OFICINA 6: Desenvolvendo um sistema de classificação

Aluno: Herik Douglas Oliveira Reinaldo

Imagine que você foi contratado por uma empresa de tecnologia para trabalhar em um projeto desafiador: melhorar o sistema de e-mail da empresa. Atualmente, muitos e-mails indesejados (spam) estão passando despercebidos pelo filtro atual, causando transtornos para os funcionários. Sua missão é desenvolver um sistema de classificação mais eficiente que consiga diferenciar com precisão os e-mails "spam" dos "não spam".

Você possui acesso a um banco de dados de e-mails antigos, já rotulados como "spam" ou "não spam", e seu chefe pediu para que você proponha uma solução usando aprendizado supervisionado. Com base no que você aprendeu sobre algoritmos de classificação, qual você escolheria para este desafio e por quê? Explique como o algoritmo funciona, por que ele seria eficaz para esse problema específico e qualquer outra consideração relevante que deva ser levada em conta ao implementar a solução. Para ajudá-lo a se guiar na sua resposta, veja alguns aspectos que você pode levar em consideração:

1. **Contextualização: Explicar o cenário do problema, identificando a necessidade de uma solução de classificação eficiente.**

   Em um ambiente corporativo, o recebimento constante de e-mails indesejados (spam) pode causar perda de produtividade, aumentar riscos de segurança e prejudicar a comunicação interna. O filtro atual da empresa não está conseguindo detectar eficientemente esses spams, o que torna essencial o desenvolvimento de um sistema de classificação automática mais eficaz. Como temos um banco de dados de e-mails já rotulados como "spam" e "não spam", podemos aplicar aprendizado supervisionado para treinar um modelo capaz de identificar com precisão os e-mails indesejados.

2. **Escolha do Algoritmo: Indicar claramente o algoritmo escolhido.**

   O algoritmo escolhido é o Naive Bayes, mais especificamente o Multinomial Naive Bayes, amplamente utilizado em problemas de classificação de texto, como a detecção de spam.

3. **Justificativa Técnica: Explicar como o algoritmo funciona, detalhando sua aplicabilidade ao problema de classificação de e-mails.**

   O Naive Bayes é um algoritmo probabilístico baseado no Teorema de Bayes, que calcula a probabilidade de um e-mail pertencer a uma determinada classe (spam ou não spam) com base nas palavras que ele contém. Apesar do nome "naive" (ingênuo), ele é muito eficaz porque assume que todas as palavras de um e-mail são independentes entre si**,** o que simplifica muito os cálculos. O modelo aprende, a partir dos dados rotulados, a frequência com que certas palavras aparecem em e-mails

spam e não spam, e usa essas informações para prever a categoria de e-mails novos.

4. **Adequação ao Problema: Justificar por que o algoritmo é adequado, considerando aspectos como eficiência, facilidade de implementação e possíveis limitações.**

   O Naive Bayes é adequado para o problema pois é rápido para treinar e prever, mesmo com grandes volumes de dados, outra característica que determina sua boa eficiência é a pouca necessidade de pré-processamento pois ele funciona bem com dados vetorizados por frequência de palavras, e-mails geralmente contém um número razoável de palavras-chave que ajudam na classificação. Sua facilidade de implementação e interpretação o destaca para projetos com prazos curtos ou recursos limitados.

5. **Considerações Adicionais: Incluir qualquer consideração adicional relevante ao contexto do problema, como a natureza dos dados ou a complexidade computacional.**

   No pré-processamento desse modelo ações como limpeza dos textos, remoção de stopwords, pontuação e transformação em vetores numéricos são cruciais para o bom funcionamento. Outro ponto a se considerar é a validação cruzada (validar com diferentes conjuntos de teste) para garantir que o modelo se generalize bem para e-mails novos.