Gaussian Processes: An Introduction

Lili MOU

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

9 April 2015

Outline

Outline

Warming up

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

Warming up

- Let Z_t be a Gaussian distribution with mean μ_t and standard deviation σ_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 μ_t and standard deviation σ_t $(t \in \mathcal{T})$.

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

- 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 μ_t and standard deviation σ_t ($t \in \mathcal{T}$).

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

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

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

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

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

- If Z dependent, what it the joint distribution?

Warming up

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

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

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

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

where $\boldsymbol{\mu} = (\mu_{i_1}, \cdots, \mu_{i_n})^T, \Sigma = \mathrm{diag}\{\sigma_{i_1}, \cdots, \sigma_{i_n}\}$

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

- If Z dependent, what it the joint distribution? Recall copulas.

Stochastic Processes

Definition. A *stochastic process* is a set of random variables $\{Z_t\}$, $t \in \mathcal{T}$. $\mathcal{T}$ 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.

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.

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.

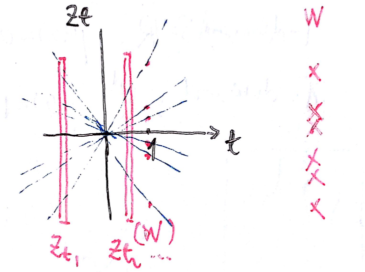
A minor question: Why not model $\{Z_t\}$ directly as a multivariate Gaussian?

- $\mathcal{T}$ 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 $w \sim \mathcal{N}(w|0,1)$

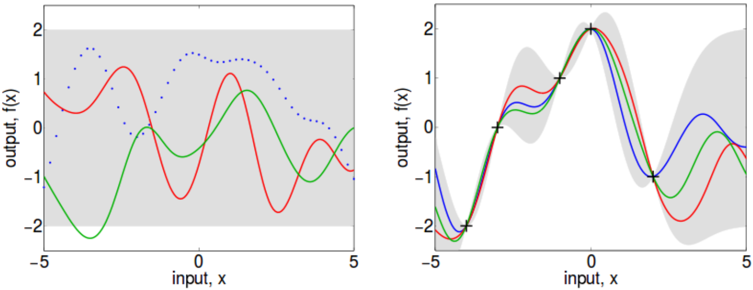
$$\begin{pmatrix} z_{t_1} \\ \vdots \\ z_{t_n} \end{pmatrix} = \begin{pmatrix} t_1 w \\ \vdots \\ t_n w \end{pmatrix} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} w \quad \sim \mathcal{N}$$



This GP defines a linear function on $\mathbb{R}$.

A Big Picture

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



[Source: NIPS-06's talk]

A prospective of Bayesianism

Outline

Existence of Gaussian Processes

Theorem. For any index set $\mathcal{T}$, any mean function $\mu : \mathcal{T} \to \mathbb{R}$ and any covariance function $k : \mathcal{T} \times \mathcal{T} \to \mathcal{R}$, 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 μ 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

$\mathcal{T} = \mathbb{R}$. $\forall t \in \mathcal{T}$, $Z_t = t \cdot w$, 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

- μ is the expectation of Z (indexed by t) rather than $t \in \mathcal{T}$
- So is Σ.

- If μ and k satisfy the above equations, for any finite subset $\{Z_{t_1}, \cdots, Z_{t_n}\}$, $\text{rank}(\Sigma) = 1$. 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

$$\begin{aligned} k(x_1, x_2) &= \exp\left\{-x_1^2 - x_2^2 + 2x_1 x_2\right\} \\ &= \exp\left\{-x_1^2\right\} \exp\left\{-x_2^2\right\} \sum_{k=0}^{\infty} \frac{2^k x_1^k x_2^k}{k!} \end{aligned}$$

