Relatório Técnico: Implementação e Análise do Algoritmo k-Nearest Neighbors (kNN) aplicado ao Instagram

Modelagem Preditiva de Influência em Redes Sociais: Análise de Instagram com k-Nearest Neighbors

Nome dos Residentes:

Onécio Araujo Ribeiro
Janailton Ferreira da Costa

Data de Entrega: 16 de novembro de 2024

Resumo

Documentação do desenvolvimento do modelo preditivo utilizando o algoritmo k-Nearest Neighbors (kNN) para análise de dados do Instagram. O objetivo é prever o influence score de influenciadores com base em variáveis como número de seguidores e engajamento. A metodologia envolve análise exploratória de dados, transformação de variáveis categóricas e ajuste de hiperparâmetros.

Introdução

O crescente uso das redes sociais tornou crucial a análise dos padrões de influência dos usuários. Este projeto se concentra na aplicação do algoritmo kNN para modelagem preditiva do influence score de influenciadores do Instagram. O uso do kNN é justificado por sua simplicidade e eficiência em problemas supervisionados. Explora-se um conjunto de dados do Instagram, que inclui características como posts, followers, avg_likes, entre outros.

Metodologia

Análise Exploratória

Iniciamos com uma análise exploratória para entender as características dos dados do Instagram. Observou-se a relação entre variáveis como followers e avg_likes, assim como o efeito do 60_day_eng_rate no engajamento. Utilizamos visualizações como mapas de calor para identificar correlações significativas.

Implementação do Algoritmo

O kNN foi implementado utilizando a biblioteca Scikit-Learn. A variável country foi transformada em categoria numérica baseada em continentes (por exemplo, América do Sul: 1-9), para refletir padrões geográficos. Esta transformação facilita a inclusão de variáveis categóricas dentro do modelo.

Validação e Ajuste de Hiperparâmetros

Aplicou-se validação cruzada para assegurar a consistência do modelo. Utilizamos o GridSearchCV para otimizar os valores de k e escolher a métrica de distância (Euclidiana ou Manhattan) que melhor se adequa aos dados. O processo de otimização indicou o valor de k e a métrica que resultaram na melhor performance do modelo.

Resultados

Métricas de Avaliação

Após a implementação e otimização do modelo kNN, foram utilizadas várias métricas para avaliar o desempenho preditivo:

Erro Absoluto Médio (MAE): Indicou a média dos erros absolutos entre as previsões e os valores reais do influence score.

Erro Quadrático Médio (MSE): Forneceu a média dos quadrados dos erros, penalizando de forma mais severa grandes discrepâncias.

Raiz do Erro Quadrático Médio (RMSE): Uma medida derivada do MSE que preserva a mesma unidade dos dados originais, proporcionando interpretação mais intuitiva.

Os resultados mostraram uma boa precisão e consistência do modelo, especialmente após o ajuste dos hiperparâmetros. A normalização dos dados contribuiu positivamente para a eficiência do algoritmo, destacando a importância de um pré-processamento robusto.

Visualizações

As visualizações criadas no projeto destacaram diversos insights significativos:

Gráficos de Dispersão: Revelaram correlações fortes entre followers e avg_likes, auxiliando na justificativa das características incluídas no modelo.

Gráficos de Barras: Mostraram comparações perspicazes entre rank e influence_score, visualizando tanto a distribuição dos dados quanto a acurácia das previsões.

Mapas de Calor de Correlação: Ajudaram a identificar variáveis altamente correlacionadas que podem influenciar o desempenho do modelo.

Esses gráficos não só contribuiram para uma análise visual abrangente, mas também validaram a seleção de características feitas pelo modelo.

Discussão

A análise revelou que o modelo kNN, embora simples, pode prever efetivamente a influência de contas Instagram por meio de dados públicos disponíveis. Contudo, o modelo é sensível a variáveis que não foram contempladas nesta versão, como conteúdos não numéricos e padrões de comportamento específicos de cada usuário. A escolha do k e da métrica de distância impactou diretamente o desempenho, indicando a importância do ajuste de hiperparâmetros.

Conclusão e Trabalhos Futuros

O projeto concluiu que o algoritmo kNN é uma ferramenta eficaz para prever scores de influência com base em variáveis públicas disponíveis no Instagram. O modelo demonstrou uma capacidade robusta de análise de dados não lineares e revelou-se flexível em adaptar-se a mudanças de estrutura de dados com eficaz pré-processamento.

Principais aprendizagens incluem:

- Importância de transformações adequadas para dados categóricos.
- Necessidade de normalização e ajuste de escalas para maximizar a eficiência algorítmica.
- Impacto significativo da escolha de parâmetros de modelagem, como o número de vizinhos e a métrica de distância.

Trabalhos Futuros

Para futuras iterações do projeto, várias avenidas de aprimoramento e expansão podem ser exploradas:

- **Integração de Dados Qualitativos:** Incluir análises de sentimentos ou tópicos de postagem para enriquecer o conjunto de dados com informações qualitativas.
- **Exploração de Algoritmos Avançados:** Investigação de algoritmos de aprendizagem não-supervisionada ou métodos de deep learning que poderiam capturar complexidades mais sutis nos dados.
- **Ampliar o Escopo Global:** Incorporar uma categorização mais completa para países de outros continentes (como Ásia e África) para análises mais abrangentes.
- **Análise Temporal:** Introdução de elementos temporais para modelar a evolução do influence score ao longo do tempo.
- **Interatividade do Modelo:** Desenvolvimento de interfaces de usuário que permitam a exploração de cenários "e se", ajudando gestores de mídias sociais a ajustarem suas estratégias baseadas em previsões automáticas.

Além disso, é benéfico realizar comparações de desempenho entre o kNN e outros algoritmos de aprendizado de máquina para reafirmar a escolha do modelo para esta aplicação. Um estudo aprofundado do impacto de combinações de variáveis diferentes também pode fornecer insights adicionais sobre a estrutura de dados mais eficaz para previsões de influenciadores de redes sociais.

Referências

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

Cunningham, P., & Delany, S. J. (2007). k-Nearest Neighbour Classifiers. University College Dublin.