Gaussian process classification for omics data

1 Gaussian Process Classification

1.1 Gaussian Process Prior

Given a set of n training samples, the input dataset is represented as $\mathbf{X} \in \mathbb{R}^{n \times d}$, where each sample $\mathbf{x}_i \in \mathbb{R}^d$ consists of three different types of features:

- Genetic features: $\mathbf{x}_{\text{genes}} \in \mathbb{R}^p$, where p represents the number of genetic features.
- Vaccine status: $x_{\text{vaccine}} \in \{0, 1\}$, a binary categorical variable indicating vaccination status.
- Days post-infection (DPI): $x_{\text{days}} \in \mathbb{R}$, representing the number of days since infection.

Thus, each sample can be expressed as:

$$\mathbf{x}_i = (\mathbf{x}_{\text{genes}}, x_{\text{vaccine}}, x_{\text{days}}) \in \mathbb{R}^{p+2}. \tag{1}$$

The corresponding binary labels $\mathbf{y} \in \{0, 1\}^n$ represent the outcomes of the disease, where $y_i = 1$ indicates the sample is challenged and $y_i = 0$ indicates unchallenged. We assume a latent function $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^T$ following a Gaussian Process:

$$\mathbf{f} \sim \mathcal{GP}(0, K(\mathbf{X}, \mathbf{X})), \tag{2}$$

where $K(\mathbf{X}, \mathbf{X})$ is the covariance matrix computed using a kernel function $k(\mathbf{x}, \mathbf{x}')$. The function $f(\mathbf{x})$ is mapped to classification probabilities through a logistic transformation.

1.2 Likelihood Function

Since classification involves discrete labels, we transform the latent function through the sigmoid function:

$$p(y_i = 1 | f_i) = \sigma(f_i) = \frac{1}{1 + e^{-f_i}}, \tag{3}$$

which gives the Bernoulli likelihood function:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} \sigma(f_i)^{y_i} (1 - \sigma(f_i))^{1-y_i}. \tag{4}$$

The log-likelihood is then:

$$\log p(\mathbf{y}|\mathbf{f}) = \sum_{i=1}^{n} \left[y_i \log \sigma(f_i) + (1 - y_i) \log(1 - \sigma(f_i)) \right]. \tag{5}$$

1.3 Posterior Distribution and MCMC Sampling

The posterior distribution over the latent function is given by:

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}). \tag{6}$$

Since the logistic likelihood is non-Gaussian, inference is performed using Markov Chain Monte Carlo (MCMC).The hyperparameters $\boldsymbol{\theta} = (\theta_{\text{genes}}, \theta_{\text{vaccine}}, \theta_{\text{days}})$ are sampled by the random walk Metropolis-Hastings (RWMH) algorithm.

2 Kernel Selection

To capture different aspects of the data, we use a hybrid kernel function that consists of three separate kernels:

- Gaussian kernel for genetic features:

$$k_{\text{genes}}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{||\mathbf{x}_1 - \mathbf{x}_2||^2}{\theta_{\text{genes}}}\right) \tag{7}$$

- Gaussian kernel for vaccine status:

$$k_{\text{vaccine}}(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{\theta_{\text{vaccine}}}\right) \tag{8}$$

- Gaussian kernel for days post-infection (DPI):

$$k_{\text{days}}(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{\theta_{\text{days}}}\right). \tag{9}$$

The combined kernel function is given by:

$$k_{\text{hybrid}}(\mathbf{x}_1, \mathbf{x}_2) = (k_{\text{genes}} + k_{\text{vaccine}}) \cdot k_{\text{days}}. \tag{10}$$

3 Prediction

Given a test point $\mathbf{x}^*$, the posterior predictive distribution is:

$$p(y^* = 1|\mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \sigma(f^*)p(f^*|\mathbf{x}^*, \mathbf{X}, \mathbf{y})df^*. \tag{11}$$

Using MCMC samples, we approximate:

$$\mathbb{E}[y^*] \approx \frac{1}{N}\sum_{i=1}^{N} \sigma(f_i^*), \tag{12}$$

where N is the number of posterior samples. The predicted class is then given by:

$$y^* = \begin{cases} 1, & \text{if } \mathbb{E}[y^*] > 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

4 Results and Performance

To evaluate the model, we conducted 100 trials where different training and test splits were randomly selected. The average classification accuracy over these trials is shown in Figure 1.

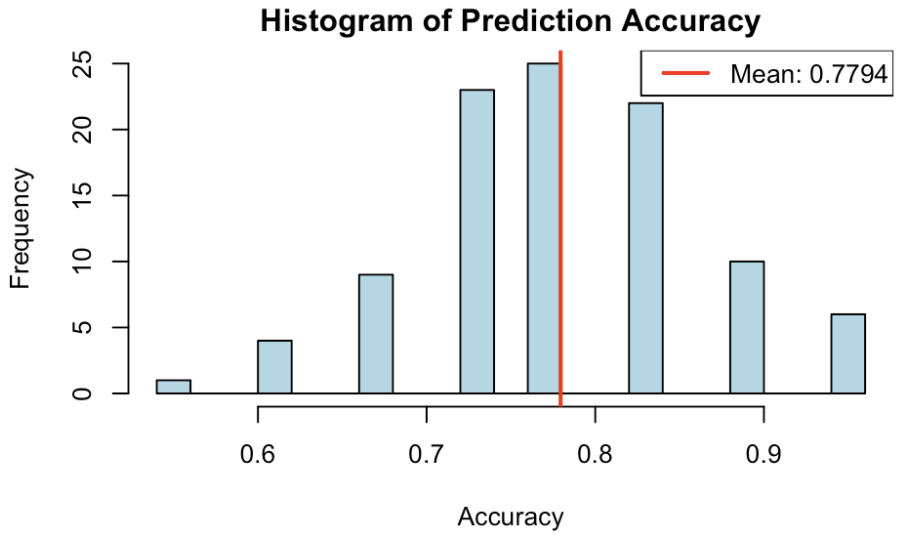

Figure 1: Average predicted accuracy over 100 trials.