6.7900 Fall 2024: Lecture Notes 16

1 Gaussian Processes

1.1 Motivation

In real-world regression problems, we often have a target variable y that varies smoothly along some observed dimensions x. Beyond smoothness, we may lack additional information about the relationship between y and x, with no parametric model available to define this relationship. These problems often involve spatio-temporal data, where x includes both spatial and temporal dimensions. We assume access to a limited set of labeled data points $(x^{(i)}, y^{(i)})$, and our goal is to predict the value of y at an *unseen* x.

Key characteristics of these problems include:

- Sparse, limited data; data may be expensive or difficult to collect
- A smooth, nonlinear relationship between x and y
- Unknown form of the relationship between x and y
- Need for uncertainty quantification: given a particular x, how confident are we in the predicted value of y?

Examples of such problems:

- Predicting ocean currents at various locations on Earth over time
- Predicting rocket lift as a function of variables like re-entry speed and angle of attack

For such settings, conventional supervised models (e.g., neural networks, decision trees) may be less effective due to 1) limited training data and 2) difficulty in capturing smoothness properties. This motivates the use of a *Gaussian process*, a type of distribution suited for these conditions.

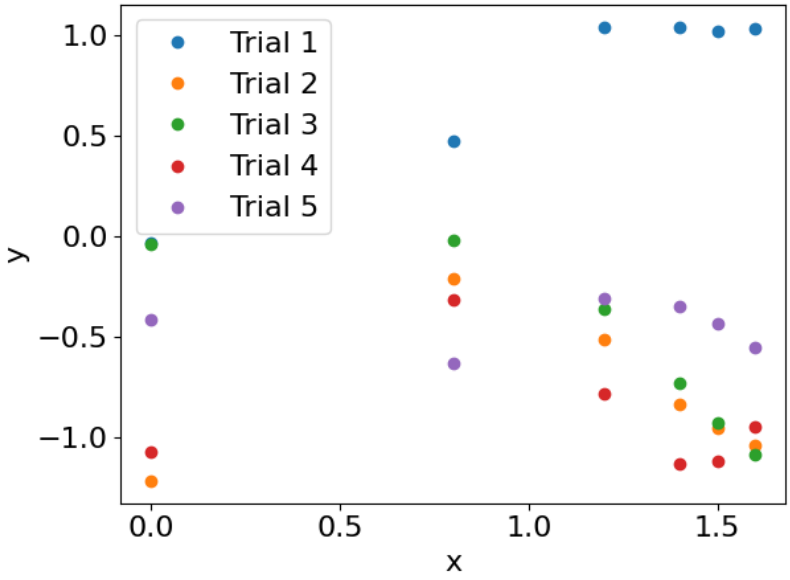

Figure 1: Five samples from a six-dimensional multivariate Gaussian with distance-dependent covariance ($M = 6$, $d = 1$). The y-axis shows the drawn values along each dimension, while the x-axis indicates the spatial locations associated with each dimension.

1.2 Multivariate Gaussian with Distance-dependent Covariance

As an intermediate step towards Gaussian processes, consider a multivariate Gaussian with *distance-dependent* covariance. Recall that a multivariate Gaussian distribution on $y \in \mathbb{R}^M$ is defined as:

$$p(y) \propto e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)} \tag{1}$$

where μ is the mean and Σ is the covariance matrix.

Now, interpret each dimension y^i as corresponding to a spatial location $x^i \in \mathbb{R}^d$. Here, $y^i \in \mathbb{R}$ (the i-th component of y) represents the value at spatial location x^i.

The covariance between y^i and y^j is defined as:

$$\mathbb{E}[(y^i - \mu^i)(y^j - \mu^j)] = \Sigma_{ij} \tag{2}$$

Suppose that the covariance depends on the *spatial distance* between x^i and x^j:

$$\Sigma_{ij} = \rho(||x^i - x^j||) \tag{3}$$

where ρ is a positive, monotonically decreasing function. If we draw y from $p(y)$, then components of y corresponding to nearby x locations will take similar values. This results in the desired smoothness property: for similar points x^i and x^j, the values y^i and y^j are likely to be similar.

Figure 1 shows samples from a multivariate Gaussian with distance-dependent covariance, illustrating smoothness in the x-y relationship.

1.3 What is a Gaussian Process?

A Gaussian process is essentially a distribution over y in the limit as $M \to \infty$, allowing x to vary continuously.

