Bayesian Nonparametric regression, especially Gaussian Process Regression

Bayesian statistics review, Parametric / semiparametric / nonparametric regression, Gaussian process regression, GPR and BNN

Sun Woo Lim

May 12, 2022

Outline

1. Introduction (Motivation)
2. Basics of Bayesian Statistics
3. Parametric and nonparametric model
4. Bayesian linear model
5. Bayesian basis function model
6. Gaussian process regression (GPR)

Outline

1. Introduction (Motivation)
2. Basics of Bayesian Statistics
3. Parametric and nonparametric model
4. Bayesian linear model
5. Bayesian basis function model
6. Gaussian process regression (GPR)

Motivation

Bayesian Nonparametric Regression : Doing "nonparametric regression" in a "Bayesian" way.

Nonparametric regression : corresponds to "what". (easily generalized into nonparametric classification too)

- Why nonparametric regression?

- **regression** for function estimation between numeric response and predictors
- **nonparametric model** : 1) scalable (dimension can get arbitrarily large as $n \to \infty$) or 2) really ∞ $\#$ of params
  $\leftrightarrow$ **parametric model** : model complexity bounded although $n \to \infty$
- try to model **complex & flexible** predictor-response relationship.
- e.g) basis function models, tree models, Gaussian process regression (GPR), neural network models

Main goal today: **GPR**, most widely used tool for nonparametric Bayesian regression. Then, briefly introduce relationships between **GPR** and **Bayesian neural network**.
GPR is an extension of **basis function model** and **basis function model** is an extension of **linear model**.

Bayesian : corresponds to "how"

- Why Bayes?

- simple framework : sum rule $\int p(x, y)dy = p(x)$ & product rule $p(x, y) = p(x)p(y|x)$
- easy framework of uncertainty quantification: intuitive procedure of interval estimation

Outline

1 Introduction (Motivation)

2 Basics of Bayesian Statistics

3 Parametric and nonparametric model

4 Bayesian linear model

5 Bayesian basis function model

6 Gaussian process regression (GPR)

Comparison with Frequentist view of statistics

Frequentist view of statistics

1. **Parameter** : unknown, but fixed. So, has no (meaningful) distribution, just a degenerate R.V.
2. **Main task in parameter estimation** : Form a statistic $T : [X_1, ..., X_n] \to \mathbb{R}$. Statistic used for estimation : estimator.
3. **Interval estimation** : Use Confidence interval $C_1(y)$ s.t. $p(\theta \in C_1(y)|\theta) = 1 - \alpha$ using distribution of data.
4. **Prediction of new data (R.V)** x_{new} : $X_{new} \sim f(x; \hat{\theta})$, where $\hat{\theta}$ is **optimal θ, usually MLE**.

Bayesian view of statistics

1. **Parameter** : unknown and random vector. So, has meaningful distributions (Prior and posterior)
2. **Main task in parameter estimation** : Derive posterior distribution
    1. "Prior" $p(\theta)$: assigns prior belief on the parameters.
    2. "Likelihood of parameter θ" : $p(\mathcal{D}|\theta)$
    3. "Posterior" $p(\theta|\mathcal{D}) = \frac{p(\theta,\mathcal{D})}{\int_\theta p(\theta,\mathcal{D})d\theta}$, or $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$: "posterior" $\propto$ "prior" $\times$ "likelihood"
3. **Interval estimation** : Use Credible interval $C_2(y)$ s.t. $p(\theta \in C_2(y)|y) = 1 - \alpha$ using posterior distribution.
4. **Prediction of new data (R.V)** x_{new} : Use **posterior predictive distribution**
   $x_{new} \sim p(x_{new}|\mathcal{D}) = \int_\theta p(x_{new}|\theta,\mathcal{D})p(\theta|\mathcal{D})d\theta = \int_\theta p(x_{new}|\theta)p(\theta|\mathcal{D})d\theta$, which is **averaging**.

   * Predictive distribution : for quantities that are observable!

Framework of Bayesian Statistics

Steps of Bayesian Data Analysis

1. **Setup full probability model**: $p(\theta, y) = p(\theta)p(y|\theta)$: joint distribution of observable and latent variables.
2. **Obtain posterior**: $p(\theta|y) = \frac{p(\theta,y)}{\int_\theta p(\theta,y)d\theta}$, or $p(\theta|y) \propto p(\theta)p(y|\theta)$: computation. "posterior" $\propto$ "prior" × "likelihood"
3. **Evaluate model fit, sensitivity analysis, Decision Analysis**: judgment.

Ways to obtain Posterior : each is fundamental part of Bayesian method, but not focus of today.

