Statistical Machine Learning

Lecture 03
Gaussian Process I

Sharif University of Technology

Spring 2021

Recall:

$D = \{x_1, x_2, \ldots x_n\}$ ← Data

Model: $p(x|\theta)$ ← likelihood function

$\underline{\theta}$: parameter Random / unknown

$p(\theta)$ ← prior on θ

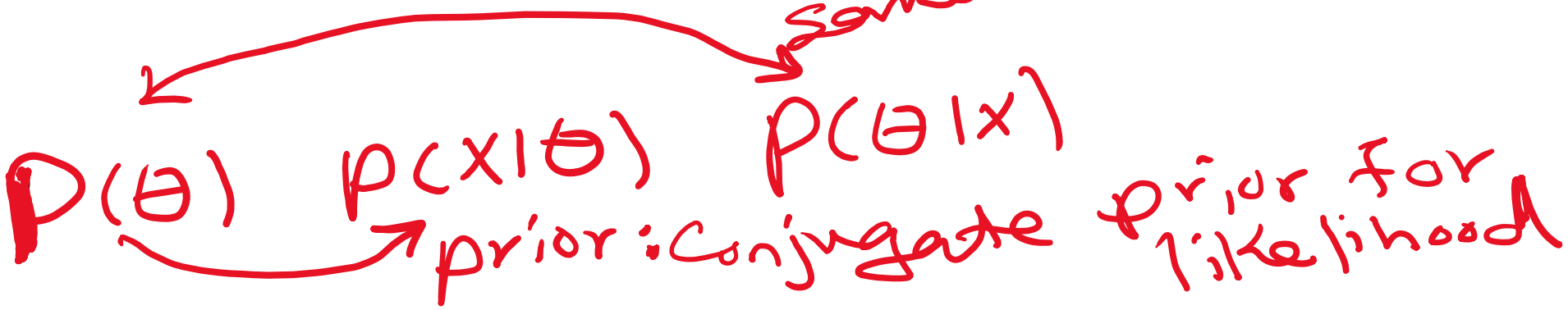
Bayes model: $p(\theta|x) = \dfrac{p(x|\theta)p(\theta)}{p(x)}$ ← posterior

θ finite size ← parametric

θ Not finite ← non parametric Bayesian (NPB)

Conjugate prior:

if $p(x|\theta)$ is in the same family of prob. density fn. as the prior $p(\theta)$ then prior and posterior called conjugate distributions or prior is the conjugate prior for the likelihood $p(x|\theta)$.



Examples of conjugate prior:

Gaussian family is conjugate to itself. (self-conjugate) with respect to a Gaussian likelihood

