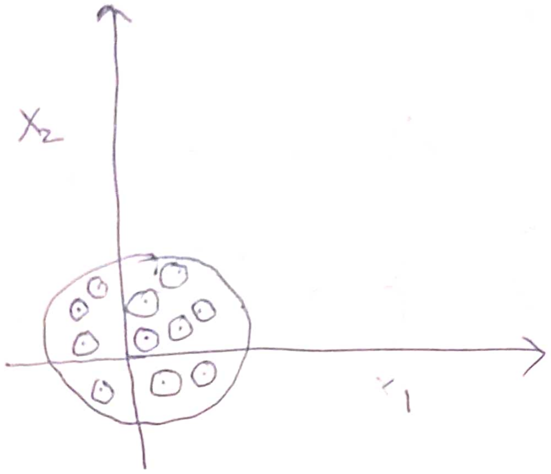
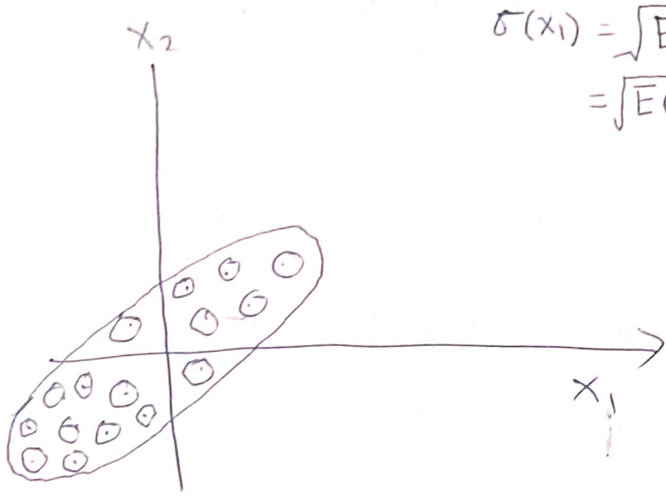
Gaussian Basics

$$\rho_{X_1, X_2} = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma(X_1)\,\sigma(X_2)}$$

$$\sigma(X_1) = \sqrt{E(X_1 - \mu_1)^2}$$
$$= \sqrt{E(X^2)} \text{ if } \mu = 0$$





$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

dot product is a measure of similarity

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$E(X_1 X_2)$

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

$P(X_2 | X_1 = x)$ can be obtained by cutting the jointy density function.

$P(X_2 | X_1 = x) = P(X_2 | X_1)$ is also known as Conditional distribution.

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

Conditional Mean and Variance of $p(X_1 | X_2)$

$$\left.\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}\right\} p(X_1|X_2)$$

Theorem

Suppose (X_1, X_2) is a jointly Gaussian distribution with parameters

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$$

$$\Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12} & \Lambda_{22} \end{bmatrix}$$

The marginal distribution can be expressed as

$$p(X_1) = N(\mu_1, \Sigma_{11})$$
$$p(X_2) = N(\mu_2, \Sigma_{22})$$

The conditional distribution can be expressed as

$$p(X_1 | X_2) = N(\mu_{1|2}\ \Sigma_{1|2})$$ where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Random Sample from multivariate Gaussian

If $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$

then draw $X_1 \sim N(0,1)$

& $X_2 \sim N(0,1)$

the $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ jointly follow $N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$

If $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}\right)$

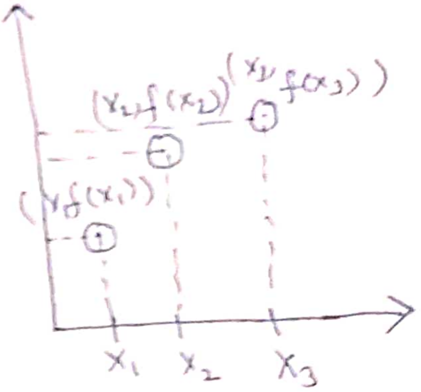
then $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + L X^*$

where $X^* \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$

& $LL^T = \Sigma$

The matrix L can be obtained by Cholesky dicomposition.

Gaussian Process



Let
$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{bmatrix} \right)$$

Let K_{ij} can be given by a measure of similarity

$$K_{ij} = e^{-\|x_i - x_j\|^2}$$

$$= \begin{cases} 0 & \text{if } \|x_i - x_j\| \to \infty \\ 1 & \text{if } x_i = x_j \end{cases}$$

Given Data $D = \{(x_1, f_1), (x_2, f_2), (x_3, f_3)\}$

and x^*

$f^* = ?$

$$f \sim N(0, K)$$

$$f^* \sim N(0, K(x_* x_*))$$

$$K(x_* x_*) = e^{-\|x_* - x_*\|^2} = 1$$

If you increase x^*, f^* should increase but, we assume that x^* is uncorrelated. So, to implement correlation, we can write

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ & K_{21} & K_{23} \\ & K_{23} & K_{33} \end{bmatrix} & \begin{matrix} K_{1*} \\ K_{2*} \\ K_{3*} \end{matrix} \\ \begin{matrix} K_{*1} & K_{*2} & K_{*3} \end{matrix} & K_{**} \end{bmatrix}\right)$$

$$K_* = \begin{bmatrix} K_{1*} \\ K_{2*} \\ K_{3*} \end{bmatrix}$$

$$E(f^*) = K_*^T K^{-1} f = \mu^*$$

$$\sigma^* = K_*^T K^{-1} K_* + K_{**}$$

So, we can estimate the mean and confidence interval.