1. Closed form obtained by analytic calculation due to conjugacy
2. Monte Carlo (Independent Monte Carlo & Markov Chain Monte Carlo (MCMC))) : obtain samples from posterior
3. Deterministic approximation (e.g, Laplace method, Variational Bayes, Expectation propagation)

Prediction: "predictive distribution": for a quantity that is **observable**!

Prior predictive $p(y) = \int_\theta p(y, \theta)d\theta = \int_\theta p(\theta)p(y|\theta)d\theta$.

Posterior predictive $p(\tilde{y}|y)$
Let $\tilde{y}$: "unknown observable" that is conditionally independent given θ.
$p(\tilde{y}|y) = \int_\theta p(\tilde{y}, \theta|y)d\theta = \int_\theta p(\tilde{y}|\theta, y)p(\theta|y)d\theta = p(\tilde{y}|\theta)p(\theta|y)d\theta$.

Example of steps of Bayesian Data Analysis (BDA3 1.4)

Situation: Somebody spelled "radom":seems awkward. Let the intended word is one of "random", "radon", "radom".

1st step: Full probability Model

1. **prior**

   θ:= discrete R.V taking value 1("random"), 2("radon"), or 3("radom") representing **intended word**.

   Obtain prior by frequency of those words in Google database.

   Relative frequency of "random": $7.6e-5$, "radon" $6.1e-6$, "radom":$3.1e-7$. Normalize them to make them $\sum = 1$!

2. **Likelihood**

   Use a Google (contextual) model that infers $p('radom'|\Theta)$.
   $\rightarrow p('radom'|\theta = 1) = 0.00193$, $p('radom'|\theta = 2) = 0.000143$, $p('radom'|\theta = 3) = 0.975$.

2nd step: Posterior

$p(\theta = 1|'radom') = 0.325, p(\theta = 2|'radom') = 0.002, p(\theta = 3|'radom') = 0.673$. What is your decision?

3rd step: Model Judgment depending on domain (context)

1) Your decision? Any reason not to choose the intended word as "radom"?

2) Model fit? Additional information to incorporate in prior?

Bayesian model selection

Setting

- Data $\mathcal{D}$ and parameter (either continuous or discrete) θ. For simplicity, two candidate models M_1, M_2.
- A discrete R.V α: only takes two values 0 (: M_1) or 1 (: M_2). The exact value of α does not matter.

Posterior for the model $p(\alpha|\mathcal{D})$ is the goal of the analysis!

Steps to obtain $p(\alpha|\mathcal{D})$

1. **prior for the model choice**: $p(\alpha)$. Common choice: equal probability.
2. **Prior of θ conditional on α**: $\theta|\alpha = 0 \sim p(\theta|\alpha = 0)$, $\theta|\alpha = 1 \sim p(\theta|\alpha = 1)$.
3. **Likelihood of the data**: $p(\mathcal{D}|\theta, \alpha)$. Keep noticing that α can take either 0 or 1.
4. **Calculate Marginal likelihood (= evidence)**: $p(\mathcal{D}|\alpha) = \int_\theta p(\mathcal{D}|\theta, \alpha)p(\theta|\alpha)d\theta$
5. **Posterior of the model**: $p(\alpha|\mathcal{D}) \propto p(\mathcal{D}|\alpha)p(\alpha)$.

The posterior odds ratio tells us relative appropriateness of two models.

$\frac{Pr(\alpha=0|\mathcal{D})}{Pr(\alpha=1|\mathcal{D})} = \frac{Pr(\alpha=0)}{Pr(\alpha=1)} \times \frac{p(\mathcal{D}|\alpha=0)}{p(\mathcal{D}|\alpha=1)}$: **Posterior odds = Prior odds × Bayes factor**

Bayesian Occam's Razor

Complex model has wide support → marginal probability of certain event is small.
If $\mathcal{D}$ is in the support of simpler model, marginal likelihood will prefer simpler one!

Example) Chapter 28 in MacKay, Information theory, inference, and learning algorithms
Data: $\mathcal{D} = (-1, 3, 7, 11)$

- Model1 (Linear): $a_n = \beta n + (\alpha - \beta)$ indicating that $\alpha = a_1, n+1 = a_n + \beta$. Assume $\alpha, \beta \in \mathbb{Z}$
- Model2 (cubic): $a_1 = a, a_{n+1} = ba_n^3 + ca_n^2 + d$. Assume $a \in \mathbb{Z}, b, c, d \in \mathbb{Q}$
- Both models has perfect fit. Which is more plausible?

Prior for the model choice : Assign equal probabilities to two models.

