Modelagem Matemática e Treinamento Supervisionado para Classificação de Culturas Bacterianas

Autor: Matheus Pullig Soranço de Carvalho

April 18, 2025

Abstract

Este artigo apresenta a modelagem matemática do crescimento bacteriano e o desenvolvimento de um modelo de aprendizado de máquina supervisionado para classificar culturas bacterianas com base em suas características físico-químicas e ambientais. A modelagem inclui tanto equações diferenciais ordinárias para representar a dinâmica populacional, quanto a representação geométrica e estatística dos dados em um pipeline de aprendizado supervisionado. São apresentadas soluções analíticas e numéricas para os modelos propostos, bem como a derivação matemática completa dos algoritmos de classificação.

1 Introdução

O crescimento bacteriano é influenciado por fatores como temperatura, pH, disponibilidade de nutrientes e oxigênio. A modelagem dessas dinâmicas pode fornecer uma base matemática sólida para classificação de culturas com técnicas de aprendizado de máquina. Neste trabalho, exploramos tanto a fundamentação teórica quanto as implementações práticas desses modelos, com ênfase na interpretação matemática dos resultados.

2 Modelagem Matemática do Crescimento Bacteriano

2.1 Modelo Logístico Clássico

O crescimento populacional pode ser descrito pela equação logística:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \tag{1}$$

onde:

- $N(t)$: população bacteriana no tempo t (unidades: UFC/mL);
- r: taxa de crescimento intrínseca (unidades: h^{-1});
- K: capacidade de suporte do meio (unidades: UFC/mL).

2.1.1 Solução Analítica

A equação logística possui solução analítica conhecida:

$$N(t) = \frac{K}{1 + \left(\frac{K-N_0}{N_0}\right) e^{-rt}} \tag{2}$$

onde $N_0 = N(0)$ é a população inicial. Esta solução pode ser derivada pelo método de separação de variáveis:

$$\begin{aligned}
\frac{dN}{N\left(1 - \frac{N}{K}\right)} &= r\,dt \\
\int \left(\frac{1}{N} + \frac{1/K}{1 - N/K}\right) dN &= \int r\,dt \\
\ln|N| - \ln|1 - N/K| &= rt + C \\
\frac{N}{1 - N/K} &= Ce^{rt}
\end{aligned}$$

Aplicando a condição inicial $N(0) = N_0$, obtemos a solução apresentada.

2.2 Modelo Generalizado com Fatores Ambientais

A equação pode ser estendida para incluir fatores ambientais:

$$\frac{dN}{dt} = r(T, \mathrm{pH}, O_2, S) \cdot N \left(1 - \frac{N}{K(T, \mathrm{pH}, S)}\right) \tag{3}$$

A taxa r pode ser modelada como:

$$r(T, \mathrm{pH}) = r_0 \cdot \exp\left(-\frac{(T - T_{\mathrm{opt}})^2}{\sigma_T^2} - \frac{(\mathrm{pH} - \mathrm{pH}_{\mathrm{opt}})^2}{\sigma_{\mathrm{pH}}^2}\right) \tag{4}$$

2.2.1 Interpretação dos Parâmetros

- T_{opt}: Temperatura ótima para crescimento (°C)
- $\mathrm{pH}_{\mathrm{opt}}$: pH ótimo para crescimento
- $\sigma_T, \sigma_{\mathrm{pH}}$: Largura das curvas de tolerância
- S: Concentração de nutrientes (g/L)
- O_2: Concentração de oxigênio dissolvido (mg/L)

2.2.2 Solução Numérica

Para o modelo generalizado, métodos numéricos como Runge-Kutta de 4ª ordem são necessários:

$$\begin{aligned}
k_1 &= h \cdot f(t_n, N_n) \\
k_2 &= h \cdot f(t_n + h/2, N_n + k_1/2) \\
k_3 &= h \cdot f(t_n + h/2, N_n + k_2/2) \\
k_4 &= h \cdot f(t_n + h, N_n + k_3) \\
N_{n+1} &= N_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)
\end{aligned}$$

onde $f(t, N)$ é o lado direito da EDO e h é o passo de integração.

3 Modelagem Estatística da Classificação

3.1 Descrição do Problema

Dado um vetor de atributos $\mathbf{x} \in \mathbb{R}^n$, buscamos uma função f que classifica a amostra:

$$f : \mathbf{x} \mapsto y \in \{1, 2, ..., C\} \tag{5}$$

onde C é o número de classes bacterianas.

