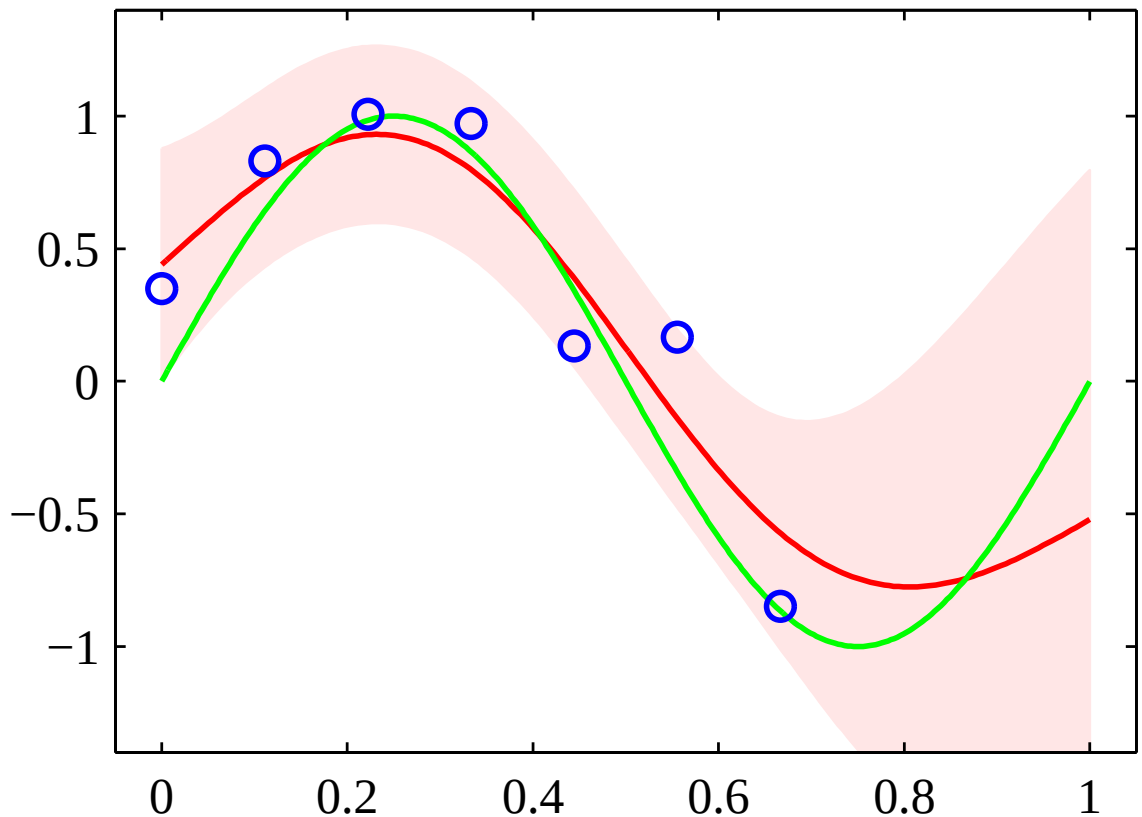
# Gaussian Process Regression

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

# Pictorial Illustration of GP Regression

Green curve: the true sinusoidal function from which the data points are obtained by sampling and addition of Gaussian noise.
Red line: the mean of GP predictive distribution and the shaded region corresponds to plus and minus two standard deviations.



# Gaussian Processes

Definition: A Gaussian Process (GP) is a collection of random variables, any finite number of which has a joint Gaussian distribution.

- A Gaussian process is a generalization of a multivariate Gaussian distribution to infinitely many variables.
- GP defines a a distribution over functions of the form $f : \mathcal{X} \mapsto \mathbb{R}$, which is completely specified by mean function $\mu(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$:

$$f(\boldsymbol{x}) \sim \mathcal{GP}\big(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big),$$

where

$$\begin{aligned} \mu(\boldsymbol{x}) &= \mathbb{E}\left\{f(\boldsymbol{x})\right\}, \\ k(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}\left\{\left(f(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)\left(f(\boldsymbol{x}') - \mu(\boldsymbol{x}')\right)\right\} \\ &= I_f \exp\left\{-\frac{1}{2\rho}(\boldsymbol{x} - \boldsymbol{x}')^\top \boldsymbol{L}(\boldsymbol{x} - \boldsymbol{x}')\right\}, \end{aligned}$$

which is referred to as squared exponential kernel, $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{l})$, $[\boldsymbol{l}]_i$ is a hyperparameter to determine a relevance of the $i$th input dimension.