Prior of θ conditional on α

- For M_1, $\alpha, \beta \sim \ iid\ Unif\{-50, -49, ..., 49, 50\}$
- For M_2, $a \sim Unif\{-50, -49, ..., 49, 50\}$ and b, c, d having form of $\frac{X}{Y}$ where $X \sim Unif\{-50, -49, ..., 49, 50\} \perp Y \sim Unif\{0, 1, ..., 49, 50\}$.

Calculate marginal likelihood (= evidence)

1. $p(\mathcal{D}|M_1) = \sum_\alpha \sum_\beta [p(\mathcal{D}|\alpha, \beta, M_1) \cdot p(\alpha, \beta|M_1)]$
   Since $p(\mathcal{D}|\alpha, \beta, M_1) = 1$ if $\alpha = -1, \beta = 4$ and $p(\mathcal{D}|\alpha, \beta, M_1) = 0$ for all other (α, β) combinations,
   $p(\mathcal{D}|M_1) = p(\alpha = -1, \beta = 4|M_1) = \frac{1}{101}\frac{1}{101} \approx 1e - 4$.
2. $p(\mathcal{D}|M_2) = \sum_a \sum_b \sum_c \sum_d [p(\mathcal{D}|a, b, c, d, M_2) \cdot p(a, b, c, d|M_2)]$
   Since $p(\mathcal{D}|a, b, c, d, M_2) = 1$ if $(a, b, c, d) = (-1, -\frac{1}{11}, \frac{9}{11}, \frac{23}{11})$, and $p(\mathcal{D}|a, b, c, d, M_2) = 0$ for all other (a, b, c, d),
   $p(\mathcal{D}|M_1) = p(a = -1, b = -\frac{1}{11}, c = \frac{9}{11}, d = \frac{23}{11}|M_2) = (\frac{1}{101}) \times (4 \cdot \frac{1}{101} \cdot \frac{1}{50}) \times (4 \cdot \frac{1}{101} \cdot \frac{1}{50}) \times (2 \cdot \frac{1}{101} \cdot \frac{1}{50}) \approx 2.5e - 12$.

Posterior model probability: Since $p(\alpha = 0) = p(\alpha = 1) = 0.5$ (equal), overwhelming evidence to choose M_1 over M_2.

Outline

1. Introduction (Motivation)
2. Basics of Bayesian Statistics
3. Parametric and nonparametric model
4. Bayesian linear model
5. Bayesian basis function model
6. Gaussian process regression (GPR)

Parametric and nonparametric Model

Parametric model : model indexed by finite dimensional parameters. model complexity bounded as $n \to \infty$

Nonparametric model : Two different categories of nonparametric models!

1. **No parameter to estimate**, Rank statistics based (median than mean).
2. **model not indexed by finite dimensional parameter : function.** e.g) Gaussian process regression
   Some call parametric, but scalable (dimension can be arbitrarily ↑ as $n \to \infty$) model as nonparametric model.
   e.g) Deep Neural Network!

cf) Semiparametric model : mean function has both parametric & nonparametric components : $E(Y_i|X_i) = \beta_1 x_1 + f(x_2)$

Do not be confused with 1) linear & nonlinear models and 2) parametric & nonparametric models.

Examples

1. Parametric regression
   1. Parametric linear regression : $E(Y_i|X_i) = X_i'\beta$. No such thing as nonparametric linear regression!
   2. Parametric nonlinear regression : $E(Y_i|X_i) = g(x_i;\beta)$. Of course, **functional form** of g is known.
      e.g, logistic growth model : $g(x_i;\beta_1,\beta_2,\beta_3) = \frac{\beta_1}{1+\beta_2 e^{-\beta_3 x_i}}$.
2. Nonparametric regression (surely nonlinear)
   $E(Y_i|X_i) = \mu(x_i)$, where the functional form $m(\cdot)$ is unknown and main goal is to estimate $m(\cdot)$
   1. Classical : **basis function models (# basis not fixed)** e.g) radial basis & spline basis, GAM, **Gaussian Process**
   2. Modern : tree based model, **neural network**, etc.

Outline

1. Introduction (Motivation)
2. Basics of Bayesian Statistics
3. Parametric and nonparametric model
4. Bayesian linear model
5. Bayesian basis function model
6. Gaussian process regression (GPR)

Linear Model

Motivation : Although today's topic is Bayesian nonparametric regression, especially GPR.

GPR is **basis function model** with ∞ basis terms and **basis function model** is an extension of **linear model**

Data: (y, X) where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$. X is assumed to be given (not considering distribution of X).
$y|X, \beta, \sigma^2 \sim MVN_n(X\beta, \sigma^2 I)$, where $\beta \in \mathbb{R}^p, \sigma^2 > 0$

$\rightarrow p(y|X, \beta, \sigma^2) = |2\pi\sigma^2 I|^{-\frac{1}{2}} exp(-\frac{1}{2}(y - X\beta)^T(\sigma^2 I)^{-1}(y - X\beta)) \propto exp(-\frac{1}{2\sigma^2} SSR(\beta))$, $SSR(\beta) := ||y - X\beta||_2^2$

Bayesian linear regression : parametric model with parameters (β, σ^2). Need prior and posterior of β, σ^2.

