Relatório do Projeto: Aplicação de Naive Bayes para Filtragem de Spam

1. Introdução

Este relatório detalha o desenvolvimento de um modelo de aprendizado de máquina para a classificação de e-mails como spam (indesejados) ou não spam (ham, desejados). O objetivo principal deste projeto é construir um modelo eficaz capaz de automatizar essa classificação, melhorando a experiência do usuário e a segurança online.

2. Algoritmo Naive Bayes

2.1 Fundamentos Teóricos

O algoritmo Naive Bayes é uma técnica de classificação probabilística baseada no Teorema de Bayes, com uma forte (ingênua) suposição de independência entre as características. Em termos simples, o Naive Bayes assume que a presença ou ausência de uma característica específica em um documento é independente da presença ou ausência de qualquer outra característica, dado que a classe do documento é conhecida.

O Teorema de Bayes é expresso pela seguinte fórmula:

$$P(classe|características) = \frac{P(características|classe)*P(classe)}{P(características)}$$

Onde:

- P(classe|características) é a probabilidade a posteriori da classe dado as características (o que queremos calcular).
- P(características|classe) é a probabilidade das características dado a classe.
- P(classe) é a probabilidade a priori da classe.
- P(características) é a probabilidade a priori das características.

2.2 Aplicação no Projeto

Neste projeto, foi utilizado o **Multinomial Naive Bayes**, usado para dados discretos, como a frequência de palavras em um texto. É adequado para classificação de texto, onde as características representam a contagem de ocorrências de palavras. Dada a natureza dos dados (frequência de palavras em e-mails). A escolha deste algoritmo se justifica pela sua eficácia em problemas de classificação de texto e sua capacidade de lidar com grandes volumes de dados.

2.3 Vantagens e Desvantagens

Vantagens:

- Simplicidade e facilidade de implementação.
- Eficiência computacional, sendo rápido para treinar e classificar.
- Bom desempenho em problemas de classificação de texto.
- Lida bem com características irrelevantes.

Desvantagens:

- A suposição de independência entre as características raramente é verdadeira na prática.
- Pode atribuir probabilidade zero a uma classe se uma característica não for observada no conjunto de treinamento (problema de frequência zero).

3. Metodologia

3.1 Conjunto de Dados

O conjunto de dados utilizado é o Spambase, disponível no UCI Machine Learning Repository. Este conjunto de dados contém informações sobre 4601 e-mails, cada um descrito por 57 características. A variável alvo é binária, indicando se um e-mail é spam ou não.

3.2 Pré-processamento dos Dados

1. **Carregamento dos Dados:** Os dados foram carregados utilizando a biblioteca Pandas, atribuindo nomes às colunas de acordo com a descrição do conjunto de dados.
2. **Separação de Características e Variável Alvo:** As características (X) foram separadas da variável alvo (y), onde X contém as colunas de `Feature_0` até `Feature_56`, e y contém a coluna `is_spam`.
3. **Vetorização do Texto:** O CountVectorizer foi utilizado para converter as características de texto em uma matriz de contagens. Cada linha das características X foi convertida em uma string e então vetorizada.
4. **Divisão dos Dados:** Os dados foram divididos em conjuntos de treinamento (80%) e teste (20%) para avaliar o desempenho do modelo. A divisão foi estratificada para manter a proporção das classes nos conjuntos de treinamento e teste.

3.3 Treinamento do Modelo

O modelo Multinomial Naive Bayes foi instanciado e treinado utilizando o conjunto de treinamento. O treinamento envolveu o cálculo das probabilidades condicionais

de cada palavra dado as classes (spam ou não spam) e as probabilidades a priori das classes.

3.4 Avaliação do Modelo

O desempenho do modelo foi avaliado no conjunto de teste utilizando as seguintes métricas:

- **Acurácia:** Proporção de classificações corretas.
- **Precisão:** Proporção de e-mails classificados como spam que realmente são spam.
- **Recall:** Proporção de e-mails spam que foram corretamente classificados como spam.
- **F1-score:** Média harmônica entre precisão e recall.
- **Matriz de Confusão:** Visualização do desempenho do modelo em termos de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

4. Resultados

Os resultados da avaliação do modelo são apresentados abaixo:

	precision	recall	f1-score	support
0	0.68	0.74	0.71	531
1	0.60	0.51	0.55	390
accuracy			0.65	921
macro avg	0.64	0.63	0.63	921
weighted avg	0.64	0.65	0.64	921

A matriz de confusão visualizada no projeto também fornece uma representação clara do desempenho do modelo em cada classe.

5. Discussão

Os resultados mostram que o modelo alcançou uma acurácia de 65% na classificação de e-mails como spam ou não spam. A precisão de 60% para a classe spam indica que o modelo tende a classificar alguns e-mails não spam como spam, enquanto o recall de 51% sugere que ele perde uma proporção significativa de e-mails spam.

A matriz de confusão oferece insights adicionais sobre os tipos de erros que o modelo está cometendo, permitindo identificar áreas onde ajustes podem ser feitos.

6. Conclusão

Este projeto demonstrou a aplicação do algoritmo Naive Bayes para a filtragem de spam. O modelo desenvolvido alcançou um desempenho satisfatório. Trabalhos futuros podem incluir a exploração de técnicas de pré-processamento mais avançadas, a utilização de outros algoritmos de classificação ou a combinação de modelos para aumentar a eficácia da filtragem de spam.