# Regression

- The regression model is described by

$$y_t = f(\boldsymbol{x}_t) + \boldsymbol{\epsilon}_t.$$

- Data: $\mathcal{D} = \{(\boldsymbol{X}, \boldsymbol{y})\}$

$$\begin{aligned} \boldsymbol{X} &= [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}, \\ \boldsymbol{y} &= [y_1, \ldots, y_N]^\top \in \mathbb{R}^N. \end{aligned}$$

- Goal: Estimate the regression function $f(\cdot)$ to make prediction of $y_* = f(\boldsymbol{x}_*)$.
  - Parametric: Maximum likelihood, Bayesian inference (parametric model $f_w(\boldsymbol{x})$).
  - Nonparametric: Kernel regression (Nadaraya-Watson estimator), GP regression (let the data explain, $f(\boldsymbol{x})$).

# Maximum Likelihood (Parametric Model)

- Model:

$$y_t = f_w(\boldsymbol{x}_t) + \epsilon_t.$$

- Gaussian likelihood:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \propto \prod_{t=1}^{N} \exp\left\{-\frac{1}{2\sigma^2}\left(y_t - f_w(\boldsymbol{x}_t)\right)^2\right\}.$$

- Maximize the log-likelihood:

$$\boldsymbol{w}_{ML} = \arg\max_{w} \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}).$$

- Make predictions by plugging in the ML estimate:

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{ML}).$$

# Bayesian Inference (Parametric Model)

- Model:

$$y_t = f_w(\boldsymbol{x}_t) + \epsilon_t.$$

- Gaussian likelihood:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \propto \prod_{t=1}^{N} \exp\left\{-\frac{1}{2\sigma^2}\left(y_t - f_w(\boldsymbol{x}_t)\right)^2\right\}.$$

- Prior over parameters:

$$p(\boldsymbol{w}|\lambda) \quad (\text{hyperparameters } \lambda).$$

- Posterior over parameters:

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \lambda) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w}|\lambda)}{p(\boldsymbol{y}|\boldsymbol{X})}.$$

# Bayesian Inference (Parametric Model)

- Make predictions by computing predictive distribution:

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}, \lambda) = \int p(y_*|\boldsymbol{w}, \boldsymbol{x}_*) p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \lambda) d\boldsymbol{w}.$$

- Marginal likelihood:

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w}|\lambda) d\boldsymbol{w}.$$

- Estimate hyperparameters by maximizing the log marginal likelihood:

$$\widehat{\lambda} = \arg\max_{\lambda} \log p(\boldsymbol{y}|\boldsymbol{X}).$$

# Kernel Regression

- Model:

$$y_t = f(\boldsymbol{x}_t) + \epsilon_t.$$

- Nadayara-Watson estimator:

$$\boxed{f(\boldsymbol{x}_*) = \frac{\sum_{t=1}^{N} y_t \, k(\boldsymbol{x}_*, \boldsymbol{x}_t, \lambda)}{\sum_{l=1}^{N} k(\boldsymbol{x}_*, \boldsymbol{x}_l, \lambda)},}$$

  where

$$k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda) = \frac{1}{Z} \exp\left\{-\lambda \|\boldsymbol{x} - \boldsymbol{x}_t\|^2\right\}.$$

- Computes the weighted average of $y_t$'s near $\boldsymbol{x}_*$.

# Nadayara-Watson Estimator (Detailed Derivation)

Recall

$$\begin{aligned} f(\boldsymbol{x}) &= \mathbb{E}\{y|\boldsymbol{x}\} \\ &= \int y\, p(y|\boldsymbol{x})\, dy \\ &= \int y\, \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})}\, dy. \end{aligned}$$

Use kernel density estimation to determine both $p(\boldsymbol{x}, y)$ and $p(\boldsymbol{x})$:

$$\begin{aligned} \widehat{p}(\boldsymbol{x}, y) &= \frac{1}{N} \sum_{t=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda_x) k(y, y_t, \lambda_y), \\ \widehat{p}(\boldsymbol{x}) &= \frac{1}{N} \sum_{t=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda_x). \end{aligned}$$

Compute

$$\begin{aligned}
\int y\,\widehat{p}(\boldsymbol{x}, y)\,dy &= \frac{1}{N}\sum_{t=1}^{N}\int k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda_x)\,y\,k(y, y_t, \lambda_y)\,dy \\
&= \frac{1}{N}\sum_{t=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda_x)\underbrace{\int y\,\frac{1}{Z_y}\exp\left\{-\lambda_y(y-y_t)^2\right\}\,dy}_{y_t} \\
&= \frac{1}{N}\sum_{t=1}^{N} y_t\,k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda_x)
\end{aligned}$$

Therefore,

$$\begin{aligned}
f(\boldsymbol{x}) &= \frac{\int y\,\widehat{p}(\boldsymbol{x}, y)\,dy}{\widehat{p}(\boldsymbol{x})} \\
&= \frac{\sum_{t=1}^{N} y_t\,k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda_x)}{\sum_{l=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_l, \lambda_x)}.
\end{aligned}$$

# Gaussian Process Regression

- Model:

$$y_t = f(\boldsymbol{x}_t) + \epsilon_t,$$

  where $f(\cdot)$ is referred to as latent function.

