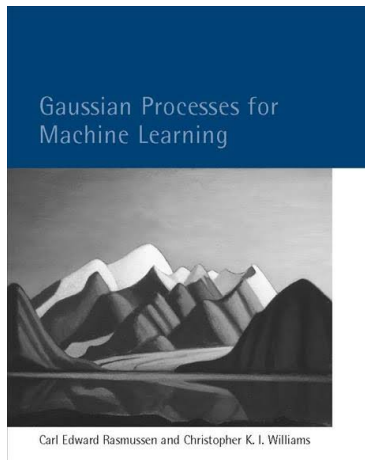


Gaussian Process

Samrudhdhi Rangrej

ECSE 626, Winter 2019



Book is available online. Today we will look at the 2nd Chapter.

- $y = f(x)$ is a deterministic function if each x corresponds to a deterministic set of y .
- $y = f(x)$ is a random function/random process if each x corresponds to a distribution over y .
- Examples:
 - Temperature of a room as a function of time is random process. Everyday a new function emerges from a distribution.
 - Human weight of as a function of age. For every person a new function emerges from a distribution.

- A **random process** is:
 - a collection of *indexed* random variables. (x is discrete variable.)
 - a time varying function. (x is continuous variable.)
- A **Gaussian process** is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Gaussian Process Regression(GPR)

- **Regression:** learn a deterministic function $g(x) = y$ given the training data pairs (x, y) .
- **Gaussian Process Regression:** learn a Gaussian process $g(x) = y$ given the training data pairs (x, y) . i.e. learn mean and variance of GP.

Two views of GPR

Weight-space view	Function-space view
Assumes parametric form of $g(\cdot)$.	No such assumption is made.
Two stages: (1) Bayesian inference of parameters of $g(\cdot)$ (2) Computation of $g(x)$ for input x .	Directly infer $g(x)$.
Similarity with Ridge Regression.	Similarity with Nearest Neighbor classifier.

Both methods give same answer.

Weight-Space View

Data:

$$\mathcal{D} = (X, \mathbf{y}); \quad X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T; \quad \mathbf{y} = [y_1, y_2, \dots, y_n]^T$$

Lets look at the case where data follows linear model and is corrupted by additive Gaussian noise.

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{w}; \quad y = f(\mathbf{x}) + \epsilon; \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

(Note: We will consider $f(\mathbf{x})$ to be a Gaussian Process and not y . Noise is considered Gaussian only for mathematical ease. Other noise models can also be used.)

*Notation: Capital letters are for matrices, bold-face small letters are for vectors, italic small letters are for scalars.

Assuming that data is i.i.d. let's derive likelihood of the observations given \mathbf{w} .

$$\begin{aligned} p(\mathbf{y}|X, \mathbf{w}) &= \prod_{i=1}^n p(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= \mathcal{N}(X^T \mathbf{w}, \sigma_n^2 I) \end{aligned}$$

Bayesian view advises us to set a prior over parameters \mathbf{w} . Before looking at the data, we believe that parameters of the model are drawn from unit Gaussian distribution.

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

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|\mathbf{w}, X)p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{w}, X)p(\mathbf{w})d\mathbf{w}}$$

Hence,

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, X) &\propto \mathcal{N}(X^T \mathbf{w}, \sigma_n^2 I) \mathcal{N}(0, \Sigma_p)^\dagger \\ &= \mathcal{N}\left(\frac{1}{\sigma_n^2} A^{-1} X \mathbf{y}, A^{-1}\right) \end{aligned}$$

Where,

$$A = \sigma_n^{-2} X X^T + \Sigma_p^{-1}$$

[†]Useful property of Gaussian: $\mathcal{N}(\mathbf{x}|\mathbf{a}, A)\mathcal{N}(\mathbf{x}|\mathbf{b}, B) = Z^{-1}\mathcal{N}(\mathbf{x}|\mathbf{c}, C)$;
 $\mathbf{c} = C(A^{-1}\mathbf{a} + B^{-1}\mathbf{b})$; $C = (A^{-1} + B^{-1})^{-1}$;
 $Z^{-1} = (2\pi)^{-\frac{D}{2}} |A + B|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{a} - \mathbf{b})^T (A + B)^{-1}(\mathbf{a} - \mathbf{b}))$.

- What does posterior variance $A^{-1} = (\sigma_n^{-2}XX^T + \Sigma_p^{-1})^{-1}$ say?
 - Posterior is thinner and sharper than the prior. i.e. prior uncertainty is resolved after looking at the data.
- As the mean of a Gaussian also serves as its mode, mean of the posterior is MAP solution.
 - Maximum a posteriori(MAP) estimate of some quantity is equal to the mode of its posterior.[‡]
- Will the mean of the posterior still be a MAP solution if the data is assumed to be corrupted by non-Gaussian (additive) noise?