Formally, a Gaussian process $\mathcal{GP}(\mu, k)$ is a random function f mapping from $x \in \mathbb{R}^d$ to $y \in \mathbb{R}$, defined by a mean function $\mu(x)$ and covariance function $k(x, x')$, such that:

$$\mu(x) = \mathbb{E}[f(x)] \tag{4}$$

$$k(x, x') = \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))] \tag{5}$$

where the distribution over $[f(x^1), f(x^2), \ldots, f(x^M)]$ is multivariate Gaussian for any finite sequence $x^1, x^2, \ldots, x^M$.

1.4 Inference with a Gaussian Process

Suppose we have a set of observed data points $(x^1, y^1), (x^2, y^2), \ldots, (x^N, y^N)$ and want to predict the y-values at new points $x^{N+1}, x^{N+2}, \ldots, x^{N+M}$.

For convenience, denote the training data as matrix $X \in \mathbb{R}^{N \times d}$ and test data as $X' \in \mathbb{R}^{M \times d}$, with training labels $Y \in \mathbb{R}^N$. The joint distribution of $[f(X), f(X')]$ is Gaussian.

Assuming μ is the zero function, the posterior distribution of $f(X')$ given $f(X) = Y$ has a mean and covariance:

$$\mathbb{E}[f(X')|f(X) = Y] = k(X', X)k(X, X)^{-1}Y \tag{6}$$

$$\mathrm{Cov}[f(X')|f(X) = Y] = k(X', X') - k(X', X)k(X, X)^{-1}k(X, X') \tag{7}$$

where $k(X, X)$ is the covariance matrix for training points, $k(X', X)$ is the covariance between test and training points, and $k(X', X')$ is the covariance for test points alone.

Exercise: (How) do the mean and covariance change when the μ is not zero?

Figure 2 illustrates that the Gaussian process produces different uncertainties of the value of $f(x)$ for different values of x. In regions far from the observed training data, there is more uncertainty in the value of $f(x)$.

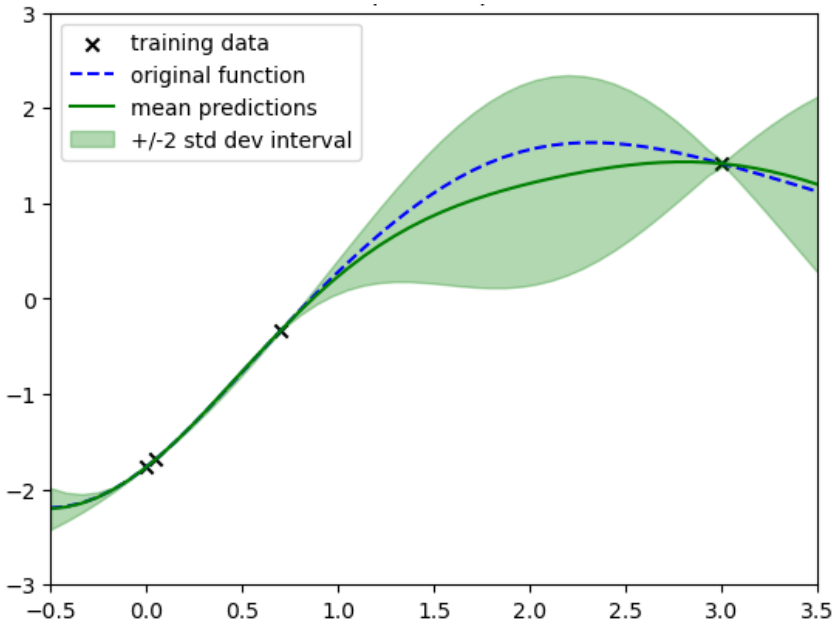

Figure 2: Gaussian process inference of $y = f(x)$ given four labeled (x, y) pairs marked by an $\times$ ($d = 1$). The green line indicates the mean of the Gaussian process while the margins indicate ± 2 standard deviations. The blue dashed line indicates the original function used to generate the data.

1.5 Design of Kernel Function k

In practice, some rules of thumb help guide kernel design. A common choice is the squared exponential kernel:

$$k(x, x') = \sigma^2 \exp\left(-\frac{1}{2\ell^2}||x - x'||^2\right) \tag{8}$$

where σ represents the variance scale of $f(x)$ values, and ℓ is the length scale in x-space, controlling how rapidly f varies. The parameters σ and ℓ are known as the *hyperparameters* of the Gaussian process and can be optimized to maximize the likelihood of the observed data.