One example of prior : $p(\sigma^2, \beta) = p(\sigma^2)p(\beta|\sigma^2)$. Inverse gamma prior on σ^2 and g-prior on β.

1. $\sigma^2 \sim InvGamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$
2. $\beta|X, \sigma^2 \sim MVN_p(0, g\sigma^2(X^T X)^{-1})$

* As I said "one example of prior", the prior above is not the only possibility.

Some results of the prior above

- Posterior of the parameters : Since we have two parameters (β, σ^2), need $p(\beta, \sigma^2|y, X) \propto p(\sigma^2|y, X)p(\beta|\sigma^2, y, X)$

  Let $SSR_g :=:= y^T[I - \frac{g}{g+1}X(X^TX)^{-1}X^T]y$

  1. $\sigma^2|y, X \sim InvGamma(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+SSR_g}{2})$
  2. $\beta|y, X, \sigma^2 \sim MVN(\frac{g}{g+1}(X^TX)^{-1}X^Ty, \frac{g}{g+1}\sigma^2(X^TX)^{-1})$

- **Model selection** : Question) Ideas to introduce a parameter that represent different linear models?

  Ans) Use indicator R.vector $z \in \mathbb{R}^p$ s.t. $y_i = \sum_{j=1}^p z_j b_j x_{ij} + \epsilon_i$. $z_j = I(b_j \neq 0)$: whether j^{th} predictor is selected.

  e.g, for $p = 4$, $E(Y|x, b, z = (1, 0, 1, 0)) = b_1x_1 + b_3x_3$ and $E(Y|x, b, z = (1, 1, 0, 0)) = b_1x_1 + b_2x_2$.

  Posterior odds $\frac{p(z_a|y,X)}{p(z_b|y,X)} = \frac{p(z_a)}{p(z_b)} \times \frac{p(y|X,z_a)}{p(y|X,z_b)}$. Without enough information $p(z_a) = p(z_b) = 0.5$.

  → **marginal likelihood** $p(y|X, z)$ critical role!

  $p(y|X, z) = \int_{\sigma^2}\int_{\beta} p(y, \beta, \sigma^2|X, z)d\beta d\sigma^2 = \int_{\sigma^2}\int_{\beta} p(y|X, z, \sigma^2, \beta)p(\beta|X, z, \sigma^2)p(\sigma^2)d\beta d\sigma^2$

Amount of information chosen by hyperparameter g: look at the form of $p(\beta|X, y, \sigma^2)$!

1. $g \to 0$: $p(\beta|X, y, \sigma^2) \approx p(\beta)$
2. $g = 1$: prior has equal information as likelihood
3. $g = n$: unit information prior
4. $g \to \infty$: $p(\beta|X, y, \sigma^2) \approx p(\beta_{MLE})$

Outline

1. Introduction (Motivation)
2. Basics of Bayesian Statistics
3. Parametric and nonparametric model
4. Bayesian linear model
5. Bayesian basis function model
6. Gaussian process regression (GPR)

Basis function model

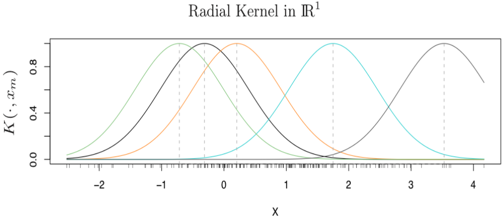
Key concept : express or approximate $\mu(x) := E(Y|x)$ by $\sum_{k=1}^{p} \beta_k b_k(x)$
: approximate the basis function by linear combination of basis expansion of X.

Form a design matrix $W = ((b_k(x_i))_{i,k} \in \mathbb{R}^{n \times p} \rightarrow y = X\beta + \epsilon$: same as linear model setting! (linear in **pararmeters**)

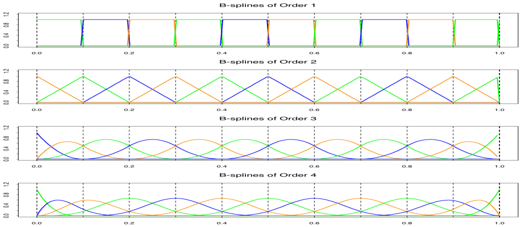
* Why is this a nonparametric regression model? b/c dimension can get ∞ as $n \rightarrow \infty$

Types of basis functions

- Original linear model basis : $X_1, X_2, ...$, polynomial basis : $X_j^2, X_j X_k, ...$
- Splines : piecewise polynomials with continuity (or smoothness of some order) constraints
- Gaussian radial basis function : $b_k(x) = exp(-\frac{(x-x_k)^2}{l^2})$, x_k : center of basis function and l : common width.



