IoT·인공지능·빅데이터 개론 및 실습

Regression (2)

서울대학교 전기정보공학부
오성회

Contents

Gaussian Random Variable

- X is a **Gaussian random variable** if X is a random variable having the following probability density function (**Gaussian or normal distribution**):

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

  - Mean: $\mathbb{E}(X) = \int x f(x) dx = \mu$
  - Variance: $\mathbf{var}(X) = \mathbb{E}\left(X - \mathbb{E}(X)\right)^2 = \sigma^2$
  - Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$

- **Central limit theorem**: Let $X_1, X_2, \ldots$ be independent and identically distributed with $\mathbb{E}(X_i) = \mu$ and $\mathbf{var}(X_i) = \sigma^2 < \infty$. If $S_n = X_1 + \cdots + X_n$, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} X,$$

where X has the standard normal distribution, i.e., $X \sim \mathcal{N}(0, 1)$.

Multivariate Gaussian Random Variable

- A random vector $\mathbf{x} = [X_1 \ \ldots \ X_n]^T$ is said to be **multivariate Gaussian** if every linear combination of the components of X is a Gaussian random variable.
  - That is, for any a_i, $\sum_{i=1}^{n} a_i X_i$ is a Gaussian random variable.
  - We also say $X_1, \ldots, X_n$ are **jointly Gaussian**.
- **Multivariate Gaussian density function**:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right) \qquad (1)$$

$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, where μ is the mean vector and Σ is the covariance matrix.

$$\mu = \mathbb{E}(\mathbf{x}) = \begin{bmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{bmatrix} \qquad \Sigma = \mathbf{cov}(\mathbf{x}) = \mathbb{E}\left(\left((\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\right)\right)$$

Conditional Density of Multivariate Gaussian

Theorem: If $\mathbf{x} \in \mathbb{R}^r$ and $\mathbf{y} \in \mathbb{R}^m$ are jointly Gaussian with $n = r + m$, mean vector $[\mathbb{E}(\mathbf{x})^T \ \mathbb{E}(\mathbf{y})^T]^T$, and covariance matrix

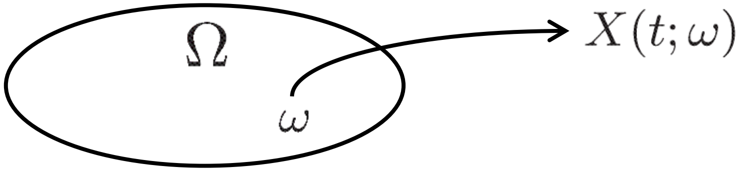
$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix},$$

then the conditional probability density function $p(\mathbf{x}|\mathbf{y})$ is also a Gaussian random vector with mean $\mathbb{E}(\mathbf{x}|\mathbf{y})$ and covariance matrix $\Sigma_{x|y}$, where

$$\begin{aligned} \mathbb{E}(\mathbf{x}|\mathbf{y}) &= \mathbb{E}(\mathbf{x}) + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \mathbb{E}(\mathbf{y})) \\ \Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}. \end{aligned}$$

Random Process

- A **random process** $X(t)$ is a collection of random variables, one for each t, defined on sample space Ω.



Two interpretations:

- For fixed t, $X(t;\omega)$ is a function of ω, i.e., $X(t,\cdot):\Omega\to\mathbb{R}$. Hence, $X(t,\cdot)$ is a random variable.
- For fixed ω, $X(\cdot;\omega):\mathbb{R}\to\mathbb{R}$ is a sample path function.

The distribution of a random process is specified by a collection of cumulative distribution functions (CDFs). More precisely, for all $k\in\mathbb{N}$ and for all $t_1,\ldots,t_k$, we need to specify the joint CDF of $X(t_1),\ldots,X(t_k)$.

Gaussian Process

- **Gaussian process**: A random process $X(t)$ is a **Gaussian process** if for all $k \in \mathbb{N}$ and for all $t_1, \ldots, t_k$, a random vector formed by $X(1), \ldots, X(t_k)$ is jointly Gaussian.
- The joint density is completely specified by
  - Mean: $m(t) = \mathbb{E}(X(t))$, where m is known as a mean function.
  - Covariance: $k(t, s) = \mathbf{cov}(X(t), X(s))$

    $$k(t, s) = \mathbb{E}\left((X(t) - m(t))(X(s) - m(s))\right),$$

    where k is known as a covariance function.
- Notation: $X(t) \sim \mathcal{GP}(m(t), k(t, s))$
- Example: $X(t) = tA$, where $A \sim \mathcal{N}(0, 1)$ and $t \in \mathbb{R}$.

