Gaussian Processes

Properties of the Multivariate Gaussian

$X \sim N(\mu, \Sigma)$

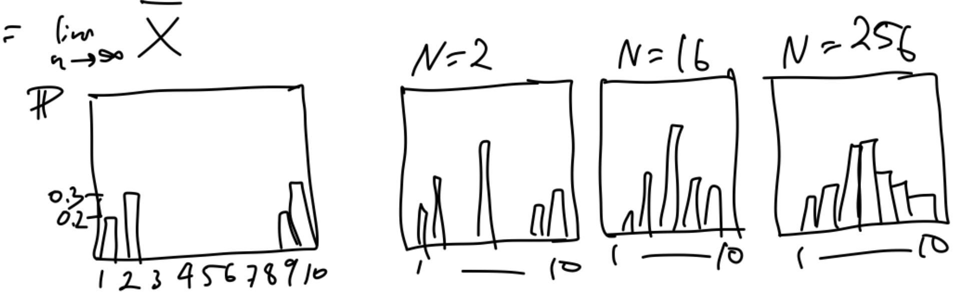
$$P(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|} e^{-\frac{1}{2}((x-\mu)^T \Sigma^{-1} (x-\mu))}$$

determinant

Central Limit Theorem

$X_1, \dots, X_n \quad \mu, \sigma^2$

$N = \lim_{n \to \infty} \overline{X}$



Once Gaussian always Gaussian

Gaussian

$y = \begin{bmatrix} y_A \\ y_B \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$

Properties

1. Marginalization

Gaussian

$P(y_A) = \int_{y_B} P(y_A, y_B)\, dy_B \quad y_A \sim N(\mu_A, \Sigma_{AA})$

$P(y_B) = \int_{y_A} P(y_A, y_B)\, dy_A \quad y_B \sim N(\mu_B, \Sigma_{BB})$

2. Summation

If $y \sim N(\mu, \Sigma)$

$y' \sim N(\mu', \Sigma')$

$y + y' \sim N(\mu + \mu', \Sigma + \Sigma')$

Gaussian!

3. Conditioning

Gaussian

$$P(y_A | y_B) = \frac{P(y_A, y_B)}{\int_{y_A} P(y_A, y_B)\, dy_A}$$

$y_A | y_B = \tilde{y} \sim N(\mu_A + \Sigma_{AB} \Sigma_{BB}^{-1} (\tilde{y} - \mu_B),$

$\Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})$

4. Multiplication

$P(y_A) P(y_B) \to$ Gaussian

Gaussian

Gaussian Process Regression

$\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2$

$|||$

$\max_w P(D|w)$

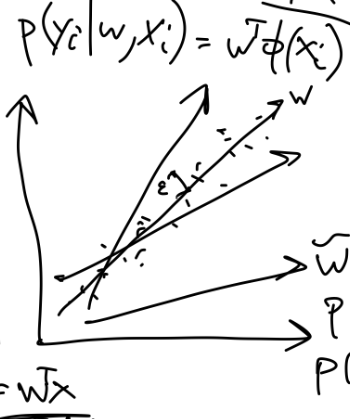
$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$

$P(y_i | w, x_i) = w^T x_i + \varepsilon$

$N(w^T x_i, \sigma^2)$

$\varepsilon \sim N(0, \sigma^2)$

$x_i \to \phi(x_i) \quad P(y_i | w, x_i) = w^T \phi(x_i) + \varepsilon$



$P(D|\tilde{w})$ is small

$P(\tilde{w}|D)$

$\cancel{D \xrightarrow{\text{training}} w \xrightarrow{\text{prediction}} y = w^T x}$

$w^T \phi(x)$

unknown

test point x^*, y^*

Sum over all models!!! → intractible

$$\underbrace{P(y | x^*, D)}_{\text{Gaussian}} = \int_w \underbrace{P(y | x^*, D, w)}_{\text{Gaussian}} \underbrace{P(w|D)}_{\text{Gaussian}} dw$$

Gaussian

assume: $P(x)$ is Gaussian

$$\frac{P(D|w) P(w)}{P(D)}$$

Gaussian

$= P(y | \cancel{X}, w)$

$P(\cancel{X} | w)$

$P(X)$

Gaussian

$P(y | \cancel{X}) P(X)$

Gaussian Gaussian Gaussian

$\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$

$|||$

$\max_w P(w|D)$

$P(w)$ is Gaussian

$P(y | x^*, D) \sim N(\mu, \sigma^2)$

Gaussian Process Regression

assumption:

$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim N(0, \Omega)$

kernel

$\Omega = \begin{bmatrix} \sigma_1^2 & 3 \\ 3 & \sigma_2^2 \end{bmatrix} \qquad \Omega = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$

p.s.d.

kernel!!!

goal:

$P(y | x^*, D) \sim N(\mu, \sigma^2)$

$\mu = K^{*T} K^{-1} y$

$\sigma^2 = K^{**} - K^{*T} K^{-1} K^*$

$\begin{bmatrix} k(x_1, x^*) \\ \vdots \\ k(x_n, x^*) \end{bmatrix} \quad k(x^*, x^*)$

$k_{ij} = k(x_i, x_j)$

$\Sigma = \begin{bmatrix} K & K^* \\ K^{*T} & K_{**} \end{bmatrix}$

$P\left(\begin{bmatrix} y_1 \\ \vdots \\ y_n \\ y \end{bmatrix} \middle| [x_1 \dots x_n\, x^*]\right)$

RBF

$K(x, z) = e^{-(x-z)^2/\sigma^2}$

similar

$x_1 \approx x_2$

$x_1 = \begin{bmatrix} \text{size} \\ \text{backyard} \\ \text{dist to} \\ \text{postbox} \end{bmatrix}$

$x_2 = \begin{bmatrix} \dots \end{bmatrix} \quad (x_1 - x_2)^2 \to 0$

$e^{-0} \to 1$

dissimilar

$x_3 = \begin{bmatrix} \dots \end{bmatrix} \quad x_1 \neq x_3$

$(x_1 - x_3)^2 \to \infty$

$e^{-\infty} \to 0$

Kernelized Least Squares!!

$h(x^*) = \phi(x^*)^T w = K^{*T} K^{-1} y$

$w = \Phi \alpha$

$K^{-1} y$

Summary

GP Regression properties:

* We get uncertainty for free σ^2
* GP work really well for small training sets
* Runtime $O(n^3)$ ← matrix inversion
  $n \to$ large
* GPs can do classification, but awkward
* Non-Gaussian likelihoods: Poisson
  $p(y | x, D, w)$ approximate inference
* GP implementation: GPyTorch, GPML (matlab)
  GPys, pyGPs, scikit learn (python)