Gaussian Processes: An Introduction

Lili MOU

moull12@sei.pku.edu.cn
http://sei.pku.edu.cn/~moull12

9 April 2015
Outline

Introduction

Kernel Tricks

Gaussian Processes for Regression

Bayesian Linear Regression
Outline

Introduction

Kernel Tricks

Gaussian Processes for Regression

Bayesian Linear Regression
Warming up

- Let $Z_t$ be a Gaussian distribution with mean $\mu_t$ and standard deviation $\sigma_t \ (t \in T)$.
Warming up

Let $Z_t$ be a Gaussian distribution with mean $\mu_t$ and standard deviation $\sigma_t$ ($t \in \mathcal{T}$).

$$p(z_t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp \left\{ -\frac{(x - \mu_t)^2}{2\sigma_t^2} \right\}$$
Warming up

- Let $Z_t$ be a Gaussian distribution with mean $\mu_t$ and standard deviation $\sigma_t$ ($t \in T$).

$$p(z_t) = \frac{1}{\sqrt{2\pi \sigma^2_t}} \exp \left\{ - \frac{(x - \mu_t)^2}{2\sigma^2_t} \right\}$$

- If $Z$ independent, what is the joint distribution of $Z_{i_1}, \cdots, Z_{i_n}$?
Warming up

Let $Z_t$ be a Gaussian distribution with mean $\mu_t$ and standard deviation $\sigma_t$ ($t \in T$).

$$p(z_t) = \frac{1}{\sqrt{2\pi \sigma_t^2}} \exp \left\{ -\frac{(x - \mu_t)^2}{2\sigma_t^2} \right\}$$

If $Z$ independent, what is the joint distribution of $Z_{i1}, \ldots, Z_{in}$?

$$(Z_{i1}, \ldots, Z_{in})^T \sim \mathcal{N}(\mu, \Sigma)$$

where $\mu = (\mu_{i1}, \ldots, \mu_{in})^T$, $\Sigma = \text{diag}\{\sigma_{i1}, \ldots, \sigma_{in}\}$

$$p(z) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp \left\{ -\frac{1}{2}(z - \mu)^T\Sigma^{-1}(z - \mu) \right\}$$

If $Z$ dependent, what is the joint distribution?
Let $Z_t$ be a Gaussian distribution with mean $\mu_t$ and standard deviation $\sigma_t$ ($t \in T$).

$$p(z_t) = \frac{1}{\sqrt{2\pi\sigma^2_t}} \exp \left\{ -\frac{(x - \mu_t)^2}{2\sigma^2_t} \right\}$$

If $Z$ independent, what is the joint distribution of $Z_{i_1}, \ldots, Z_{i_n}$?

$$(Z_{i_1}, \ldots, Z_{i_n})^T \sim \mathcal{N}(\mu, \Sigma)$$

where $\mu = (\mu_{i_1}, \ldots, \mu_{i_n})^T$, $\Sigma = \text{diag}\{\sigma_{i_1}, \ldots, \sigma_{i_n}\}$

$$p(z) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp \left\{ -\frac{1}{2}(z - \mu)^T\Sigma^{-1}(z - \mu) \right\}$$

If $Z$ dependent, what is the joint distribution? Recall copulas.
Definition. A *stochastic process* is a set of random variables \( \{Z_t\} \), \( t \in T \). \( T \) is called an *index set*.

- Trivial process: \( Z_t \) independent
- Brownian process
- Poisson process

The relationships between \( Z_t \) are a distinguishing feature in the field of stochastic processes.
Definition. A Gaussian process \( \{ Z_t \} \), \( (t \in T) \) is a stochastic process, each subset of \( \{ Z_t \} \) forming a (multivariate) Gaussian.
Gaussian Processes

Definition. A Gaussian process \( \{Z_t\}, (t \in \mathcal{T}) \) is a stochastic process, each subset of \( \{Z_t\} \) forming a (multivariate) Gaussian.

A minor question: Why not model \( \{Z_t\} \) directly as a multivariate Gaussian?
Gaussian Processes

**Definition.** A *Gaussian process* \( \{Z_t\}, (t \in \mathcal{T}) \) is a stochastic process, each subset of \( \{Z_t\} \) forming a (multivariate) Gaussian.

A minor question: Why not model \( \{Z_t\} \) directly as a multivariate Gaussian?