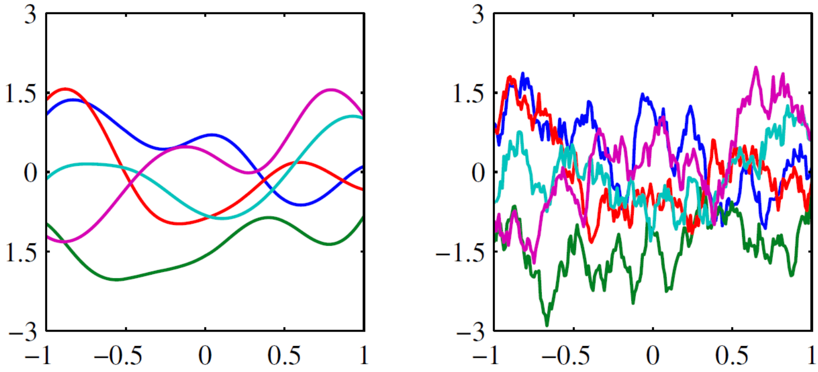
$$\Phi : x \mapsto \left(\sqrt{\frac{2}{1}} \cdot \frac{x^0}{\exp\{-x^2\}}, \sqrt{\frac{2^2}{2!}} \cdot \frac{x^1}{\exp\{-x^2\}}, \sqrt{\frac{3^2}{3!}} \cdot \frac{x^2}{\exp\{-x^2\}}, \dots \right)$$

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
- $\exp\{k(x, y)\}$
- $f(x) k(x, y) \overline{f(y)}$, $\forall f : \mathcal{T} \to \mathbb{R}$, or $f : \mathcal{T} \to \mathbb{C}$
- $k(\psi(x), \psi(y))$, $\forall \psi : \mathcal{T} \to \mathcal{S}$

Examples



[Source: *Pattern Recognition and Machine Learning*]

Generating the Random Functions

To generate the previous beautiful figures, i.e., random functions defined by $\mathcal{GP}(\mu, k)$, 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 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$
- Interpolate, which is valid intuitively as long as the kernel is "smooth."

Outline

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

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

  i.e.,

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

- 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,

$$\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$$

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

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 $\mathbf{Y}_a | \mathbf{Y}_b = \mathbf{y}_b$?

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!

- $\mathbf{Y} = \mathbf{Z} + \boldsymbol{\epsilon}$, where $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
  $\mathbf{Z}$ and $\boldsymbol{\epsilon}$ are independent
- $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K} + \sigma^2 \mathbf{I}) \triangleq \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$
- Denote $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix}$, then
  $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$, 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 $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, and partition it into two parts

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix}$$

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

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix}$$

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

- Gaussian! $\mathcal{N}(\mathbf{m}, \mathbf{D})$
- $\mathbf{m} = \boldsymbol{\mu}_a + \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}(\mathbf{y}_b - \boldsymbol{\mu}_b)$
- $\mathbf{D} = \mathbf{C}_{aa} - \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}\mathbf{C}_{ba}$

Conditional Gaussian Distributions

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

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix}$$

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

- Gaussian! $\mathcal{N}(\mathbf{m}, \mathbf{D})$
- $\mathbf{m} = \boldsymbol{\mu}_a + \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}(\mathbf{y}_b - \boldsymbol{\mu}_b)$
- $\mathbf{D} = \mathbf{C}_{aa} - \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}\mathbf{C}_{ba}$

For GP regression,

- $\mathbf{C}_{aa} = \mathbf{K}_{aa} + \sigma^2\mathbf{I}, \mathbf{C}_{ab} = \mathbf{K}_{ab}, \mathbf{C}_{ba} = \mathbf{K}_{ba}, \mathbf{C}_{bb} = \mathbf{K}_{bb} + \sigma^2\mathbf{I}$

More realistically, $\boldsymbol{\mu} = \mathbf{0}$, and thus

- $\mathbf{m} = \mathbf{K}_{ab}(\mathbf{K}_{bb} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_b$

Conditional Gaussian Distributions

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

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix}$$

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

- Gaussian! $\mathcal{N}(\mathbf{m}, \mathbf{D})$
- $\mathbf{m} = \boldsymbol{\mu}_a + \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}(\mathbf{y}_b - \boldsymbol{\mu}_b)$
- $\mathbf{D} = \mathbf{C}_{aa} - \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}\mathbf{C}_{ba}$

For GP regression,

- $\mathbf{C}_{aa} = \mathbf{K}_{aa} + \sigma^2\mathbf{I}, \mathbf{C}_{ab} = \mathbf{K}_{ab}, \mathbf{C}_{ba} = \mathbf{K}_{ba}, \mathbf{C}_{bb} = \mathbf{K}_{bb} + \sigma^2\mathbf{I}$

More realistically, $\boldsymbol{\mu} = \mathbf{0}$, and thus

- $\mathbf{m} = \mathbf{K}_{ab}(\mathbf{K}_{bb} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_b$

$\mathbf{Y}_b$ dependent even given $\mathbf{X}_b$?

