Stochastic Processes

Lecture Notes

Week 05: Gaussian Processes

Payam Taebi

Fall 2024

https://stoch-sut.github.io/

Contents

1 Motivation

1.1 Introduction

- In many real-world applications, we are confronted with the need for a robust and interpretable model.
- Traditional approaches often rely on applying knowledge of the underlying physics to deduce specific model forms.
- However, our understanding of the underlying physical processes can be limited, involve too many assumptions, or include variables that are difficult to measure.
- In such cases, machine learning provides an alternative by learning relationships directly from existing data or measurements, resulting in empirical models.
- Deep Neural Networks (DNNs) have shown impressive results but are often considered black boxes (not interpretable) and are vulnerable to adversarial attacks.

1.2 Limitations of Deep Neural Networks (DNNs)

- **Data Availability:** DNNs typically require large amounts of labeled data to perform effectively.
- **Processing Complexity:** Training and deploying DNNs can be computationally intensive.
- **Robustness:** DNNs can be sensitive to noise and adversarial perturbations.
- **Interpretability:** Understanding the decision-making process within DNNs is challenging, limiting their applicability in fields requiring transparency.

1.3 Advantages of Gaussian Processes (GPs)

- **Probabilistic Framework:** GPs provide a principled way to quantify uncertainty in predictions.
- **Flexibility:** GPs are non-parametric models, allowing them to adapt their complexity based on the data.
- **Interpretability:** GPs offer insights into the underlying data through the covariance function (kernel).
- **Bayesian Approach:** GPs integrate seamlessly with Bayesian inference, facilitating the incorporation of prior knowledge.

2 The Gaussian Distribution

2.1 Gaussian Density Function

- The Gaussian distribution, also known as the normal distribution, is the most common probability density function.

- It is completely specified by its mean μ and variance σ^2:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

2.2 Gaussian PDF Example

Example 1 (Gaussian PDF with Specific Parameters)**.** ***Problem:*** *Consider a Gaussian distribution with mean $\mu = 1.6$ and variance $\sigma^2 = 0.125$. Describe its properties and plot the probability density function.*

Solution: The Gaussian distribution with $\mu = 1.6$ and $\sigma^2 = 0.125$ has the following properties:

- **Mean (μ):** The peak of the distribution is centered at 1.6.
- **Variance (σ^2):** Indicates the spread of the distribution. A variance of 0.125 implies a standard deviation of $\sigma = \sqrt{0.125} \approx 0.354$.
- **Probability Density Function (PDF):**

$$f(x) = \frac{1}{\sqrt{2\pi \times 0.125}} \exp\left(-\frac{(x-1.6)^2}{2 \times 0.125}\right)$$

The PDF is bell-shaped, symmetric around the mean, and its width is determined by the variance. □

2.3 Important Gaussian Properties

- **Sum of Independent Gaussians:** The sum of independent Gaussian random variables is also Gaussian.

  If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent, then $X+Y \sim \mathcal{N}(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$.
- **Central Limit Theorem (CLT):** As the sum of a large number of independent and identically distributed (i.i.d.) random variables with finite mean and variance, the distribution of the sum tends towards a Gaussian distribution, regardless of the original distribution of the variables.
- **Scaling:** Scaling a Gaussian random variable by a constant results in another Gaussian random variable.

$$\text{If } X \sim \mathcal{N}(\mu, \sigma^2), \text{ then } aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

3 Covariance Functions

3.1 Definition and Importance

- The covariance matrix in Gaussian Processes is constructed by evaluating a covariance function (kernel) over pairs of input points.
- **Covariance Function (Kernel):** A function that defines the covariance between any two points in the input space.

$$K(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right]$$

- Covariance functions are the building blocks of covariance matrices and play a crucial role in determining the properties of the Gaussian Process.

3.2 Example: Exponentiated Quadratic Kernel Function

- Also known as the Radial Basis Function (RBF), Squared Exponential, or Gaussian kernel.
- It is one of the most widely used kernels due to its smoothness and infinite differentiability.
- **Form:**
$$K(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2l^2}\right)$$
where:
  - σ_f^2 is the signal variance.
  - l is the length-scale parameter.
- **Properties:**
  - **Smoothness:** Implies that the functions drawn from the GP are smooth.
  - **Stationarity:** The covariance depends only on the distance between input points, not their absolute positions.
  - **Isotropy:** The covariance is the same in all directions.