Gaussian Process: A distribution over functions

A Gaussian process is a Gaussian distribution over function

$$f(x) \sim GP\left(m(x), K(x\ x')\right)$$

$$m(x) = E\left[f(x)\right]'$$

$$K(x\ x') = E\left[\left(f(x) - m(x)\right)\left(f(x') - m(x')\right)^T\right]$$

Usually $\quad K(x, x') = \exp\left(-\frac{1}{2}(x - x')^2\right)$

Simulation from Gaussian Process

(I) Create $X_{1:N}$

(II) Create $\mu = 0_N$ & $K_{N\times N} = \left(\left(\exp - \frac{1}{2}\|x_i - x_j\|^2\right)\right)_{n\times n}$

(III) Decompose $K = LL'$ by Cholesky Method

(III) Generate $f^*_N \sim N(0_N, I_{n\times n})$

(IV) $f = L \cdot f^*_N$

Gaussian Process Prior and Posterior

Let $D = \{(x_i, f_i), i = 1, 2, \dots, N\}$

$$P(f|D) = \frac{P(D|f)\,P(f)}{P(D)}$$

$$P(f) = N(\mu(x), K(x, x'))$$

Noiseless Gaussian Process Regression

We observe a training set $D = \{(x_i, f_i), i = 1, \dots n\}$ where $f_i = f(x_i)$. Given a test set X_* of size $N_* \times D$, we want to predict the outputs f_*

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix} \right)$$

$$K_* = k(X, X^*)$$

$$K_{**} = K(X_*, X_*)$$

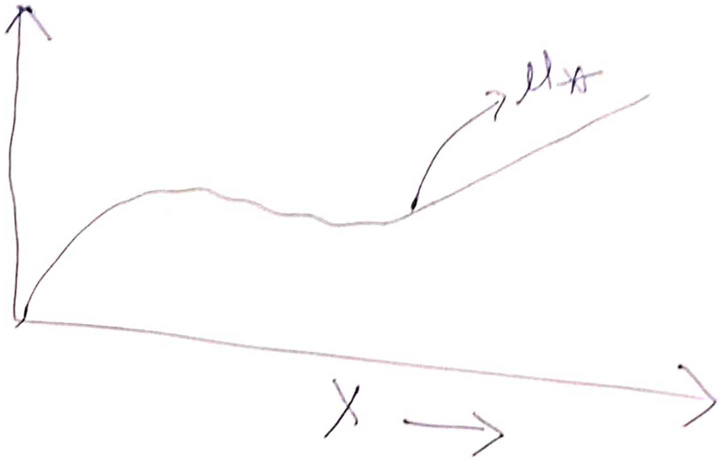
$$K(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x - x')^2\right)$$

Noiseless Gaussian Process Regression can be estimated as

$$p(f_*|X_*, X, f) = N(\mu_*, \Sigma_*)$$

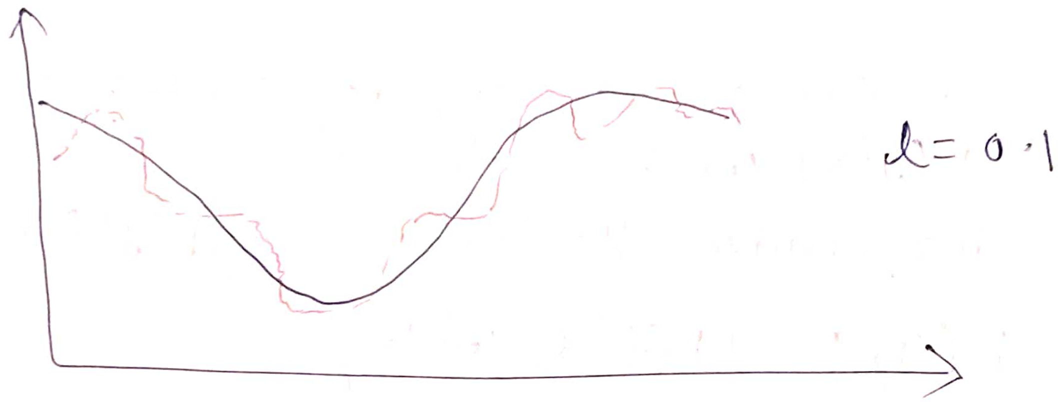
$$\mu_* = \mu(X_*) + K_*^T K^{-1}(f - \mu(X))$$