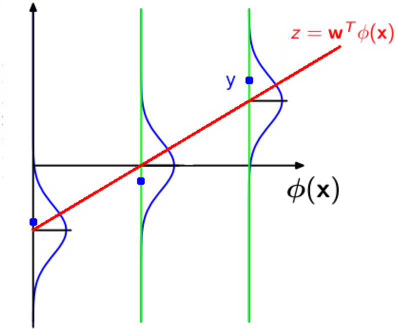
Outline

Linear Regression

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

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

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



[Modified from *Pattern Recognition and Machine Learning*.]

Frequentism v.s. Bayesianism

- Frequentism
  - Estimate $\mathbf{w}^* = \text{argmax}_{\mathbf{w}}\, p(\mathbf{y}|\mathbf{x};\mathbf{w})$
  - 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

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

Frequentism v.s. Bayesianism

- Frequentism
  - Estimate $\mathbf{w}^* = \operatorname{argmax}_{\mathbf{w}} p(\mathbf{y}|\mathbf{x};\mathbf{w})$
  - 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

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

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 $\mathbf{x}$ omitted for clarity)

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

- Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

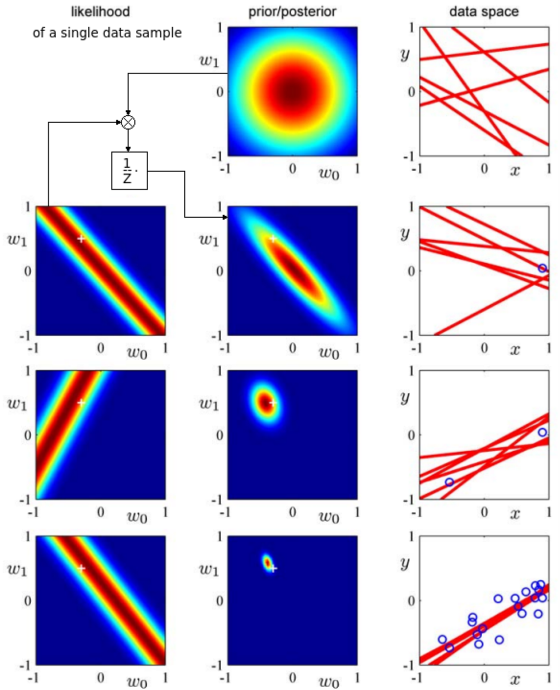
- Posterior

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

where

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{y}) \\
&= \beta\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{y} \\
\mathbf{S}_N &= (\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}
\end{aligned}$$

The subscript N denotes the number of samples seen.
In practice, $\mathbf{m}_0 = \mathbf{0}$.



[Modified from
Pattern Recognition and Machine Learning]

The Predictive Density

Cheat sheet

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

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\begin{aligned}
p(y^{(t)}|\mathbf{y}, \alpha, \beta) &= \int p(y^{(t)}|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{y}, \alpha, \beta) \, \mathrm{d}\,\mathbf{w} \\
&= \int \mathcal{N}(y^{(t)}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}) \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \, \mathrm{d}\,\mathbf{w} \\
&\propto \exp\{\cdot\} \exp\{\cdot\} \, \mathrm{d}\,\mathbf{w} \\
&\propto \int \mathcal{N}(\mathbf{w}|\cdot) g(y) \, \mathrm{d}\,\mathbf{w} \\
&= g(y) \int \mathcal{N}(\mathbf{w}|\cdot) \, \mathrm{d}\,\mathbf{w} \\
&\propto \mathcal{N}(y|\cdot)
\end{aligned}$$