(a) Gaussian RBF (from Hastie et al., 2009)



(b) B-spline basis function with different orders (from Hastie et al., 2009)

Splines : piecewise polynomials with continuity (or smoothness of some order) constraints

Cons of global polynomial regression
All data points involved in the estimation of the regression coefficients. Thus, remote parts involved in prediction on a point.

Def) Knots : Letting x ranging from $[a, b]$, let $a < \xi_1 < ... < \xi_K < b$. $\{\xi_1, ..., \xi_K\}$ called as knots.

Def) Order-M splines with fixed knots $\{\xi_1, ..., \xi_K\}$: Piecewise polynomials of order M & cont. derivative up to order M-2 : e.g) $M = 1$: piecewise linear, $M = 2$: continuous piecewise linear,..., $M = 4$: piecewise cubic w/ continuous 1st,2nd derivatives.

Two most famous bases for order-M splines with fixed knots

1. **Truncated power basis** : intuitively great, but computationally bad.
   $\{1, x, ..., x^{M-1}, (x - \xi_1)_+^{M-1}, ..., (x - \xi_K)_+^{M-1}\}$ is a basis set for order-M splines with knots $\{\xi_1, ..., \xi_K\}$.
   $\rightarrow f(x) = \sum_{j=0}^{M-1} \beta_j x^j + \sum_{j=1}^{K} \theta_j (x - \xi_j)_+^{M-1}$
2. **B-spline basis** : intuitively worse, but computationally great

Number of knots and knot positions : Suppose we use order-M splines. Then, the models are decided by 1) number of knots, 2) knot positions.

- Number of knots related to model complexity.
- Given the number of knots, knot positions can be either **uniform** or **free**. If we use uniform knot positions, the knot positions are completely decided by number of knots.

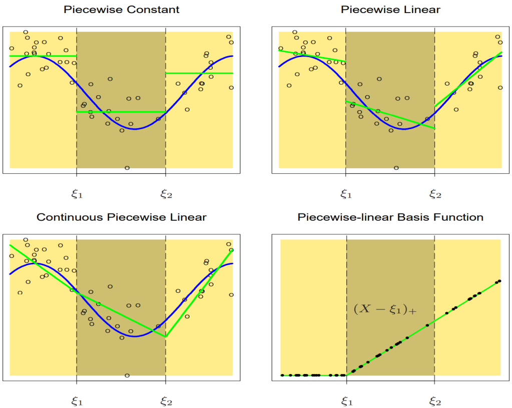


FIGURE 5.1. *The top left panel shows a piecewise constant function fit to some artificial data. The broken vertical lines indicate the positions of the two knots ξ_1 and ξ_2. The blue curve represents the true function, from which the data were generated with Gaussian noise. The remaining two panels show piecewise linear functions fit to the same data—the top right unrestricted, and the lower left restricted to be continuous at the knots. The lower right panel shows a piecewise–linear basis function, $h_3(X) = (X - \xi_1)_+$, continuous at ξ_1. The black points indicate the sample evaluations $h_3(x_i)$, $i = 1, \ldots, N$.*

(a) from Hastie et al., 2009

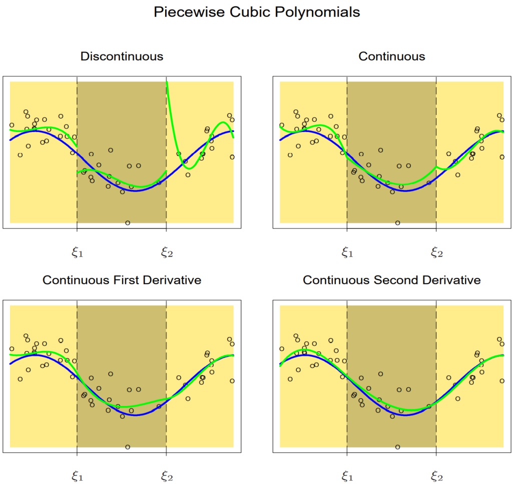
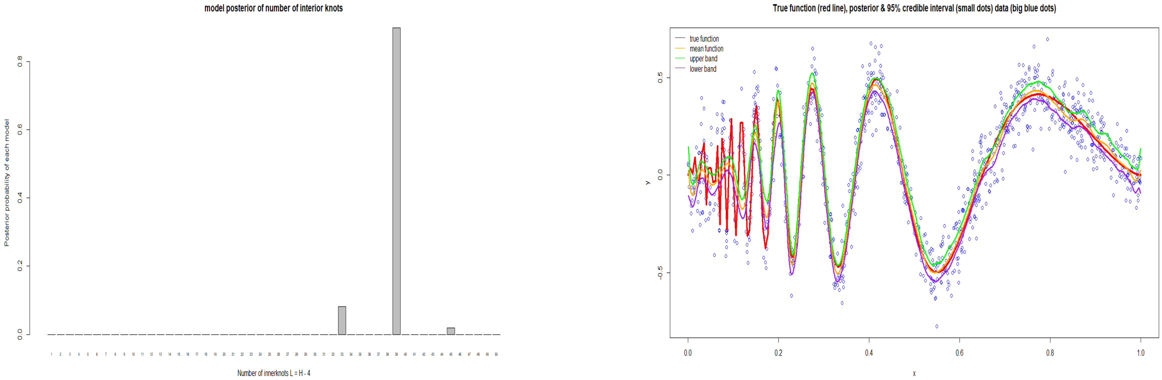


FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

(b) from Hastie et al., 2009

Example: B-spline basis functions fitting for Doppler function

- $f(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right), x \in [0,1]$
- I generated data by $y_i \sim N(f(x_i), \sigma^2)$, where $\sigma^2 = 0.1^2$, $y_1, ..., y_n$ independent. $n = 1000$.
- B-spline basis functions and uniform knots, so, model completely specified by number of knots!
- g-prior with $g = n$ and Jeffrey's prior on $\sigma^2 : p(\sigma^2) \propto \frac{1}{\sigma^2}$



(a) Posterior probabilities of models (number of knots)

(b) Pointwise posterior and 95% confidence band

Outline

1. Introduction (Motivation)
2. Basics of Bayesian Statistics
3. Parametric and nonparametric model
4. Bayesian linear model
5. Bayesian basis function model
6. Gaussian process regression (GPR)

Gaussian Process Regression

Following Gaussian process regression is unarguably nonparametric, with even more flexibility than basis function models.

More flexibility : wider family of functions contained in parameter space
Note, parameter in the nonparametric regression is a function, compared to linear regression, GLM, etc.

Based on the Bayes rule (as always), but the parameter is a function, requiring challenging math: functional analysis.

Model setting and data

- Data $\mathcal{D} = (X, y)$. Since X is deterministic always in the regression setting, only consider distribution on y.
- $y_i = f(x_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \rightarrow$ Likelihood $y_i | f \sim N(f(x_i), \sigma^2)$. Goal is to learn f with uncertainty estimation
- Bayes rule : $p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})} \propto p(f)p(\mathcal{D}|f)$
- GP Prior on f : $f \sim GP(m = 0, K)$, where m is the prior mean function and K is the prior covariance function. Realizations (or, samples) of GP is a random function, which makes GP used as prior distribution over functions.

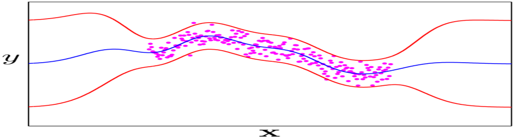


Figure: GPR example, Figure from Zoubin, 2011

Introduction to Stochastic Processes

Basic Terms

1. **Stochastic process**: Family of random variables indexed by index set $\mathcal{I}$ is called a stochastic process.
   : S valued family of random variables $\{X(t)|t \in \mathcal{I}\}$
2. **Index Set**: $\mathcal{I}$ is called the index set, or the parameter set (common to be a set of time).
3. **State Space** "S": The set of different values that the stochastic processes can take.
   - Discrete State Space(Finite or countable) vs Continuous State Space(uncountable)
4. **Sample function = Trajectory = Path function = Path**: single outcome (realization) of a stochastic process.

Types of Stochastic Processes

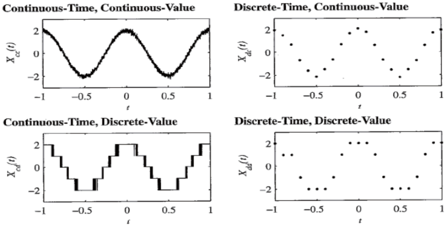


Figure: From https://www.ee.ryerson.ca/ courses/ee8103/chap4.pdf

Gaussian Processes

Def) A **continuous time** stochastic process $\{X_t | t \in \mathcal{I}\}$ is called **Gaussian Process** if **every finite collection** of times $t_1, ..., t_k$ follow **multivariate normal distribution**.