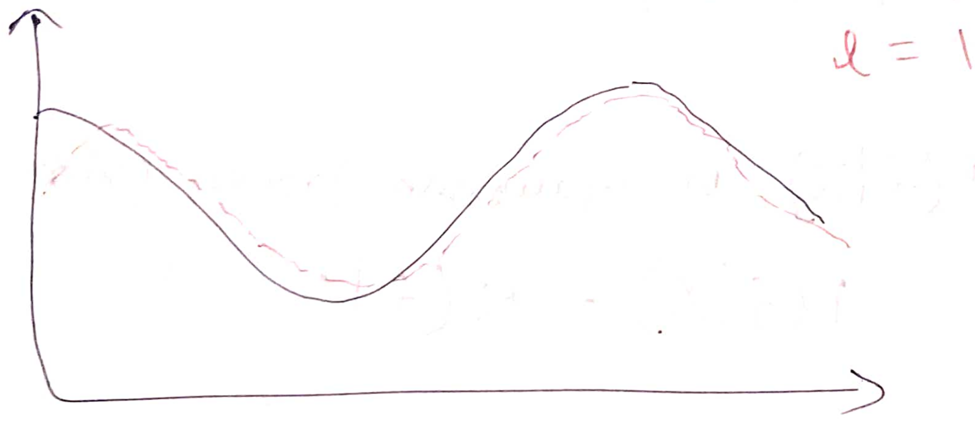
$$\Sigma_* = K_{**} - K_*^T K^{-1} K_*$$



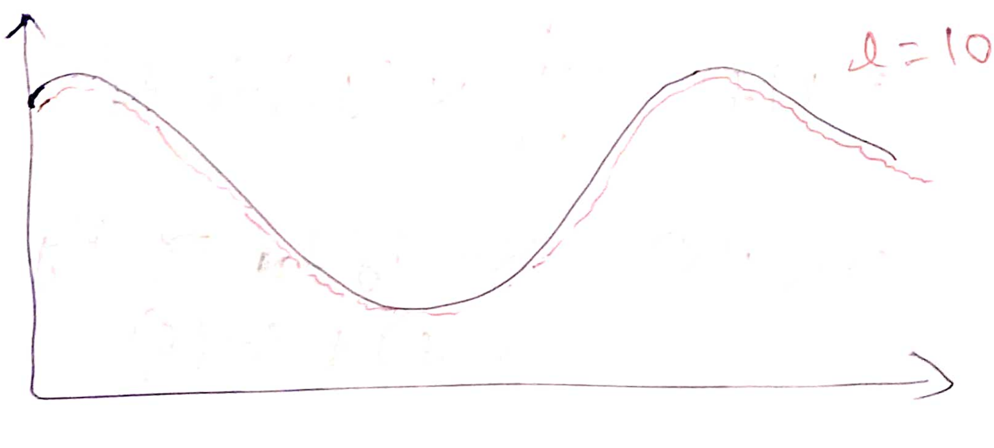
Effect of Choosing Kernal width parameter

If l^2 is very small, means Kernal is very thin only points that are nearby similar to each other and we can expect more wiggly function.





As we increase the Kernal width the function will become smooth.



One way of choose Kernal width is cross-validation. The blue curve is true function.

Noisy Gaussian Process Regression

$$y = f(x) + \epsilon$$

where $\epsilon \sim N(0, \sigma_y^2)$

y is called noisy function measurement

In order to compute $p(y|X)$ we need to marginalize the joint $p(y|f, X)\, p(f|X)$ with respect to f.

$$p(y|X) = \int p(y|f, X)\, p(f|X)\, d$$

$$= \int p(y|f, X)\, p(f|X)\, df$$

$p(f|X)$ is Gaussian Process prior

$$p(f|X) = N(f|0, K)$$

Similarity Kernal

$$p(y|f) = \prod_{i=1}^{n} N(y_i | f_i\ \sigma_y^2)$$

$$\mathrm{cov}(y|X) = K + \sigma_y^2 I_N \approx K_y$$

$$= \mathrm{cov}(f) + \mathrm{cov}(\epsilon)$$

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K_y & K_* \\ K_*^T & K_{**} \end{bmatrix}\right)$$

$$p(f^* \mid X_*, X, y) = N(f_* \mid \mu_*, \Sigma_*)$$

$$\mu_* = K_*^T K_y^{-1} y$$

$$\Sigma_* = K_{**} - K_*^T K_y^{-1} K_*$$

If you want to fit Gaussian process to data you just construct a matrix K using the similarity kernal (If you have noise, you just need to add σ) and obtain μ_* and Σ_*.

This method is sometimes called ML-II method.