3.3 Additional Covariance Functions

- **Linear Kernel:**
$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' + c$$
where c is a constant.
- **Matérn Kernel:** Controls the smoothness of the functions. It introduces a parameter ν that determines the differentiability of the functions.
$$K(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{l}\right)$$
where K_ν is the modified Bessel function of the second kind.
- **Periodic Kernel:**
$$K(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{2\sin^2(\pi||\mathbf{x} - \mathbf{x}'||/p)}{l^2}\right)$$
where p is the period.

4 Gaussian Process

4.1 Definition

- A **Gaussian Process (GP)** is a stochastic process where any finite set of random variables has a joint Gaussian distribution.
- A GP is completely specified by its mean function $m(\mathbf{x})$ and covariance function $K(\mathbf{x}, \mathbf{x}')$:
$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$$

- **Mean Function:**
$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$
- **Covariance Function:**
$$K(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right]$$

4.2 GP as Distribution over Functions

- A GP can be viewed as a distribution over functions, providing a probabilistic framework for function estimation.
- **Bayesian Approach:** GPs adopt a Bayesian perspective, allowing for the incorporation of prior knowledge and the updating of beliefs based on observed data.
- **Uncertainty Quantification:** GPs naturally provide uncertainty estimates for predictions, which are crucial for decision-making processes.

4.3 Relation to Neural Networks

- GPs can be interpreted as neural networks with infinitely many hidden units, where each weight has a Gaussian distribution.
- This perspective links GPs to deep learning, highlighting their flexibility and expressiveness.
- Unlike traditional neural networks, GPs do not require explicit parameter training, as their parameters are implicitly defined by the covariance function.

4.4 Example: Multivariate Gaussian Distribution

Example 2 (Multivariate Gaussian Density Function)**.** ***Problem:*** *Describe the multivariate Gaussian distribution for a k-dimensional random vector and its density function.*

Solution: *The multivariate Gaussian distribution for a k-dimensional random vector* $\mathbf{X} = [X_1, X_2, \ldots, X_k]^T$ *is defined by its mean vector* $\boldsymbol{\mu} \in \mathbb{R}^k$ *and covariance matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$*.*

The probability density function (PDF) is given by:

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right)$$

where:

- $|\boldsymbol{\Sigma}|$ *is the determinant of the covariance matrix.*
- $\boldsymbol{\Sigma}^{-1}$ *is the inverse of the covariance matrix.*

4.5 Case Study: Overdetermined and Underdetermined Systems

Example 3 (Overdetermined Systems)**.** ***Problem:*** *Consider a system of two equations with two unknowns. Now, add an additional observation leading to an overdetermined system. How can we solve this using a noise model?*

Solution:

- ***System of Equations:***
$$\begin{cases} a_1x + b_1y = c_1 \\ a_2x + b_2y = c_2 \\ a_3x + b_3y = c_3 \end{cases}$$

- ***Overdetermined System:*** *With three equations and two unknowns, the system may not have an exact solution.*

- ***Noise Model:*** *Introduce a noise term to account for discrepancies:*
$$a_ix + b_iy = c_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- ***Probabilistic Formulation:*** *Model the observations probabilistically, leading to a likelihood function that can be maximized to find the best estimates for x and y.*

- ***Gaussian Process Approach:*** *Treat the unknown parameters as random variables with a Gaussian prior, allowing for a Bayesian treatment of the problem.*

Example 4 (Underdetermined Systems)**.** ***Problem:*** *Consider a system with two unknowns and one observation. How can we compute the distribution of solutions?*
Solution:

- ***System of Equations:***
$$a_1x + b_1y = c_1$$

- ***Underdetermined System:*** *With one equation and two unknowns, there are infinitely many solutions.*

- ***Probabilistic Approach:*** *Assign a probability distribution to the parameters (e.g., x and y) to quantify the uncertainty.*

- ***Gaussian Process Model:*** *Use a GP to model the relationship between the variables, allowing us to compute a distribution over the possible solutions.*

4.6 Probability for Under- and Overdetermined Systems

- **Overdetermined Systems:** Introduce a probability distribution for the variables to handle the excess equations.

