Classificação de Spams: Uma Aplicação de Conceitos Básicos de Probabilidade no Ensino Médio

Lucas de Oliveira Silva [1]
Areli Mesquita da Silva [2]
Joelson da Cruz Campos [3]

INTRODUÇÃO

O ensino de matemática no Brasil tem sido um desafio constante, especialmente em conceitos abstratos como a probabilidade. Muitas vezes, a abordagem tradicional se limita à memorização de fórmulas e à resolução de exercícios fora da realidade, o que contribui para a desmotivação dos alunos e a falta de compreensão sobre a relevância do tema. As diretrizes curriculares, como os Parâmetros Curriculares Nacionais (PCN) (BRASIL, 1997) e a Base Nacional Comum Curricular (BNCC) (BRASIL, 2018), solicitam abordagens mais contextualizadas, que relacionem os conceitos matemáticos com o cotidiano e que preparem o estudante para um mundo com uma enorme quantidade de dados e consequentemente de incertezas.

Nesse cenário, é crucial a adoção de metodologias ativas e inovadoras que transformem a sala de aula em um ambiente de descoberta e aplicação. O uso de ciência de dados e aprendizagem de máquina (*machine learning*) pode ser uma poderosa ferramenta para atingir esse objetivo, tornando o ensino de probabilidade mais concreto e atrativo. Uma proposta pedagógica nesse contexto é a construção de um classificador de spam utilizando o método de Naive Bayes.

O Naive Bayes é um algoritmo de classificação que se baseia no Teorema de Bayes, um conceito fundamental da probabilidade condicional. A sua aplicação para a detecção de spam é um exemplo clássico e intuitivo, pois o problema de classificar um *e-mail* como spam ou não-spam pode ser facilmente compreendido pelos alunos. O ponto central do método está em calcular a probabilidade de um *e-mail* ser spam, dado que certas palavras estão presentes nele. Ao conectar a matemática com a tecnologia, o ensino

[1] Graduando do Curso de Estatística da Universidade Federal de Campina Grande - UFCG, lucas.oliveira1@estudante.ufcg.edu.br;
[2] Professora Orientadora: Professora da Universidade Federal de Campina Grande - UFCG, arelimesquita@uaest.ufcg.edu.br;
[3] Professor Orientador: Professor da Universidade Federal de Campina Grande - UFCG, joelson.cruz@professor.ufcg.edu.br;

de probabilidade se torna mais atrativo, envolvente e, acima de tudo, eficaz, preparando os alunos não apenas para o vestibular, mas para os desafios de um mundo cada vez mais orientado por dados e que norteiam atualmente a tomada de decisões.

METODOLOGIA

Esse trabalho foi desenvolvido em uma iniciação científica vinculada ao Programa de Educação Tutorial (PET) de Matemática e Estatística da Universidade Federal de Campina Grande (UFCG). Para sua realização houve encontros semanais com o propósito de estudar os principais conceitos relacionados à Teoria das Probabilidades, tomando como referência Meyer (2003). Posteriormente, elaborou-se a fundamentação teórica necessária à construção de um classificador de spams por meio do método Naive Bayes, conforme Grus (2016), bem como sua aplicação a um conjunto de dados reais.

REFERENCIAL TEÓRICO

Conforme mencionado, neste trabalho, propõe-se uma abordagem alternativa, em que os conceitos de probabilidade são introduzidos a partir de um problema real de interesse geral: a classificação de *e-mails* como spam ou não-spam. Tal proposta tem como objetivo promover uma aprendizagem significativa, contextualizando os conceitos teóricos de probabilidade e conduzindo o estudante, de forma gradual, até a construção de um classificador probabilístico baseado no Teorema de Bayes.

A introdução ao conceito de experimento aleatório pode ser feita considerando o processo de recebimento de *e-mails* em uma caixa de entrada. Cada novo *e-mail* pode ou não ser classificado como spam, configurando um evento incerto (aleatório). Assim, o conjunto de todos os *e-mails* possíveis de serem recebidos constituem o espaço amostral, Ω, enquanto que cada possível classificação (spam ou não-spam) representa um evento.

Essa analogia permite que o estudante compreenda de maneira intuitiva a natureza aleatória dos fenômenos probabilísticos, estabelecendo a base conceitual necessária para o desenvolvimento posterior dos modelos. Nesse contexto, a classificação inicial dos *e-mails* pode ser tratada como aleatória, de modo a introduzir a ideia de variabilidade e incerteza.

A partir dessa motivação, é possível introduzir a definição axiomática de probabilidade, conforme proposta por Kolmogorov (1933), e aplicá-la a um conjunto de

dados hipotético composto por *e-mails* previamente classificados. A Tabela 1 apresenta um conjunto rotulado de *e-mails* que será a base para a classificação de novos *e-mails*.

Tabela 1: Conjunto de *e-mails* rotulados (base para treino).