2 Gaussian Process Regression

- $\mathcal{X}$: index set (e.g., time $\mathbb{R}$, space $\mathbb{R}^3$)
- $f(x)$: a collection of random variables with $x \in \mathcal{X}$.
- $f(x)$ is a **Gaussian process** if for any finite set $\{x_1, \ldots, x_n\}$, $\{f(x_1), \ldots, f(x_n)\}$ has a multivariate Gaussian distribution, with mean $\mu \in \mathbb{R}^n$ and covariance $K \in \mathbb{R}^{n \times n}$.
- The mean μ and covariance K depend on the chosen finite set $\{x_1, \ldots, x_n\}$.
- **Gaussian process regression**: A nonparametric regression method using properties of Gaussian processes.
- Two views to interpret Gaussian process regression:
  - Weight-space view
  - Function-space view

(Source: Gaussian Processes for Machine Learning, C. E. Rasmussen and C. K. I. Williams, MIT Press, 2006.)

Function-Space View

- $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$, i.e., $f(x)$ is a Gaussian process
- $m(x) = \mathbb{E}(f(x))$, mean function
- $k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$, covariance function
- Example: $f(x) = \phi(x)^T w$ with $w \sim \mathcal{N}(0, \Sigma_p)$.
    - $\mathbb{E}(f(x)) = \phi(x)^T \mathbb{E}(w) = 0.$
    - $\mathbb{E}(f(x)f(x')) = \phi(x)^T \mathbb{E}(ww^T)\phi(x') = \phi(x)^T \Sigma_p \phi(x').$
    - Hence, $f(x)$ and $f(x')$ are jointly Gaussian.
    - It is also true for $f(x_1), \ldots, f(x_n)$ for any $x_1, \ldots, x_n$ and n.
    - Therefore, $f(x)$ is a Gaussian process.
- If $K(x_p, x_q) = \mathbf{cov}(f(x_p), f(x_q))$, then (assuming $m(x) = 0$)

$$f_* \sim \mathcal{N}(0, K(x_*, x_*)).$$

2 Gaussian Process Regression

Prediction

f and f_* are jointly Gaussian, hence, for any finite number of measurements at $x_1, \ldots, x_n$ and x_*, (again assuming $m(x) = 0$)

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix}\right),$$

where $[K(X,X)]_{ij} = k(x_i, x_j)$.

Recall that the conditional distribution of a jointly Gaussian random vector $[\mathbf{x}^T \ \mathbf{y}^T]^T$ is such that $\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mathbb{E}(\mathbf{x}|\mathbf{y}), \Sigma_{x|y})$, where

$$\begin{aligned} \mathbb{E}(\mathbf{x}|\mathbf{y}) &= \mathbb{E}(\mathbf{x}) + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \mathbb{E}(\mathbf{y})) \qquad (1) \\ \Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}. \qquad (2) \end{aligned}$$

By conditioning, we get

$$f_*|X_*, X, f \sim \mathcal{N}(K(X_*,X)K(X,X)^{-1}f, \\ K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*))$$

Prediction with Noise

Let $y(x) = f(x) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.

Then $\mathbf{cov}(y(x_p), y(x_q)) = K(x_p, x_q) + \sigma_n^2 \delta_{pq}$ or in a matrix form

$$\mathbf{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 \mathbb{I}$$

The joint distribution between y and f_* is

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(X,X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix}\right)$$

By conditioning, we get

$$f_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{f}_*, \mathrm{cov}(f_*)),$$

$$\bar{f}_* = K(X_*, X)\left(K(X,X) + \sigma_n^2 \mathbb{I}\right)^{-1} \mathbf{y}$$

$$\mathbf{cov}(f_*) = K(X_*, X_*) - K(X_*, X)\left(K(X,X) + \sigma_n^2 \mathbb{I}\right)^{-1} K(X, X_*)$$

Learning

Since $\mathbf{y} \sim \mathcal{N}(0, K + \sigma_n^2 \mathbb{I})$, the log marginal likelihood is

$$\log P(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T (K + \sigma_n^2 \mathbb{I})^{-1} \mathbf{y} - \frac{1}{2} \log |K + \sigma_n^2 \mathbb{I}| - \frac{n}{2} \log 2\pi,$$

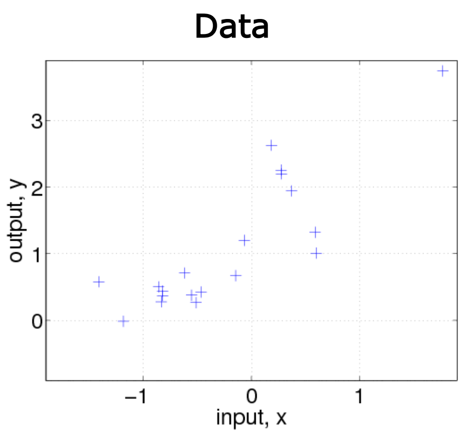
which can be used to estimate σ_n^2 and parameters for the kernel function (using a gradient based method).

