de cada um dos K vizinhos selecionados. A classe que aparecer com mais frequência entre os vizinhos (a moda) é atribuída ao novo ponto X_{novo}. Para evitar empates, é comum escolher um valor de K ímpar para problemas de classificação binária.

5. **Média (para Regressão):** Em problemas de regressão, em vez de uma votação, o valor previsto para X_{novo} é a média (ou mediana) dos valores dos K vizinhos mais próximos.

4 Considerações Importantes

4.1 A Escolha de K

A escolha de K é um trade-off entre viés e variância:

- **K pequeno:** O modelo é muito sensível a ruídos e outliers, resultando em alta variância e uma fronteira de decisão complexa (overfitting).
- **K grande:** O modelo se torna mais robusto a ruídos, mas pode ignorar padrões locais, resultando em alto viés e uma fronteira de decisão muito simplificada (underfitting).

O valor ideal de K é geralmente encontrado por meio de técnicas de validação cruzada.

4.2 A Importância da Normalização dos Dados

Como o KNN se baseia em distâncias, atributos com escalas muito diferentes podem dominar o cálculo. Por exemplo, um atributo 'salário' (na casa dos milhares) terá uma influência muito maior na distância do que um atributo 'idade' (na casa das dezenas).

Portanto, é uma prática **essencial e obrigatória** normalizar ou padronizar os dados antes de aplicar o KNN. Técnicas como a **Normalização Min-Max** (colocando todos os atributos na escala $[0, 1]$) ou a **Padronização (Z-score)** (transformando os dados para terem média 0 e desvio padrão 1) garantem que todos os atributos contribuam de forma equilibrada para o cálculo da distância.

5 Conclusão

O K-Nearest Neighbors é um algoritmo conceitualmente simples, mas poderoso para tarefas de classificação e regressão. Sua fundamentação matemática está centrada no cálculo de distâncias, o que o torna sensível à escala dos dados e à "maldição da dimensionalidade"(seu desempenho degrada em espaços com muitos atributos). Apesar de seus custos computacionais na fase de predição, sua simplicidade, interpretabilidade e ausência de premissas sobre os dados o mantêm como uma ferramenta valiosa no arsenal de qualquer cientista de dados.