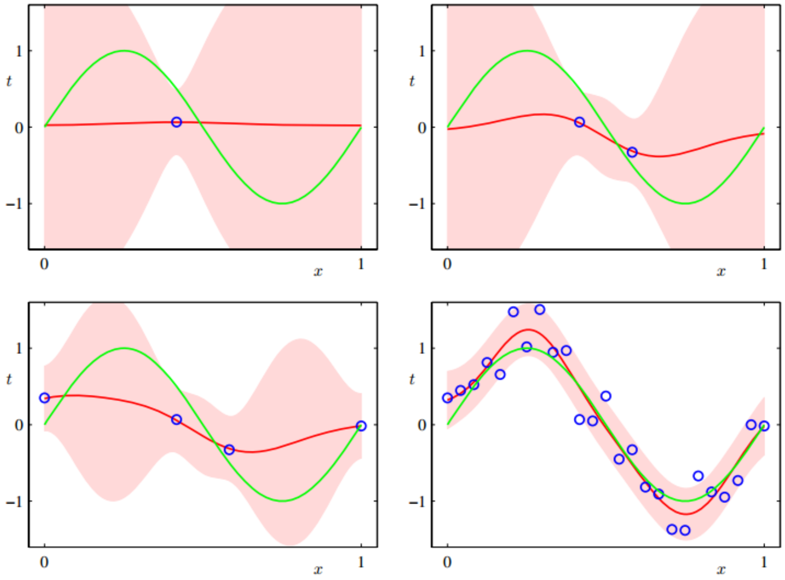
Predictive Density

$$p(y^{(t)}|\mathbf{y}) = \mathcal{N}(y^{(t)}|\mathbf{m}_N^T\phi(x), \sigma_N^2(x))$$

where

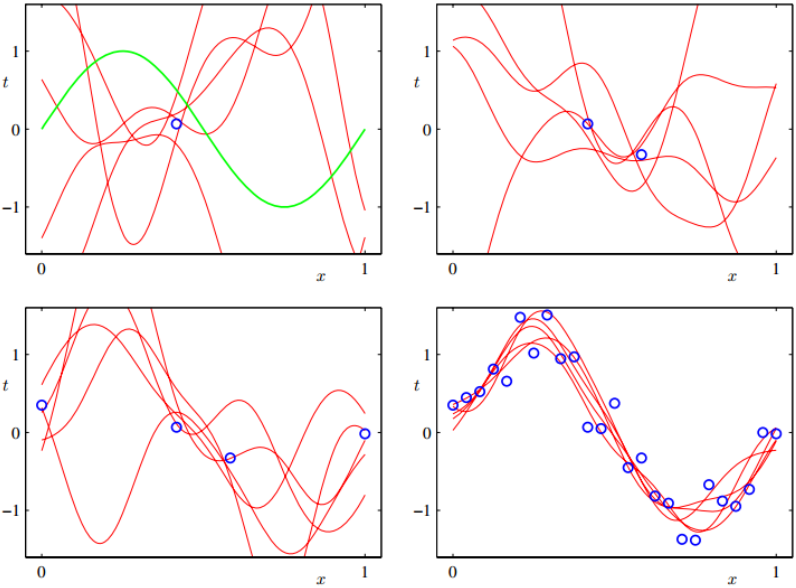
$$\sigma_N^2(x) = \frac{1}{\beta} + \mathbf{\Phi}(\mathbf{x})^T\mathbf{S}_N\mathbf{\Phi}(x)$$

An Example of Predictive Density with RBF Bases



[Source: *Pattern Recognition and Machine Learning*]

Function samples $y(x, \mathbf{w})$ Dawn from the Posterior over $\mathbf{w}$



[Source: *Pattern Recognition and Machine Learning*]

The Equivalent Kernel

The predicted density has mean

$$\begin{aligned}
\mathbb{E}[y(x)] &= \mathbb{E}[\phi(x)^T \mathbf{w}] \\
&= \phi(x)^T \mathbf{m}_N \\
&= \beta \phi(x)^T \mathbf{S}_N \mathbf{\Phi}^T \mathbf{y} \\
&= \sum_{n=1}^{N} \beta \phi(x)^T \mathbf{S}_N \phi(x_n) y_n \\
&\triangleq \sum_{n=1}^{N} k(x, x_n) y_n
\end{aligned}$$

where $k(x, x') = \beta \phi(x)^T \mathbf{S}_N \phi(x')$, depending on the input $\mathbf{X}$

$$\begin{aligned}
\operatorname{cov}\left(y(x), y(x')\right) &= \operatorname{cov}\left(\phi(x)^T \mathbf{w}, \mathbf{w}^T \phi(x')\right) \\
&= \phi(x)^T \mathbf{S}_N \phi(x') \\
&= \beta^{-1} k(x, x')
\end{aligned}$$

Gaussian Process and Bayesian Linear Regression

- In a Gaussian process regression, the predictive density has mean

$$m = \mathbf{K}_{ab}(\mathbf{K}_{bb} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}_b$$

- In Bayesian linear regression,

$$m = \phi(x)^T \left(\frac{\alpha}{\beta} \mathbf{I} + \mathbf{\Phi}^T \mathbf{\Phi} \right)^{-1} \mathbf{\Phi}^T \mathbf{y}$$

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*
- http://www.gaussianprocess.org/
- https://www.youtube.com/user/mathematicalmonk