By definition, when $f \sim GP(m, K)$, any n finite points $x_1, ..., x_n \in \mathbb{R}^p$,
$(f(x_1), ..., f(x_n)) \sim MVN_n((m(x_1), ..., m(x_n)), K(x_1, ..., x_n))$, where $K(x_1, ..., x_n)$ is a $n \times n$ matrix with (i, j) element $K(x_i, x_j) = Cov(f(x_i), f(x_j))$.

✓ For mean function, $m(x) = 0$ is usually used.

✓ Infinitely many possiblities of covariance function:

1. Isotropic Squared exponential covariance function : $K(x, x') = \tau^2 exp(-\frac{||x-x'||_2^2}{l^2}))$
2. Anisotropic squared exponential covariance function : $K(x, x') = \tau^2 exp(-\sum_{j=1}^{p} \frac{(x_j - x_j')^2}{l_j^2})$

Possible tasks

1. Model selection : tune hyperparameters σ^2, τ^2, l^2 by comparing **marginal likelihood (= evidence)** $p(y|X) = \int_f p(y|f, X)p(f)df$
2. Obtain posterior of f : $p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})} \propto p(f)p(\mathcal{D}|f)$
3. Obtain posterior predictive $p(y_*|x_*, \mathcal{D}) = \int_f p(y_*|x_*, f, \mathcal{D})p(f|\mathcal{D})df$, where y_*, x_* denote new data.

Posterior of f

Goal : Find $p(f|\mathcal{D})$, which is a distribution over function, which is impossible to directly find.

Thus, find $p((f(\tilde{x}_1), ..., f(\tilde{x}_m))|y, \sigma^2)$: posterior of arbitrary finite object $(f(\tilde{x}_1), ..., f(\tilde{x}_m))$ given required hyperparams.

Step 1) By def'n of GP, $(f(x_1), ..., f(x_n), f(\tilde{x}_1), ..., f(\tilde{x}_m)) \sim MVN_{n+m}(0_{n+m}, \begin{bmatrix} K_{x,x} & K_{x,\tilde{x}} \\ K_{\tilde{x},x} & K_{\tilde{x},\tilde{x}} \end{bmatrix})$, where $K_{x,x} \in \mathbb{R}^{n \times n}$ has i, j element $k(x_i, x_j) = Cov(f(x_i), f(x_j))$, $K_{x,\tilde{x}} \in \mathbb{R}^{n \times m}$ having i, j element $k(x_i, \tilde{x}_j) = Cov(f(x_i), f(\tilde{x}_j))$, $K_{\tilde{x},x} = K(x, \tilde{x})'$, $K_{\tilde{x},\tilde{x}} \in \mathbb{R}^{m \times m}$ has i, j element $k(\tilde{x}_i, \tilde{x}_j) = Cov(f(\tilde{x}_i), f(\tilde{x}_j))$

Step 2) Denoting $\tilde{f} := (f(\tilde{x}_1), ..., f(\tilde{x}_m)) \in \mathbb{R}^m$ $\begin{pmatrix} y \\ \tilde{f} \end{pmatrix} |\sigma^2 \sim MVN_{n+m}(0, \begin{bmatrix} K_{x,x} + \sigma^2 I_n & K_{x,\tilde{x}} \\ K_{\tilde{x},x} & K_{\tilde{x},\tilde{x}} \end{bmatrix}$

Step 3) Applying conditional distribution of MVN, $\tilde{f}|y, \sigma^2 \sim MVN_m(E(\tilde{f}|y, \sigma^2), Cov(\tilde{f}|y, \sigma^2))$, where $E(\tilde{f}|y, \sigma^2) := K_{\tilde{x},x}(K_{\tilde{x},x} + \sigma^2 I_n)^{-1} y$, and $Cov(\tilde{f}|y, \sigma^2) := K_{\tilde{x},\tilde{x}} - K_{\tilde{x},x} + \sigma^2 I_n)^{-1} K_{x,\tilde{x}}$

Posterior predictive of new response y_*

Step 1) $(f(x_1), ..., f(x_n), f(\tilde{x}_1), ..., f(\tilde{x}_m)) \sim MVN_{n+m}(0_{n+m}, \begin{bmatrix} K_{x,x} & K_{x,x_*} \\ K_{x_*,x} & K_{x_*,x_*} \end{bmatrix})$

$(\epsilon_1, ..., \epsilon_n, \epsilon_{1*}, ..., \epsilon_{m*}) \sim MVN_{n+m}(0_{n+m}, \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & \sigma^2 I_m \end{bmatrix})$ by iid noise assumption

Step 2) $\begin{pmatrix} y \\ y_* \end{pmatrix} | X, X_*, \sigma^2 = \begin{pmatrix} f \\ f_* \end{pmatrix} + \begin{pmatrix} \epsilon \\ \epsilon_* \end{pmatrix} \sim MVN_{n+m}(0, \begin{bmatrix} K_{X,X} + \sigma^2 I_n & K_{X,X_*} \\ K_{X_*,X} & K_{X_*,X_*} + \sigma^2 I_m \end{bmatrix})$