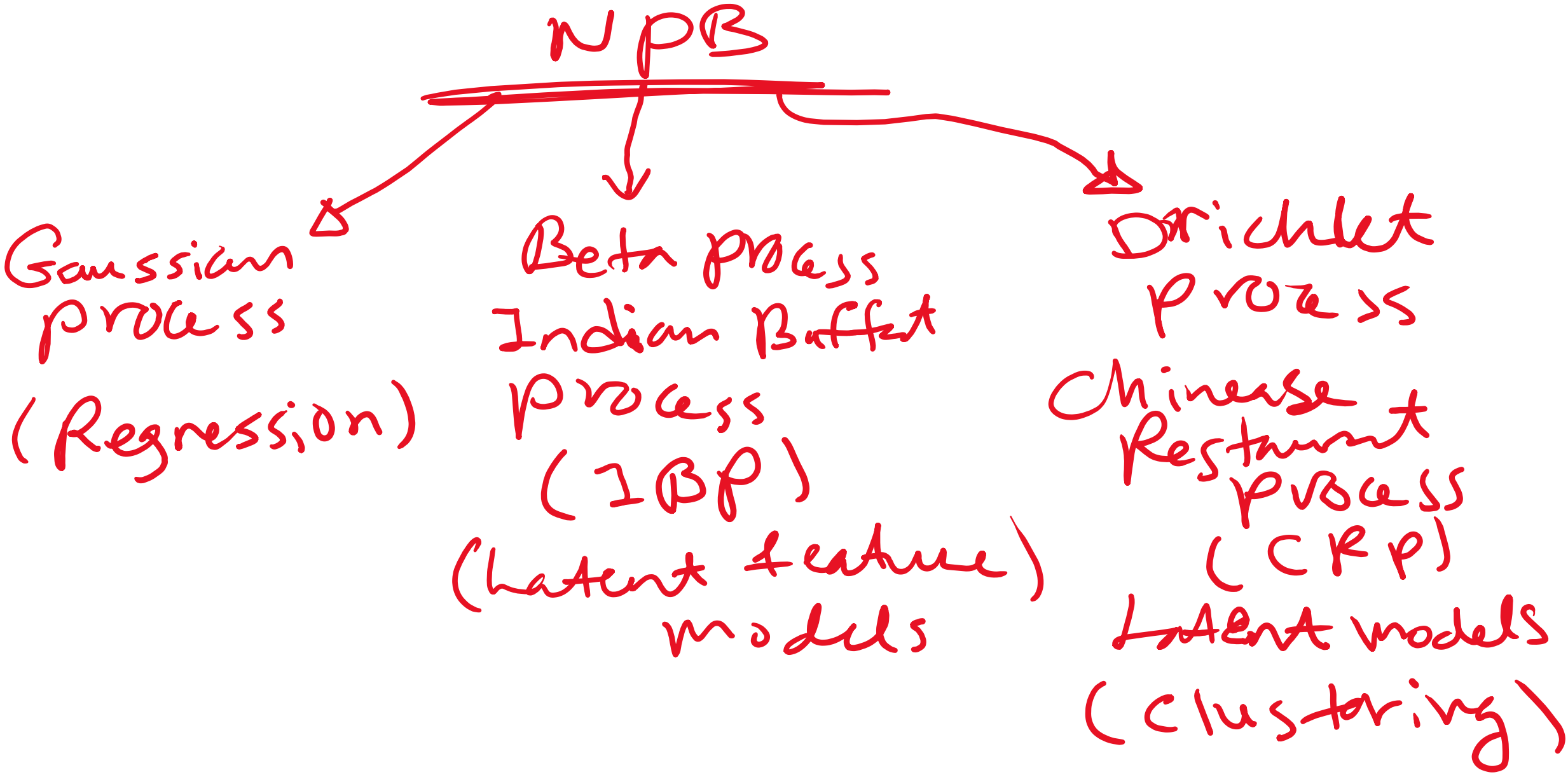
Normal prior → Normal likelihood (mean) → Normal posterior

Dirichlet dist. → multinomial likelihood → Dirichlet posterior

Dirichlet ↑ prior

multinomial likelihood ↑ likelihood

Dirichlet posterior ↑ posterior


NPB
Gaussian process
(Regression)
Beta process
Indian Buffet process
(IBP)
(latent feature) models
Dirichlet process
Chinese Restaurant process
(CRP)
LATENT models
(clustering)

Gaussian process:

Defines a dist. $p(f)$ over function f where f is a function mapping some input space $\mathcal{X}$ to $\underline{\underline{\mathbb{R}}}$: $f: \mathcal{X} \rightarrow \mathbb{R}$

Let $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))$

vector n-dim. $x_i \in \mathcal{X}$

$\mathbf{f}$ is a R.V.

Then $p(f)$ is a Gaussian process (GP) if for any finite subset $\{x_1, \dots x_n\} \subset X$ the marginal dist. over that subset $p(f)$ has multivariate Gaussian dist.

GPs are parametrized by a:

$\mu(x)$ ← vector

$C(x, x')$ ← covariance fn

x, x'

$(\theta = (\nu_1, \nu_2, \lambda, \alpha)$

$$p(f(x), f(x')) \sim N(\underline{\mu}, \underline{\Sigma})$$

$$\mu = \begin{bmatrix} \mu(x) \\ \mu(x') \end{bmatrix} \qquad \Sigma = \begin{bmatrix} C(x,x) & C(x,x') \\ C(x',x) & C(x',x') \end{bmatrix}$$

$$\rightarrow C(x_i, x_j) = \nu_0 \exp\left\{-\left(\frac{|x_i - x_j|}{\lambda}\right)^{\alpha} + \nu_1 + \nu_2 \delta_{ij}\right\}$$

⊛ How to use GP for nonlinear regression?