- **Underdetermined Systems:** Introduce a probability distribution for the parameters to handle the insufficient equations.

- **Bayesian Treatment:** Utilize Gaussian Processes to model the random line example, allowing for a distribution over possible solutions.

- **Multivariate Priors:** In Bayesian inference, multivariate Gaussian priors are often used to model the distribution over parameters and variables.

Example 5 (Multivariate Linear Regression)**.** ***Problem:*** *Describe the multivariate linear regression model using Gaussian priors.*
Solution:

- ***Model Formulation:***
$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$$
where:

- $\mathbf{y}$ *is the vector of observations.*
- $\mathbf{X}$ *is the design matrix.*
- $\mathbf{w}$ *is the weight vector (parameters).*
- $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ *is the noise.*

- ***Prior Distribution:*** *Assign a Gaussian prior to the weight vector* $\mathbf{w}$*:*

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

where $\mathbf{K}$ *is the covariance matrix.*

- ***Posterior Distribution:*** *Combining the prior and the likelihood using Bayes' theorem results in a posterior distribution for* $\mathbf{w}$ *that is also Gaussian.*

5 Basis Function Representations

5.1 Basis Function Form

- Radial Basis Functions (RBFs) are commonly used in GP models. They have the form:

$$\phi_i(\mathbf{x}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{c}_i||^2}{2\sigma^2}\right)$$

where:

- $\mathbf{c}_i$ are the centers of the basis functions.
- σ is the width parameter.

- A set of RBFs maps data into a high-dimensional feature space, enabling the modeling of complex nonlinear relationships.

5.2 Random Functions via Basis Functions

- Functions can be represented as a linear combination of basis functions:

$$f(\mathbf{x}) = \sum_{i=1}^{m} w_i \phi_i(\mathbf{x})$$

where m is the number of basis functions and w_i are the weights.

- **Weight Distribution:** The weights $\mathbf{w} = [w_1, w_2, \ldots, w_m]^T$ are sampled from a Gaussian distribution:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- **Sample Functions:** Each sample of $f(\mathbf{x})$ corresponds to a different realization of the weights $\mathbf{w}$.

Example 6 (Sample Functions from Basis Function Representation)**.** ***Problem:*** *Given a set of basis functions and weights sampled from a Gaussian distribution, describe how to generate sample functions.*

Solution:

- ***Step 1:*** *Define the basis functions* $\phi_i(\mathbf{x})$ *for* $i = 1, 2, \ldots, m$.
- ***Step 2:*** *Sample the weights* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- ***Step 3:*** *Compute the function:*

$$f(\mathbf{x}) = \sum_{i=1}^{m} w_i \phi_i(\mathbf{x})$$

- ***Result:*** *Each set of sampled weights* $\mathbf{w}$ *generates a unique function* $f(\mathbf{x})$*, allowing us to visualize the variability and uncertainty in the model.*

6 Constructing Covariance

6.1 Matrix Notation for Functions

- Using matrix notation, the function $f(\mathbf{x})$ can be written as:

$$\mathbf{f} = \mathbf{\Phi}\mathbf{w}$$

  where:

  - $\mathbf{\Phi}$ is the design matrix with elements $\Phi_{ij} = \phi_j(\mathbf{x}_i)$.
  - $\mathbf{w}$ is the weight vector.

- **Assumptions:**
  - The weights $\mathbf{w}$ are Gaussian distributed.
  - The basis functions $\mathbf{\Phi}$ are fixed and non-stochastic for a given training set.

6.2 Infinite Feature Space

- A GP with an infinite number of basis functions corresponds to a non-parametric model.
- As the number of basis functions m approaches infinity, the model becomes more flexible, allowing it to capture complex patterns in the data.
- **Covariance Function:** In the infinite limit, the covariance function of the GP can be seen as the limit of the covariance constructed from finite basis functions.
- **Functional Forms:** The functional forms for covariance functions and basis functions are similar but distinct. For example, the RBF kernel can be derived from an infinite set of RBF basis functions.