- \( \mathcal{T} \) may have infinite elements (or even uncountable).
- What computers can deal with is finite Gaussian processes, degraded to multivariate Gaussian distributions.
An Example

Random lines: \( \mathcal{T} = \mathbb{R} \). \( \forall t \in \mathcal{T} \), let \( Z_t = t \cdot w \), where \( w \in \mathbb{R} \) and \( w \sim \mathcal{N}(w|0,1) \)

\[
\begin{pmatrix}
Z_{t_1} \\
\vdots \\
Z_{t_n}
\end{pmatrix} =
\begin{pmatrix}
t_1w \\
\vdots \\
t_nw
\end{pmatrix} =
\begin{pmatrix}
t_1 \\
\vdots \\
t_n
\end{pmatrix} w \sim \mathcal{N}
\]

This GP defines a linear function on \( \mathbb{R} \).
Consider a regression problem. Let a GP \( \{ Z_t \} \) define a random function (not necessarily linear), where \( t \) comes from an arbitrary index set \( T \) of the input space.

[Source: NIPS-06’s talk]

A prospective of Bayesianism
Outline

Introduction

Kernel Tricks

Gaussian Processes for Regression

Bayesian Linear Regression
Existence of Gaussian Processes

**Theorem.** For any index set \( \mathcal{T} \), any mean function \( \mu : \mathcal{T} \rightarrow \mathbb{R} \) and any covariance function \( k : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R} \), there exists a Gaussian process \( \{ Z_t \} \) on \( \mathcal{T} \) such that \( \mathbb{E}[Z_t] = \mu(t) \) and \( \text{cov}(Z_s, Z_t) = k(s, t), \forall s, t \in \mathcal{T} \).

\( \Rightarrow \) A Gaussian process is fully characterized by \( \mu \) and \( k \).

- \( k \) is also called a *kernel function*.
- When evaluated on a finite subset, \( k \) defines a *kernel matrix* \( K \).
- Mercer’s Theorem: If \( K \) is symmetric and positive semi-definite, then \( K \) can be represented as an inner-product in some Hilbert space.
Random Line Revisit

\[ \mathcal{T} = \mathbb{R}. \, \forall t \in \mathcal{T}, \, Z_t = t \cdot w, \text{ where } w \sim \mathcal{N}(w|0, 1) \]

- \[ \mu(t) = \mathbb{E}[z_t] = \mathbb{E}[t \cdot w] = t \cdot \mathbb{E}[w] = 0 \]

- \[ k(s, t) = \text{cov}(Z_s, Z_t) = \mathbb{E}[Z_s Z_t] - \mathbb{E}[Z_s] \mathbb{E}[Z_t] = s \cdot t \]

Note that

- \( \mu \) is the expectation of \( Z \) (indexed by \( t \)) rather than \( t \in \mathcal{T} \)

- So is \( \Sigma \).

- If \( \mu \) and \( k \) satisfy the above equations, for any finite subset \( \{Z_{t_1}, \ldots, Z_{t_n}\} \), \( \text{rank}(\Sigma) = 1 \). We are happy for that. ☺️
Kernels

- Standard Brownian motion
  \( \mathcal{T} = [0, \infty), \mu(t) = 0, k(s, t) = \min(s, t) \)
- Gaussian kernel \( \mathcal{T} = \mathbb{R}^d, \mu(t) = 0, k(x, y) = \exp\{-\alpha \|x - y\|^2\} \)
- Laplacian kernel \( \mathcal{T} = \mathbb{R}^d, \mu(t) = 0, k(x, y) = \exp\{-\alpha \|x - y\|\} \)
Kernels

- **Standard Brownian motion**
  \( \mathcal{T} = [0, \infty), \mu(t) = 0, k(s, t) = \min(s, t) \)

- **Gaussian kernel** \( \mathcal{T} = \mathbb{R}^d, \mu(t) = 0, k(x, y) = \exp\{-\alpha \|x - y\|^2\} \)

- **Laplacian kernel** \( \mathcal{T} = \mathbb{R}^d, \mu(t) = 0, k(x, y) = \exp\{-\alpha \|x - y\|\} \)

Basis expansion for the Gaussian kernel

\[
k(x_1, x_2) = \exp \left\{ -x_1^2 - x_2^2 + 2x_1x_2 \right\} \\
= \exp \left\{ -x_1^2 \right\} \exp \left\{ -x_2^2 \right\} \sum_{k=0}^{\infty} \frac{2^k x_1^k x_2^k}{k!}
\]

\( \Phi : x \mapsto \left( \sqrt{\frac{2}{1}} \cdot \frac{x^0}{\exp\{-x^2\}}, \sqrt{\frac{2^2}{2!}} \cdot \frac{x^1}{\exp\{-x^2\}}, \sqrt{\frac{3^2}{3!}} \cdot \frac{x^2}{\exp\{-x^2\}}, \ldots \right) \)
Operations on Kernels

Let $k, k_1, k_2$ be valid kernels, and $x, y \in \mathcal{T}$. The followings are also valid kernels.