: sum of independent normals : normals

Step 3) Applying conditional distribution of MVN, $y_* | y, X, X_*, \sigma^2 \sim MVN_m(\mu_*, \Sigma_*)$, where

$\mu_* = K_{X_*,X}(K_{X,X} + \sigma^2 I_n)^{-1} y$ and $\Sigma_* = K_{X_*,X_*} + \sigma^2 I_m - K_{X_*,X}(K_{X,X} + \sigma^2 I_n)^{-1} K_{X,X_*}$

Useful demo in http://chifeng.scripts.mit.edu/stuff/gp-demo/ and https://bookdown.org/rbg/surrogates/chap5.html

GP in classification

In classification task, GPR can do things that Support Vector Machines (SVM) do and moreover, provide uncertainty quantification (Zoubin, 2011)

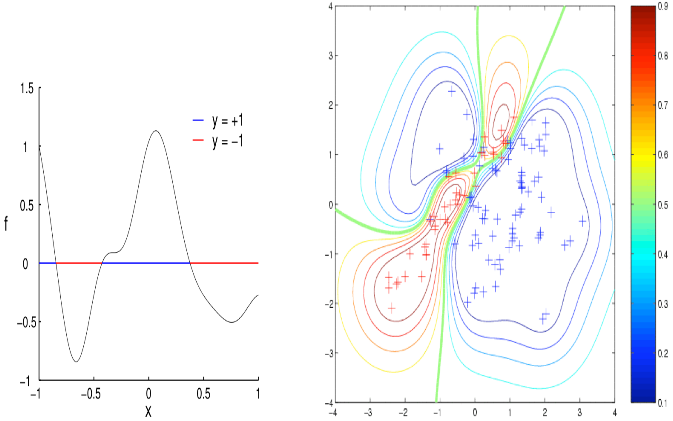


Figure: Left : one dimensional GPR classification, Right : two dimensional GPR classification. Figure from Zoubin, 2011

Connection between linear regression, basis function model, GPR

1. **Linear regression** : parametric regression, linear in paraemeters and linear in X
2. **Basis function model** : nonparametric regression b/c number of basis functions can grow to ∞. However, same framework as linear regression. linear in parameters, nonlinear in X
3. **GPR** : basis function model with truly ∞ number of basis functions. (This part requires functional analysis)
   - $E(Y|X) = \sum_{k=1}^{p} \beta_k b_k(x) = \beta' b(x)$. $\beta := (\beta_1, ..., \beta_p)' \sim MVN_p(\beta_0, \Sigma_0)$
   - $(f(x_1), ..., f(x_n)) \sim MVN((m(x_1), ..., m(x_n)), K(x_1, ..., x_n))$, $(m(x_1), ..., m(x_n)) = \beta_0' b(x)$ and $K(x_i, x_j) = b(x_i)' \Sigma_0 b(x_j)$
4. **BNN and GPR**
   - **BNN** : neural network (either deep or shallow) with priors on weights. Many times, **overparameterized model**.
   - **GPR** : truly infinite # of parameters
   - **Relationships** :
     1. One layer NN and infinite width and gaussian prior on weights equivalent to **GPR**
     2. A NN with arbitrary depth and nonlinearities, with dropout within every hidden layer equivalent to **deep GPR**
   - **pros and cons** :
     1. Although GPR is truly overparameterized, less "hyperparameters" to handle, so, need less data than BNN
     2. Runtime bad in GPR ($O(n^3)$ by matrix inversion), better time complexity in BNN.

References

https://www.youtube.com/watch?v=naN41kICcEQlist=PLUAbWHMZe0KFJ5WYnN1O5rg4TAxz9UphXindex=3

Gelman, A., Carlin, Jd.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20110553.

Z. Ghahramani, A Tutorial on Gaussian Processes (or why I don't use SVMs), in: Machine Learning Summer School (MLSS), 2011.

https://www.youtube.com/watch?v=IEpc2ClaYH8list=PLUAbWHMZe0KFJ5WYnN1O5rg4TAxz9UphXindex=6

Advanced Bayesian methods (2022, Yonsei university) lecture note

https://www.youtube.com/watch?v=Z9cdIQ-WDLMlist=PLUAbWHMZe0KFJ5WYnN1O5rg4TAxz9UphXindex=7

http://chifeng.scripts.mit.edu/stuff/gp-demo/

https://bookdown.org/rbg/surrogates/chap5.html

https://towardsdatascience.com/deep-neural-networks-vs-gaussian-processes-similarities-differences-and-trade-offs-18647376d799