[‡]What maximum likelihood is to frequentist is MAP to Bayesian. Sometimes frequentists disguise prior as regularization term and describe MAP as penalized maximum likelihood i.e. Ridge Regression.

Lets derive distribution of prediction(f_*) for a test sample \mathbf{x}_* .

$$\begin{aligned} p(f_*|\mathbf{x}_*, X, \mathbf{y}) &= \int p(f_*|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|X, \mathbf{y})d\mathbf{w} \\ &= \mathcal{N}\left(\frac{1}{\sigma_n^2}\mathbf{x}_*^T A^{-1}X\mathbf{y}, \mathbf{x}_*^T A^{-1}\mathbf{x}_*\right) \end{aligned}$$

- Mean(\bar{f}_*) of the prediction samples $f_* \sim p(f_*|\mathbf{x}_*, X, \mathbf{y})$ is equal to a single prediction at mean of the posterior $p(\mathbf{w}|X, \mathbf{y})$, i.e.
 $\bar{f}_* = \mathbf{x}_*^T \bar{\mathbf{w}} = \frac{1}{\sigma_n^2}\mathbf{x}_*^T A^{-1}X\mathbf{y}$. Which assumption induces this behavior?

Visualization

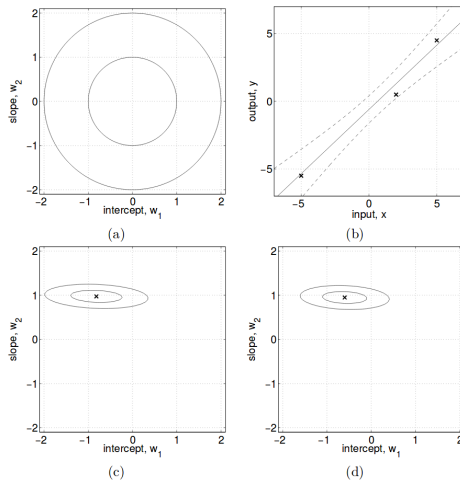


Figure: Visualization of Bayesian linear model.[1] (a) prior (b) fitted linear model with three training points. Notice how uncertainty grows with magnitude of x . (c) likelihood (d) posterior.

Projection of Data into Higher Dimensional Feature Space

- Problem: Data may not always be explained by a linear model.
- Solution: Project data into higher dimensional feature space and then apply linear model.

$$\mathbf{x} \rightarrow \phi(\mathbf{x})$$

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}$$

$$A = \sigma_n^{-2} \Phi \Phi^T + \Sigma_p^{-1}$$

$$p(f_* | \mathbf{x}_*, X, \mathbf{y}) = \mathcal{N}\left(\frac{1}{\sigma_n^2} \phi(\mathbf{x}_*)^T A^{-1} \Phi \mathbf{y}, \phi(\mathbf{x}_*)^T A^{-1} \phi(\mathbf{x}_*)\right)$$

Kernel or Covariance Function

When you replace formula of A is the last formula, you will notice the terms $\phi_*^T \Sigma_p \Phi$ and $\Phi^T \Sigma_p \Phi$. This form is called *Kernel* or *Covariance function*. Kernels facilitates easy and fast computation. How?

$$\begin{aligned}k(x, x') &= \phi^T \Sigma_p \phi' \\&= \phi^T (\Sigma_p^{1/2})^T (\Sigma_p^{1/2}) \phi' \\&= (\Sigma_p^{1/2} \phi)^T (\Sigma_p^{1/2} \phi') \\&= \psi^T \psi' \\&= \psi \cdot \psi'\end{aligned}$$

$k(x, x')$ can be any function with two arguments as far as it can be written is the form $\psi \cdot \psi'$.

- Radial Basis Functions (RBF) are widely used Kernels.

$$k(x, x') = \exp\left(-\frac{|x - x'|^2}{2\ell^2}\right)$$

- ℓ is a length-scale parameter. It indicates how far points x and x' can be in order to affect each other significantly.
- Can you derive $k(x, x') = \psi \cdot \psi'$ form for RBF?

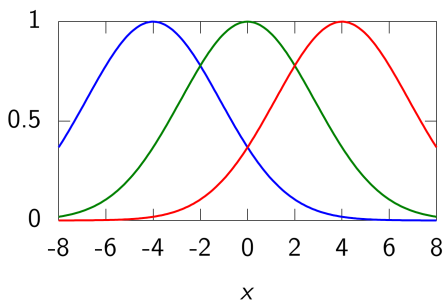


Figure: three RBF Kernels with $\ell = 2$ for: (red) $x' = 4$ (green) $x' = 0$ (blue) $x' = -4$.