- $\alpha k(x, y)$
- $k_1(x, y) + k_2(x, y)$
- $k_1(x, y)k_2(x, y)$
- $p(k(x, y))$, where $p$ is a polynomial with non-negative coefficients
- $\exp\{k(x, y)\}$
- $f(x)k(x, y)f(y)$, $\forall f : \mathcal{T} \rightarrow \mathbb{R}$, or $f : \mathcal{T} \rightarrow \mathbb{C}$
- $k(\psi(x), \psi(y))$, $\forall \psi : \mathcal{T} \rightarrow S$
Examples

[Source: Pattern Recognition and Machine Learning]
Generating the Random Functions

To generate the previous beautiful figures, i.e., random functions defined by $GP(\mu, k)$, we need to
Generating the Random Functions

To generate the previous beautiful figures, i.e., random functions defined by $\mathcal{GP}(\mu, k)$, we need to

- Take discrete points $Z_{x_1}, \cdots, Z_{x_n}$ in an interval
- Sample $z_{x_1}, \cdots, z_{x_n}$ from $\mathcal{N}(\mu, \Sigma)$
- Interpolate, which is valid intuitively as long as the kernel is “smooth.”
Outline

Introduction

Kernel Tricks

Gaussian Processes for Regression

Bayesian Linear Regression
The Gaussian Process Model for Regression Problem

To predict \( \{y(i)\}_{i=1}^m \) given \( \{x(i)\}_{i=1}^m \), with \( \{x(i), y(i)\}_{i=m+1}^{m+n} \) known (\( n \) training samples, \( m \) test samples)

- Assume \( Y(i) = Z(i) + \epsilon(i) \), where \( \epsilon(i) \) is the random noise

\[
\epsilon(i) \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2)
\]

i.e.,

\[
\epsilon \sim \mathcal{N}(0, \sigma^2 I)
\]

- Assume \( \{Z_x\} \) is a \( \mathcal{GP}(\mu, k) \), where \( x \in \mathcal{T} \).
  \( \mathcal{T} \) is the sample space, which is arbitrary. (Think of \( \mathbb{R}^d \))

- As we always have finite samples,

\[
Z \sim \mathcal{N}(\mu, K)
\]

where \( \mu \) and \( K \) are defined by \( \mathcal{GP}(\mu, k) \), evaluated at \( X = (X^{(1)}, \ldots, X^{(m)}, X^{(m+1)}, \ldots, X^{(n+m)}) \)
Inference

Let $Y_a = \{ y^{(i)} \}_{i=1}^m$ (test set), and $Y_b = \{ y^{(i)} \}_{i=m+1}^{m+n}$ (training set).

What is the distribution of $Y_a | Y_b = y_b$?
Inference

Let $\mathbf{Y}_a = \{y^{(i)}\}_{i=1}^m$ (test set), and $\mathbf{Y}_b = \{y^{(i)}\}_{i=m+1}^{m+n}$ (training set).

What is the distribution of $\mathbf{Y}_a | \mathbf{Y}_b = \mathbf{y}_b$? Gaussian!
Inference

Let \( \mathbf{Y}_a = \{y^{(i)}\}_{i=1}^m \) (test set), and \( \mathbf{Y}_b = \{y^{(i)}\}_{i=m+1}^{m+n} \) (training set).

What is the distribution of \( \mathbf{Y}_a|\mathbf{Y}_b = \mathbf{y}_b \)? **Gaussian!**

\[ \mathbf{Y} = \mathbf{Z} + \mathbf{\epsilon}, \text{ where } \mathbf{Z} \sim \mathcal{N}(\mu, \mathbf{K}) \text{ and } \mathbf{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \]

\( \mathbf{Z} \) and \( \mathbf{\epsilon} \) are independent

\[ \mathbf{Y} \sim \mathcal{N}(\mu, \mathbf{K} + \sigma^2 \mathbf{I}) \triangleq \mathcal{N}(\mu, \mathbf{C}) \]

Denote \( \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \), then

\[ \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \text{ and } \mathbf{C} = \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix} \]

The solution is analytic!
Conditional Gaussian Distributions

Let \( \mathbf{Y} \sim \mathcal{N}(\mu, \mathbf{C}) \), and partition it into two parts

\[
\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix}
\]

What is the distribution of \( \mathbf{Y}_a \) given \( \mathbf{Y}_b = \mathbf{y}_b \)?