E-mail	Conteúdo	Classificação
1	"ganhe dinheiro fácil"	Spam
2	"oferta imperdível grátis"	Spam
3	"reunião amanhã escritório"	Não spam
4	"projeto relatório prazo"	Não spam
5	"ganhe prêmios grátis"	Spam
6	"reunião projeto cliente"	Não spam

Com base nesse conjunto de dados, é possível propor questionamentos iniciais, tais como:

- Qual a probabilidade de um *e-mail* escolhido ao acaso ser classificado como spam? $P(Spam) = n(Spam)/n(\Omega) = 3/6 = 1/2$.
- Qual a probabilidade de um *e-mail* conter a palavra "ganhe" e ser spam?

$$P(\{Ganhe\} \cap Spam) = n(\{Ganhe\} \cap Spam)/n(\Omega) = 2/6 = 1/3.$$

Essas perguntas permitem aplicar a definição clássica de probabilidade e introduzir, de maneira contextualizada, o conceito de frequência relativa e de probabilidade empírica (ou seja, abordagem frequentista).

O passo seguinte consiste em introduzir o conceito de probabilidade condicional, fundamental para a compreensão dos classificadores probabilísticos. A partir da Tabela 1, pode-se calcular, por exemplo, a probabilidade de um *e-mail* conter a palavra "ganhe" dado que é spam, ou a probabilidade de conter a palavra "loteria" dado que não é spam.

- $P(\{ganhe\}|spam) = P(\{Ganhe\} \cap Spam)/P(Spam) = 2/3.$
- $P(\{loteria\}|spam) = P(\{loteria\} \cap Spam)/P(Spam) = 0.$

Esses cálculos favorecem a compreensão da relação entre eventos e permitem discutir a noção de dependência. A partir desse ponto, pode-se destacar a importância da independência condicional, conceito central no modelo Naive Bayes, que assume que as características (no caso, as palavras) são independentes entre si, condicionadas à classe (spam ou não-spam). Matematicamente, essa suposição pode ser expressa da seguinte forma: sejam $\{w_1, w_2, \dots, w_n\}$ as palavras que caracterizam um *e-mail*, e C a classe à qual ele pertence. O modelo independência condicional nos diz que:

$$P(\{w_1, w_2, \dots w_n\}|C) = \prod_{i=1}^{n} P(w_i|C).$$

Essa formulação simplifica o cálculo da probabilidade conjunta, pois evita a necessidade de estimar a probabilidade de ocorrência simultânea de todas as combinações de palavras. Em um conjunto de *e-mails* reais, com vocabulários possivelmente compostos por milhares de termos, o número de combinações possíveis seria extremamente elevado, tornando o cálculo direto inviável.

Assim, ao assumir a independência condicional das palavras dado a classificação do *e-mail*, temos como objetivo classificar um novo *e-mail* dado o conjunto de palavras do mesmo. Em outras palavras, devemos calcular $P(C|\{w_1, w_2, \dots w_n\})$ e verificar o que é mais provável de ocorrer e isso corresponde ao classificador de Bayes. Notemos que:

$$P(C|\{w_1, w_2, \dots w_n\}) = \frac{P(C \cap \{w_1, w_2, \dots w_n\})}{P(\{w_1, w_2, \dots w_n\})} \propto P(C) \prod_{i=1}^{n} P(\{w_i\}|C).$$

Essa simplificação, embora baseada em uma hipótese forte, permite que o classificador Naive Bayes seja computacionalmente eficiente e apresente bom desempenho em tarefas práticas de filtragem de spam.

Na aplicação prática de classificadores probabilísticos, um desafio recorrente consiste em lidar com palavras que não aparecem na base de treinamento, como foi o caso da palavra loteria mencionada anteriormente. O surgimento de uma palavra desse tipo faz com que $P(C|\{w_1, w_2, \dots w_n\}) = 0$, para ambas as classes. Para contornar essa limitação, utiliza-se a suavização de Laplace que evita que probabilidades nulas sejam atribuídas a eventos não observados.

A introdução dessa técnica pode ser feita de forma didática, ressaltando que, ao observar uma nova palavra, mesmo que não esteja presente nos *e-mails* anteriores, o modelo deve atribuir-lhe uma probabilidade mínima, preservando a coerência do cálculo probabilístico e o foco apenas na classificação mais provável. Essa discussão também abre espaço para refletir sobre as limitações dos modelos probabilísticos, como a suposição de independência e a sensibilidade ao tamanho e à representatividade da amostra. Matematicamente a suavização de Laplace nos diz que

$$P(\{w_i\}|C) = \frac{Número\ de\ ocorrências\ da\ palavra\ da\ classe\ C\ +\ 1}{Total\ de\ palavras\ na\ classe + Número\ de\ palavras\ únicas}.$$