3.2 Pipeline de Processamento

$$\mathbf{x} \xrightarrow{\text{scaling}} \mathbf{x}' \xrightarrow{\text{SMOTE}} \mathbf{x}'' \xrightarrow{\text{selection}} \mathbf{x}''' \xrightarrow{\text{PCA}} \mathbf{z} \xrightarrow{f_\theta} \hat{y} \tag{6}$$

3.2.1 Detalhamento Matemático

1. **Scaling**: Normalização z-score
$$x'_i = \frac{x_i - \mu_i}{\sigma_i} \tag{7}$$

2. **SMOTE**: Geração de amostras sintéticas para classes minoritárias
$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda(\mathbf{x}_j - \mathbf{x}_i) \tag{8}$$
onde $\lambda \sim U(0, 1)$ e $\mathbf{x}_j$ é um vizinho próximo de $\mathbf{x}_i$.

3. **PCA**: Projeção em componentes principais
$$\mathbf{z} = \mathbf{W}^T(\mathbf{x}''' - \boldsymbol{\mu}) \tag{9}$$
onde $\mathbf{W}$ contém os autovetores da matriz de covariância $\boldsymbol{\Sigma}$.

3.3 Interpretação Geométrica e Probabilística

Após a aplicação do PCA, os dados são representados em $\mathbb{R}^k$, e os classificadores aprendem fronteiras que separam regiões de classes.

Para modelos baseados em boosting (XGBoost):

$$P(y = c \mid \mathbf{z}) = \frac{e^{s_c(\mathbf{z})}}{\sum_{j=1}^{C} e^{s_j(\mathbf{z})}} \tag{10}$$

onde $s_c(\mathbf{z}) = \sum_{m=1}^{M} \eta f_m^c(\mathbf{z})$ é o score da classe c, combinando M árvores de decisão com taxa de aprendizado η.

3.3.1 Função de Perda

A otimização utiliza a entropia cruzada:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log p_{i,c} \tag{11}$$

com regularização L2:

$$\mathcal{L}_{\text{total}} = \mathcal{L}(\theta) + \lambda \|\theta\|_2^2 \tag{12}$$

4 Possíveis Expansões Futuras

4.1 ODEs Neurais (Neural ODEs)

$$\frac{d\mathbf{h}(t)}{dt} = f(\mathbf{h}(t), t, \theta) \tag{13}$$

onde f é uma rede neural parametrizada por θ. A solução é obtida por:

$$\mathbf{h}(t_1) = \mathbf{h}(t_0) + \int_{t_0}^{t_1} f(\mathbf{h}(t), t, \theta) dt \tag{14}$$

4.1.1 Aplicação ao Crescimento Bacteriano

Podemos modelar:

$$\frac{dN}{dt} = f_{\text{NN}}(N, T, \text{pH}, O_2, S; \theta) \tag{15}$$

onde f_{NN} aprende a dinâmica não-linear diretamente dos dados.

5 Conclusão

A combinação de modelagem dinâmica e aprendizado supervisionado permite uma compreensão mais profunda e acurada das culturas bacterianas. O uso de modelos matemáticos fundamentados amplia a interpretabilidade e potencial de generalização. As expansões apresentadas, como Neural ODEs, abrem caminho para modelos híbridos que integram conhecimento de domínio com aprendizado de máquina.

Apêndice: Derivações Matemáticas

A. Derivação Completa da Equação Logística

Partindo da equação:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right)$$

Separamos as variáveis:

$$\frac{dN}{N(1 - N/K)} = r\,dt$$

Expandimos em frações parciais:

$$\int \left(\frac{1}{N} + \frac{1/K}{1 - N/K}\right) dN = \int r\,dt$$

Integrando ambos os lados:

$$\ln N - \ln\left(1 - \frac{N}{K}\right) = rt + C$$

Exponenciando:

$$\frac{N}{1 - N/K} = Ce^{rt}$$

Resolvendo para $N(t)$:

$$N(t) = \frac{K}{1 + \frac{K - N_0}{N_0} e^{-rt}}$$

B. Cálculo dos Autovalores para PCA

Dada a matriz de covariância $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

Os autovetores $\mathbf{w}_i$ satisfazem:

$$\mathbf{\Sigma}\mathbf{w}_i = \lambda_i \mathbf{w}_i$$

A variância explicada por cada componente é:

$$\text{VE}_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$