- Gaussian! \( \mathcal{N}(\mathbf{m}, \mathbf{D}) \)
Conditional Gaussian Distributions

Let $Y \sim \mathcal{N}(\mu, C)$, and partition it into two parts

$$Y = \begin{pmatrix} Y_a \\ Y_b \end{pmatrix}, \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, C = \begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix}$$

What is the distribution of $Y_a$ given $Y_b = y_b$?

- Gaussian! $\mathcal{N}(m, D)$
- $m = \mu_a + C_{ab}C_{bb}^{-1}(y_b - \mu_b)$
- $D = C_{aa} - C_{ab}C_{bb}^{-1}C_{ba}$

For GP regression,

$C_{aa} = K_{aa} + \sigma^2 I$, $C_{ab} = K_{ab}$, $C_{ba} = K_{ba}$, $C_{bb} = K_{bb} + \sigma^2 I$

More realistically, $\mu = 0$, and thus

$m = K_{ab}(K_{bb} + \sigma^2 I)^{-1}y_b - \mu_b$
Conditional Gaussian Distributions

Let \( Y \sim \mathcal{N}(\mu, C) \), and partition it into two parts

\[
Y = \begin{pmatrix} Y_a \\ Y_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad C = \begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix}
\]

What is the distribution of \( Y_a \) given \( Y_b = y_b \)?

- Gaussian! \( \mathcal{N}(m, D) \)
- \( m = \mu_a + C_{ab}C_{bb}^{-1}(y_b - \mu_b) \)
- \( D = C_{aa} - C_{ab}C_{bb}^{-1}C_{ba} \)

For GP regression,

- \( C_{aa} = K_{aa} + \sigma^2 I, \quad C_{ab} = K_{ab}, \quad C_{ba} = K_{ba}, \quad C_{bb} = K_{bb} + \sigma^2 I \)

More realistically, \( \mu = 0 \), and thus

- \( m = K_{ab}(K_{bb} + \sigma^2 I)^{-1}y_b \)
Conditional Gaussian Distributions

Let $Y \sim \mathcal{N}(\mu, C)$, and partition it into two parts

$$Y = \begin{pmatrix} Y_a \\ Y_b \end{pmatrix}, \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, C = \begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix}$$

What is the distribution of $Y_a$ given $Y_b = y_b$?

- Gaussian! $\mathcal{N}(m, D)$
- $m = \mu_a + C_{ab}C_{bb}^{-1}(y_b - \mu_b)$
- $D = C_{aa} - C_{ab}C_{bb}^{-1}C_{ba}$

For GP regression,

- $C_{aa} = K_{aa} + \sigma^2 I$, $C_{ab} = K_{ab}$, $C_{ba} = K_{ba}$, $C_{bb} = K_{bb} + \sigma^2 I$

More realistically, $\mu = 0$, and thus

- $m = K_{ab}(K_{bb} + \sigma^2 I)^{-1}y_b$

$Y_b$ dependent even given $X_b$?
Outline

Introduction

Kernel Tricks

Gaussian Processes for Regression

Bayesian Linear Regression
Linear Regression

Let $\phi(x)$ be a set of basis functions. The target variable $y$ is a linear combination of $\phi(x)$ with coefficients $w$, plus a Gaussian noise.

$$p(y|x, w) = \mathcal{N}(y|w^T \phi(x), \beta^{-1})$$

$$p(y|x, w) = \mathcal{N}(y|\Phi w, \beta^{-1}I)$$

[Modified from Pattern Recognition and Machine Learning.]
Frequentism v.s. Bayesianism

▶ Frequentism
- Estimate $w^* = \operatorname*{arg\,max}_w p(y|x; w)$
- Predict $\hat{p}(y^{(t)}|x^{(t)}) = p(y^{(t)}|x^{(t)}; w^*)$

▶ Bayesianism
- Have some prior $p(w)$ on $w$
- Adjust our belief with data $\mathcal{D} = \{x, y\}$

$$p(w|\mathcal{D}) = \frac{p(w)p(\mathcal{D}|w)}{p(\mathcal{D})}$$

- Derive the predictive density

$$p(y^{(t)}|x^{(t)}, \mathcal{D}) = \int_W p(y^{(t)}|w)p(w|\mathcal{D}) \, dw$$
Frequentism v.s. Bayesianism

▶ Frequentism
  - Estimate $\mathbf{w}^* = \operatorname{argmax}_w p(\mathbf{y} | \mathbf{x}; \mathbf{w})$
  - Predict $\hat{p}(y^{(t)} | x^{(t)}) = p(y^{(t)} | x^{(t)}; \mathbf{w}^*)$