- Latent vector:

$$\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)]^\top \in \mathbb{R}^N.$$

  Note that "parameters" are function itself in GPR model.

- Gaussian likelihood:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2 \boldsymbol{I}).$$

- Gaussian process prior (zero mean):

$$\begin{aligned} f(\boldsymbol{x}) &\sim \mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}')), \\ p(\boldsymbol{f}|\boldsymbol{X}) &= \mathcal{N}(\boldsymbol{f}|0, \boldsymbol{K}). \end{aligned}$$

# GP Regression (Cont'd)

- Gaussian process posterior:

$$f(\boldsymbol{x})|\boldsymbol{X},\boldsymbol{y} \sim \mathcal{GP}(\overline{\mu}(\boldsymbol{x}), \overline{k}(\boldsymbol{x},\boldsymbol{x}')),$$

where

$$\begin{aligned}
\overline{\mu}(\boldsymbol{x}) &= k(\boldsymbol{x},\boldsymbol{X})\left[k(\boldsymbol{X},\boldsymbol{X}) + \sigma^2\boldsymbol{I}\right]^{-1}\boldsymbol{y}, \\
\overline{k}(\boldsymbol{x},\boldsymbol{x}') &= k(\boldsymbol{x},\boldsymbol{x}') - k(\boldsymbol{x},\boldsymbol{X})\left[k(\boldsymbol{X},\boldsymbol{X}) + \sigma^2\boldsymbol{I}\right]^{-1}k(\boldsymbol{X},\boldsymbol{x}).
\end{aligned}$$

- Predictive distribution:

$$\begin{aligned}
p(y_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x}_*) = \mathcal{N}\Bigg(y_*\Big|&k(\boldsymbol{x}_*,\boldsymbol{X})\left[k(\boldsymbol{X},\boldsymbol{X}) + \sigma^2\boldsymbol{I}\right]^{-1}\boldsymbol{y}, \\
&k(\boldsymbol{x}_*,\boldsymbol{x}_*) - k(\boldsymbol{x}_*,\boldsymbol{X})\left[k(\boldsymbol{X},\boldsymbol{X}) + \sigma^2\boldsymbol{I}\right]^{-1}k(\boldsymbol{X},\boldsymbol{x}_*) + \sigma^2\Bigg),
\end{aligned}$$

where

$$\begin{aligned}
k(\boldsymbol{x}_*,\boldsymbol{X}) &= [k(\boldsymbol{x}_*,\boldsymbol{x}_1),\ldots,k(\boldsymbol{x}_*,\boldsymbol{x}_N)] \in \mathbb{R}^{1\times N}, \\
k(\boldsymbol{X},\boldsymbol{X}) &= [k(\boldsymbol{x}_i,\boldsymbol{x}_j)] \in \mathbb{R}^{N\times N}.
\end{aligned}$$

# GP Regression: Detailed Derivation

Let $\boldsymbol{f}_* \in \mathbb{R}^T$ be latent function values evaluated at test data points $\boldsymbol{X}_* \in \mathbb{R}^{D \times T}$.

We first write the joint distribution of the observed target values and the function values at the test locations under the prior:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I} & k(\boldsymbol{X}, \boldsymbol{X}_*) \\ k(\boldsymbol{X}_*, \boldsymbol{X}) & k(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix} \right),$$

It follows from the Gaussian Identity that we have

$$\begin{aligned} \boldsymbol{f}_* | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_* \quad \sim \quad & \mathcal{N}\Big( \boldsymbol{f}_* \Big| k(\boldsymbol{X}_*, \boldsymbol{X}) \left[ k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}_N \right]^{-1} \boldsymbol{y}, \\ & k(\boldsymbol{X}_*, \boldsymbol{X}_*) - k(\boldsymbol{X}_*, \boldsymbol{X}) \left[ k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}_N \right]^{-1} k(\boldsymbol{X}, \boldsymbol{X}_*) \Big), \end{aligned}$$

leading to

$$\begin{aligned} p(\boldsymbol{y}_* | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*) \quad = \quad & \mathcal{N}\Big( \boldsymbol{f}_* \Big| k(\boldsymbol{X}_*, \boldsymbol{X}) \left[ k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}_N \right]^{-1} \boldsymbol{y}, \\ & k(\boldsymbol{X}_*, \boldsymbol{X}_*) - k(\boldsymbol{X}_*, \boldsymbol{X}) \left[ k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}_N \right]^{-1} k(\boldsymbol{X}, \boldsymbol{X}_*) + \sigma^2 \boldsymbol{I}_T \Big). \end{aligned}$$

# References


- C. E. Rasmussen and C. K. I. Williams, "Gaussian Processes for Machine Learning," MIT Press, 2006.
- C. E. Rasmussen, "Advances in Gaussian Processes," NIPS-2006 Tutorial.