7 Gaussian Noise

7.1 Gaussian Noise Model

- In regression models, Gaussian noise is introduced to account for the mismatch between the true underlying function and the observed data.

- The noise model can be represented as:

$$y = f(\mathbf{x}) + \epsilon$$

  where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.

- **Covariance Representation:**

$$K_y = K + \sigma_n^2 \mathbf{I}$$

  where K is the covariance matrix derived from the kernel function, and $\mathbf{I}$ is the identity matrix.

- **Additive Nature:** Due to the additive nature of Gaussian noise, the noise term can be simply added to the existing covariance matrix, facilitating straightforward inference.

8 Gaussian Process Limitations

8.1 Computational Complexity

- Inference in GPs involves inverting the covariance matrix, which has a computational complexity of $\mathcal{O}(n^3)$, where n is the number of training points.
- For large datasets, this becomes computationally infeasible.
- **Solutions:**
  - **Sparse GPs:** Approximate methods that reduce computational complexity by using a subset of the data.
  - **Inducing Points:** Introduce a set of inducing points to approximate the full covariance matrix.
  - **Stochastic Variational Inference:** Combine variational methods with stochastic optimization for scalability.

8.2 Handling Discontinuities

- GPs with standard covariance functions (e.g., RBF) assume smoothness in the underlying function.
- This assumption limits their ability to model functions with discontinuities or abrupt changes, such as:
  - Financial crises
  - Phase transitions like phosphorylation
  - Collisions or edges in images
- **Solution:** Utilize covariance functions that can handle non-smooth behavior or combine multiple kernels to capture different characteristics.

8.3 Covariance Function Limitations

- The commonly used exponentiated quadratic (RBF) covariance function imposes strong smoothness assumptions, which may be too restrictive for certain applications.
- **Alternatives:**
  - **Matérn Kernel:** Introduces a smoothness parameter ν that controls the differentiability of the functions.
  - **Periodic Kernel:** Suitable for modeling periodic phenomena.
  - **Linear Kernel:** Useful for capturing linear trends in the data.
- **Composite Kernels:** Combine multiple kernels to capture various aspects of the data.

9 Summary of Gaussian Processes

- **Broad Introduction to Gaussian Processes:**
  - Started with the Gaussian distribution to build intuition.
  - Motivated Gaussian processes through the multivariate density function.
- **Role of Covariance:**
  - Emphasized the significance of the covariance function in defining the properties of the GP.
  - Covariance functions determine the smoothness, periodicity, and other characteristics of the functions modeled by the GP.
- **Nonlinear Regression with Uncertainty:**
  - GPs perform nonlinear regression while providing uncertainty estimates (error bars) for predictions.
- **Optimization of Covariance Parameters:**
  - Parameters of the covariance function (kernel) can be optimized using maximum likelihood, facilitating model selection and fitting.
- **Demos and Further Resources:**
  - Interactive demonstrations:
    * `https://edward-rees.com/gp`
    * `http://chifeng.scripts.mit.edu/stuff/gp-demo/`

10 References

- G. Della Gatta, M. Bansal, A. Ambesi-Impiombato, D. Antonini, C. Missero, and D. di Bernardo. "Direct targets of the trp63 transcription factor revealed by a combination of gene expression profiling and reverse engineering." *Genome Research*, 18(6): 939–948, Jun 2008.

- A. A. Kalaitzis and N. D. Lawrence. "A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression." *BMC Bioinformatics*, 12(180), 2011.

- R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996. Lecture Notes in Statistics 118.

- J. Oakley and A. O'Hagan. "Bayesian inference for the uncertainty distribution of computer model outputs." *Biometrika*, 89(4): 769–784, 2002.

- C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. [Google Books].

- C. K. I. Williams. "Computation with infinite neural networks." *Neural Computation*, 10(5):1203–1216, 1998.

11 Next Week

- **Point Estimation:** Exploration of methods for estimating the parameters of stochastic processes, including maximum likelihood estimation and Bayesian inference.

Have a good day!