Mathematical Modeling and Supervised Learning for Bacterial Culture Classification

Author: Matheus Pullig Soranço de Carvalho

April 18, 2025

Abstract

This article presents the mathematical modeling of bacterial growth and the development of a supervised machine learning model to classify bacterial cultures based on their physicochemical and environmental characteristics. The modeling includes both ordinary differential equations to represent population dynamics and the geometric and statistical representation of data in a supervised learning pipeline. Analytical and numerical solutions for the proposed models are presented, along with complete mathematical derivations of the classification algorithms.

1 Introduction

Bacterial growth is influenced by factors such as temperature, pH, nutrient availability, and oxygen. Modeling these dynamics can provide a solid mathematical foundation for culture classification with machine learning techniques. In this work, we explore both the theoretical foundations and practical implementations of these models, with emphasis on the mathematical interpretation of results.

2 Mathematical Modeling of Bacterial Growth

2.1 Classical Logistic Model

Population growth can be described by the logistic equation:

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

where:

- $N(t)$: bacterial population at time t (units: CFU/mL);
- r: intrinsic growth rate (units: h^{-1});
- K: carrying capacity of the medium (units: CFU/mL).

2.1.1 Analytical Solution

The logistic equation has a known analytical solution:

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

where $N_0 = N(0)$ is the initial population. This solution can be derived using separation of variables:

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

Applying the initial condition $N(0) = N_0$, we obtain the presented solution.

2.2 Generalized Model with Environmental Factors

The equation can be extended to include environmental factors:

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

The rate r can be modeled as:

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

2.2.1 Parameter Interpretation

- T_{opt}: Optimal growth temperature (°C)
- pH_{opt}: Optimal growth pH
- $\sigma_T, \sigma_{\text{pH}}$: Width of tolerance curves
- S: Nutrient concentration (g/L)
- O_2: Dissolved oxygen concentration (mg/L)

2.2.2 Numerical Solution

For the generalized model, numerical methods like 4th-order Runge-Kutta are required:

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

where $f(t, N)$ is the right-hand side of the ODE and h is the integration step.

3 Statistical Modeling of Classification

3.1 Problem Description

Given a feature vector $\mathbf{x} \in \mathbb{R}^n$, we seek a function f that classifies the sample:

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

where C is the number of bacterial classes.

3.2 Processing Pipeline

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

3.2.1 Mathematical Details

1. **Scaling**: Z-score normalization

$$x_i' = \frac{x_i - \mu_i}{\sigma_i} \tag{7}$$

2. **SMOTE**: Synthetic sample generation for minority classes

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda(\mathbf{x}_j - \mathbf{x}_i) \tag{8}$$

   where $\lambda \sim U(0, 1)$ and $\mathbf{x}_j$ is a nearest neighbor of $\mathbf{x}_i$.

3. **PCA**: Projection onto principal components

$$\mathbf{z} = \mathbf{W}^T(\mathbf{x}''' - \boldsymbol{\mu}) \tag{9}$$

   where $\mathbf{W}$ contains the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$.

3.3 Geometric and Probabilistic Interpretation

After PCA application, the data is represented in $\mathbb{R}^k$, and classifiers learn boundaries that separate class regions.

For boosting-based models (XGBoost):

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

where $s_c(\mathbf{z}) = \sum_{m=1}^{M} \eta f_m^c(\mathbf{z})$ is the class c score, combining M decision trees with learning rate η.

3.3.1 Loss Function

Optimization uses cross-entropy:

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

with L2 regularization:

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

4 Possible Future Expansions

4.1 Neural ODEs

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

where f is a neural network parameterized by θ. The solution is obtained by:

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

4.1.1 Application to Bacterial Growth

We can model:

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

where f_{NN} learns the nonlinear dynamics directly from data.

5 Conclusion

The combination of dynamic modeling and supervised learning enables deeper and more accurate understanding of bacterial cultures. The use of mathematically grounded models enhances interpretability and generalization potential. The presented expansions, such as Neural ODEs, pave the way for hybrid models that integrate domain knowledge with machine learning.

Appendix: Mathematical Derivations

A. Complete Derivation of the Logistic Equation

Starting from the equation:

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

We separate variables:

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

Expand using partial fractions:

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

Integrating both sides:

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

Exponentiating:

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

Solving for $N(t)$:

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

B. Eigenvalue Calculation for PCA

Given the covariance matrix $\mathbf{\Sigma}$:

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

The eigenvectors $\mathbf{w}_i$ satisfy:

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

The explained variance for each component is:

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