Considerando os *e-mails* rotulados na Tabela 1 observamos que o número total de palavras únicas é 14, onde na classe spam temos 7 palavras únicas e na classe não-spam também temos 7 palavras. Dado um *e-mail* com as palavras oferta, grátis e resgatar,

podemos então nos questionar se o mesmo deve ser classificado como spam ou não-spam. Notemos então que, considerando o *e-mail* com as palavras oferta, grátis e resgatar como o evento $A = \{oferta, grátis, resgatar\}$ e denotando por S o evento ser spam temos:

$$P(S|A) \propto P(S)P(\{oferta\}|S)P(\{grátis\}|S)P(\{resgatar|S)$$
$$= \frac{1}{2}\frac{2}{23}\frac{3}{23}\frac{1}{23} = \frac{6}{24334}$$

e

$$P(\bar{S}|A) \propto P(\bar{S})P(\{oferta\}|\bar{S})P(\{grátis\}|\bar{S})P(\{resgatar|\bar{S})$$
$$= \frac{1}{2}\frac{1}{23}\frac{1}{23}\frac{1}{23} = \frac{1}{24334}.$$

Como $P(S|A) > P(\bar{S}|A)$, segue então esse *e-mail* é classificado como spam.

RESULTADOS E DISCUSSÕES

Na prática, a implementação manual do algoritmo de Naive Bayes é inviável, especialmente quando se trabalha com bases de dados extensas e com múltiplos atributos. Esse processo envolve o cálculo de inúmeras probabilidades condicionais e combinações de variáveis, tornando o procedimento extremamente trabalhoso e suscetível a erros. Por esse motivo, a utilização de *softwares* estatísticos ou linguagens de programação como R ou Python garantem resultados mais rápidos, precisos e reprodutíveis.

Como aplicação prática, utilizamos a base *Spambase*, disponibilizada pela UCI *Machine Learning Repository*, composta por 4.601 *e-mails* em inglês, classificados como spam (1) ou não-spam (0). Cada *e-mail* é descrito por 57 atributos numéricos, representando frequências de palavras e caracteres especiais, além de medidas relacionadas a letras maiúsculas. O objetivo é aplicar o Naive Bayes para prever automaticamente se um novo *e-mail* é spam ou não-spam.

No R, o procedimento foi o seguinte: primeiramente, os dados foram carregados diretamente do repositório *online* e armazenados em um *data frame*. Em seguida, foram atribuídos nomes às variáveis, representando as frequências de palavras, caracteres e medidas de capitalização. A variável alvo (“spam”) foi convertida para o tipo fator, requisito necessário para o reconhecimento das categorias de classificação.

Para garantir resultados reprodutíveis, foi definida uma semente aleatória (set.seed(123)), e o pacote caret foi utilizado para dividir o conjunto de dados em 70%

para treino e 30% para teste. O modelo foi então ajustado com a função naiveBayes() do pacote e1071, utilizando todas as variáveis explicativas para prever a variável-alvo. Após o treinamento, o modelo foi aplicado ao conjunto de teste por meio da função predict(), gerando previsões de classificação para cada *e-mail*.

Em seguida, foi construída uma matriz de confusão com a função confusionMatrix() do pacote caret. Essa matriz apresenta os acertos e erros de classificação, além de métricas como acurácia, sensibilidade e especificidade.

A acurácia obtida foi de 72,3%, o que indica que o modelo classificou corretamente cerca de 72% dos *e-mails*. Esse resultado é considerado satisfatório, tendo em vista que, a priori, uma classificação aleatória teria aproximadamente 50% de acertos. Assim, o algoritmo Naive Bayes demonstrou ser uma ferramenta útil para detecção automática de spams, mesmo sendo um método simples e de fácil implementação computacional.

CONSIDERAÇÕES FINAIS

O estudo do classificador Naive Bayes mostrou-se uma ferramenta útil para aproximar conceitos de probabilidade à realidade dos estudantes do ensino médio. A resolução manual de um exemplo simples favoreceu a compreensão da lógica probabilística por trás das decisões, enquanto a aplicação prática com a base *Spambase* da UCI mostrou a relevância desses conceitos no tratamento de problemas reais, como a filtragem de *e-mails*. Assim, essa abordagem contribuiu não apenas para o ensino de probabilidade de forma contextualizada, mas também para despertar o interesse dos alunos em aplicações tecnológicas.

REFERÊNCIAS

BRASIL. **Parâmetros Curriculares Nacionais (PCN)**: Ensino Médio. Brasília: MEC/SEF, 1997.

BRASIL. **Base Nacional Comum Curricular (BNCC)**. Brasília: MEC, 2018.

GRUS, Joel. Data science do zero: primeiras regras com o Python. Rio de Janeiro: Alta Books, 2016.

MEYER, Paul L. Probabilidade: aplicações à estatística. Tradução de Ruy de Carvalho Bergström Lourenço Filho. 2. ed. Rio de Janeiro: LTC, 2003.