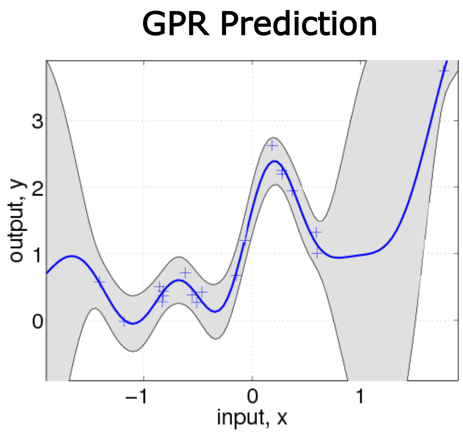
For example, if the following squared exponential kernel is used, the kernel parameters are (σ_f^2, σ_l^2).

$$K(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2\sigma_l^2} ||x_p - x_q||^2\right)$$

In practice, selecting the right kernel for a given problem is also an important task.

Comments on GPR

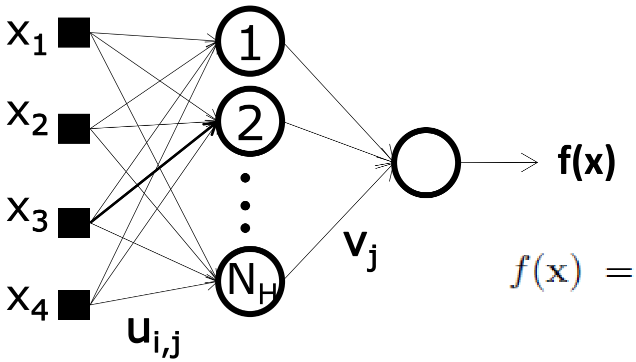




Pros: principled, probabilistic, predictive uncertainty
Cons: computationally intensive (n x n matrix inversion)

Neural Network

- Neural network with a single hidden layer with N_H units



$$f(\mathbf{x}) = b + \sum_{j=1}^{N_H} v_j h(\mathbf{x}; \mathbf{u}_j)$$

[Cybenko 1989, Hornik 1993]

- Neural network with one hidden layer is a **universal approximator** as $N_H \to \infty$
- That is, it can approximate any continuous function on a compact support under mild conditions.

NN Converges to a GP

$$\boxed{f(\mathbf{x}) = b + \sum_{j=1}^{N_H} v_j h(\mathbf{x};\mathbf{u}_j)}$$

Suppose that

- $b \sim (0, \sigma_b^2)$ and $v_j \sim (0, \sigma_v^2)$
- $\mathbf{u}_j$ are independently and identically distributed
- σ_v^2 scales as ω^2/N_H

$$\begin{aligned}
\mathbb{E}(f(\mathbf{x})) &= 0 \\
\mathbb{E}\left(f(\mathbf{x})f(\mathbf{x}')\right) &= \sigma_b^2 + \sum \sigma_v^2 \mathbb{E}_{\mathbf{u}}\left(h(\mathbf{x};\mathbf{u}_j)h(\mathbf{x}';\mathbf{u}_j)\right) \\
&= \sigma_b^2 + \omega^2 \mathbb{E}_{\mathbf{u}}\left(h(\mathbf{x};\mathbf{u}_j)h(\mathbf{x}';\mathbf{u}_j)\right)
\end{aligned}$$

[Neal 1996] By the central limit theorem, $f(\mathbf{x})$ converges to a Gaussian process as $N_H \to \infty$.
If $h(\mathbf{x};\mathbf{u}) = \operatorname{erf}(u_0 + \sum u_j x_j)$ and $\mathbf{u} \sim \mathcal{N}(0,\Sigma)$, then the covariance function of the neural network is

$$k_{NN}(\mathbf{x},\mathbf{x}') = \frac{2}{\pi}\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^T\Sigma\tilde{\mathbf{x}}'}{((1+2\tilde{\mathbf{x}}^T\Sigma\tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^T\Sigma\tilde{\mathbf{x}}'))^{1/2}}\right),$$

where $\tilde{\mathbf{x}} = [1 \ x_1 \ \ldots \ x_d]^T$.

Summary

Linear regression:

- Parametric regression method

Gaussian process regression:

- Nonparametric regression method
- Weight-space view: Bayesian approach to linear regression (with the kernel trick)
- Function-space view: MMSE estimate, linear predictor
- Provides the predictive variance for an unseen data
- Computationally intensive (for prediction, $O(n^3)$)
- A single hidden layer neural network converges to a Gaussian process