▶ Bayesianism
  ▶ Have some prior $p(\mathbf{w})$ on $\mathbf{w}$
  ▶ Adjust our belief with data $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$

$$p(\mathbf{w} | \mathcal{D}) = \frac{p(\mathbf{w}) p(\mathcal{D} | \mathbf{w})}{p(\mathcal{D})}$$

▶ Derive the predictive density

$$p(y^{(t)} | x^{(t)}, \mathcal{D}) = \int_{\mathcal{W}} p(y^{(t)} | \mathbf{w}) p(\mathbf{w} | \mathcal{D}) \, d\mathbf{w}$$

Note that

▶ Mathematicians are happy 😊 if prior and posterior distributions take the same form. (Called conjugate priors.)
▶ Most problems do not have closed-form solutions.
Bayesian Linear Regression

- Likelihood function (with $x$ omitted for clarity)

$$p(y|w) = \mathcal{N}(y|\Phi w, \beta^{-1} I)$$

- Prior

$$p(w) = \mathcal{N}(w|m_0, S_0)$$

- Posterior

$$p(w|y) = \mathcal{N}(w|m_N, S_N)$$

where

$$m_N = S_N(S_0^{-1}m_0 + \beta\Phi^T y)$$
$$= \beta S_N\Phi^T y$$
$$S_N = (S_0^{-1} + \beta\Phi^T \Phi)^{-1}$$

The subscript $N$ denotes the number of samples seen. In practice, $m_0 = 0$. 
The Predictive Density

Cheat sheet

\[ p(y|w) = \mathcal{N}(y|\Phi w, \beta^{-1} I) \]
\[ p(w|y) = \mathcal{N}(w|m_N, S_N) \]

\[ p(y^{(t)}|y, \alpha, \beta) = \int p(y^{(t)}|w, \beta)p(w|y, \alpha, \beta) \, d\, w \]
\[ = \int \mathcal{N}(y^{(t)}|\Phi w, \beta^{-1} I)\mathcal{N}(w|m_N, S_N) \, d\, w \]
\[ \propto \exp \{ \cdot \} \exp \{ \cdot \} \, d\, w \]
\[ \propto \int \mathcal{N}(w|\cdot)g(y) \, d\, w \]
\[ = g(y) \int \mathcal{N}(w|\cdot) \, d\, w \]
\[ \propto \mathcal{N}(y|\cdot) \]
Predictive Density

\[ p(y^{(t)}|y) = \mathcal{N}(y^{(t)}|m_N \Phi(x), \sigma^2_N(x)) \]

where

\[ \sigma^2_N(x) = \frac{1}{\beta} + \Phi(x)^T S_N \Phi(x) \]
An Example of Predictive Density with RBF Bases

[Source: Pattern Recognition and Machine Learning]
Function samples $y(x, w)$ Dawn from the Posterior over $w$

[Source: Pattern Recognition and Machine Learning]
The Equivalent Kernel

The predicted density has mean

\[ \mathbb{E}[y(x)] = \mathbb{E}[\phi(x)^T w] \]
\[ = \phi(x)^T m_N \]
\[ = \beta \phi(x)^T S_N \Phi^T y \]
\[ = \sum_{n=1}^{N} \beta \phi(x)^T S_N \phi(x_n) y_n \]
\[ \Delta = \sum_{n=1}^{N} k(x, x_n) y_n \]

where \( k(x, x') = \beta \phi(x)^T S_N \phi(x') \), depending on the input \( X \)

\[ \text{cov} (y(x), y(x')) = \text{cov} \left( \phi(x)^T w, w^T \phi(x') \right) \]
\[ = \phi(x)^T S_N \phi(x') \]
\[ = \beta^{-1} k(x, x') \]
Gaussian Process and Bayesian Linear Regression

- In a Gaussian process regression, the predictive density has mean
  \[ m = K_{ab}(K_{bb} + \sigma^2 I)^{-1}y_b \]

- In Bayesian linear regression,
  \[ m = \phi(x)^T \left( \begin{pmatrix} \alpha \\ \beta \end{pmatrix} I + \Phi^T \Phi \right)^{-1} \Phi^T y \]

My notes

- Both Gaussian process regression and Bayesian linear regression stem from a prospective of Bayesianism, taking similar forms.

- Provided a training set, Bayesian linear regression can be fully represented by an equivalent kernel, which inspires the Gaussian process regression.

- However, the two models seems to be NOT equivalent in general.

Disclaimer: If I were wrong, please feel free to tell me.
References

- *Pattern Recognition and Machine Learning*
- *Machine Learning: A Probabilistic Prospective*
- [https://www.youtube.com/user/mathematicalmonk](https://www.youtube.com/user/mathematicalmonk)