Function-Space View

- Recall our model.

$$y = f(\mathbf{x}) + \epsilon; \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$



- Gaussian Process is a Gaussian distribution over functions.

$$p(f|\mathbf{x}) = \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

- Now we have to find:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]^\S$$

[§] Given $f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}$ and $m(\mathbf{x}) = 0$, can you prove that $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \Sigma_p \phi(\mathbf{x}')$?  

- Marginal and Conditional Distribution of Multivariate Gaussian are given below.

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right)$$

$$\mathbf{y} \sim \mathcal{N}(\mu_y, B)$$

$$\mathbf{x} \sim \mathcal{N}(\mu_x, A)$$

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mu_y + CA^{-1}(\mathbf{x} - \mu_x), B - CA^{-1}C^T)$$

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mu_x + CB^{-1}(\mathbf{y} - \mu_y), A - CB^{-1}C^T)$$

Prior and Joint Probability

Our prior belief is that the observations are Gaussian distributed whether they belong to train or test set.

$$\mathbf{f} \sim \mathcal{N}(0, K(X, X)); \quad \mathbf{f}_* \sim \mathcal{N}(0, K(X_*, X_*))$$

Lets write this in terms of joint distribution.

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

or

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

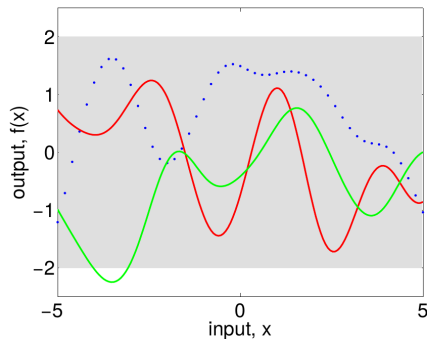
Posterior can be derived from Joint probability using the property of multivariate Gaussian.

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

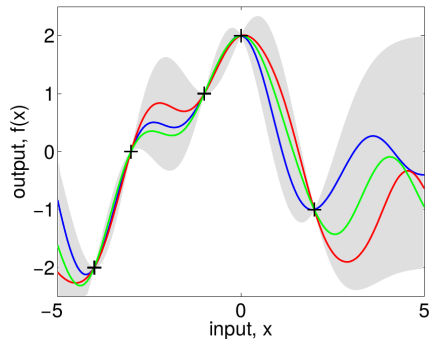
$$\bar{\mathbf{f}}_* = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y}$$

$$\text{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

Visualization



(a), prior



(b), posterior

Figure: Prior and Posterior for **noise-free** case. Five training points are indicated with '+' sign in panel (b). Gray region is scaled according to uncertainty. Notice how uncertainty is reduced near the training points. Figure from [1].

Visualization

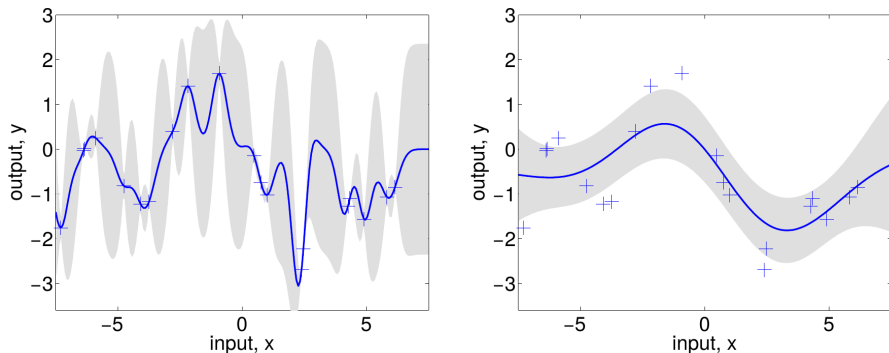


Figure: Posterior for different values of (ℓ, σ_n) **(left)** $(\ell, \sigma_n) = (0.3, 0.00005)$ **(right)** $(\ell, \sigma_n) = (3.0, 0.89)$. Increasing ℓ (length-scale of RBF kernel) increases the extent by which one training sample affects its neighboring region^{||}. σ_n defines the uncertainty around a training sample. Figure from [1].

^{||} Does this reminds you of some form of Nearest Neighbour classifier?



Rasmussen, Carl Edward. "Gaussian processes in machine learning." Advanced lectures on machine learning. Springer, Berlin, Heidelberg, 2004. 63-71.