Data Set given $D = \{(x_i, y_i)_{i=1}^n\} = (\underline{x}, \underline{y})$ $f(\cdot)$

Model: $y_i = f(x_i) + \epsilon_i$

$f \sim GP(\cdot | \underline{0}, \underline{\underline{C}})$

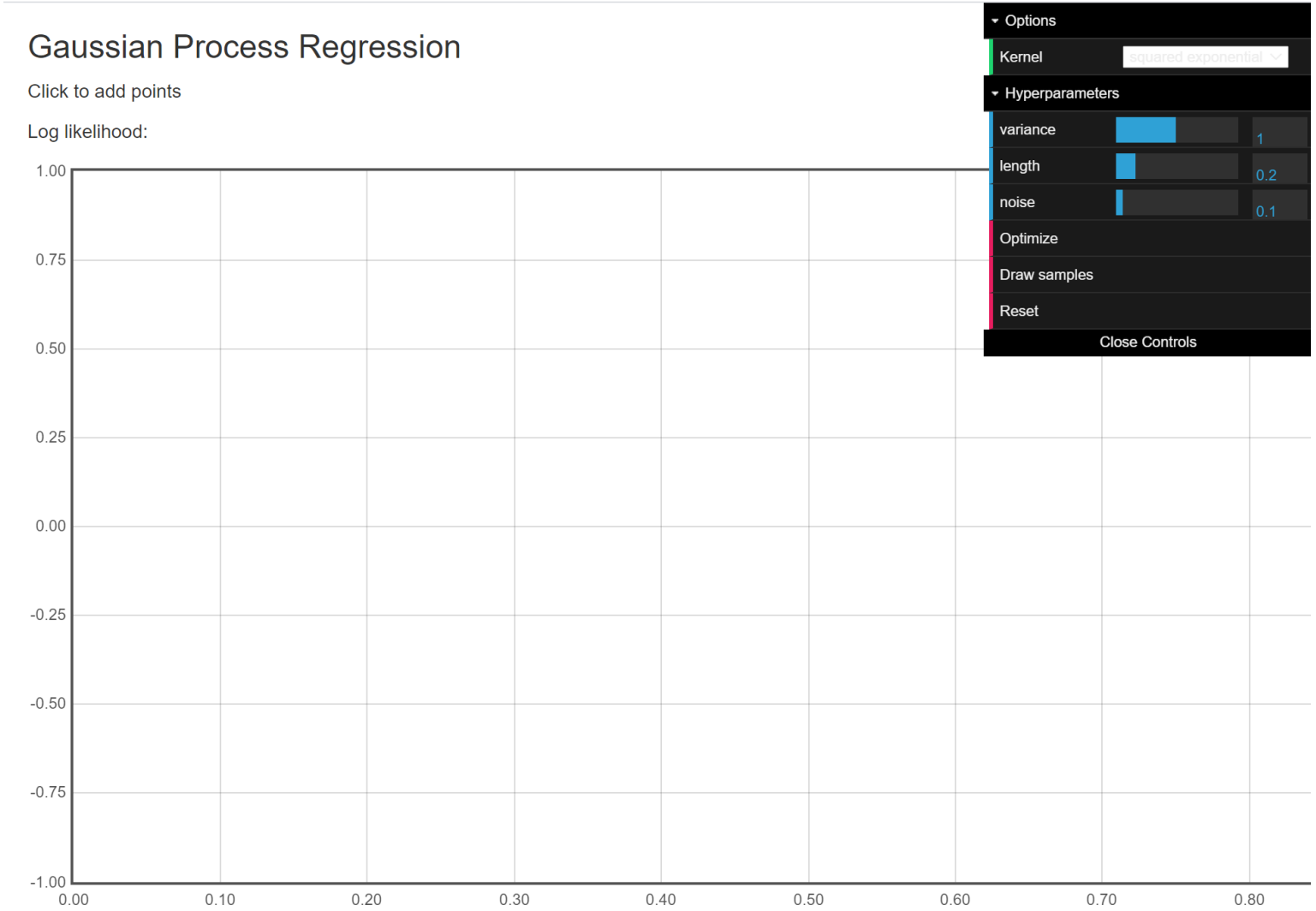
$\epsilon_i \sim N(\cdot | 0, \delta^2)$ ← Noise

prior of f is GP
likelihood is Normal
↓
posterior is GP

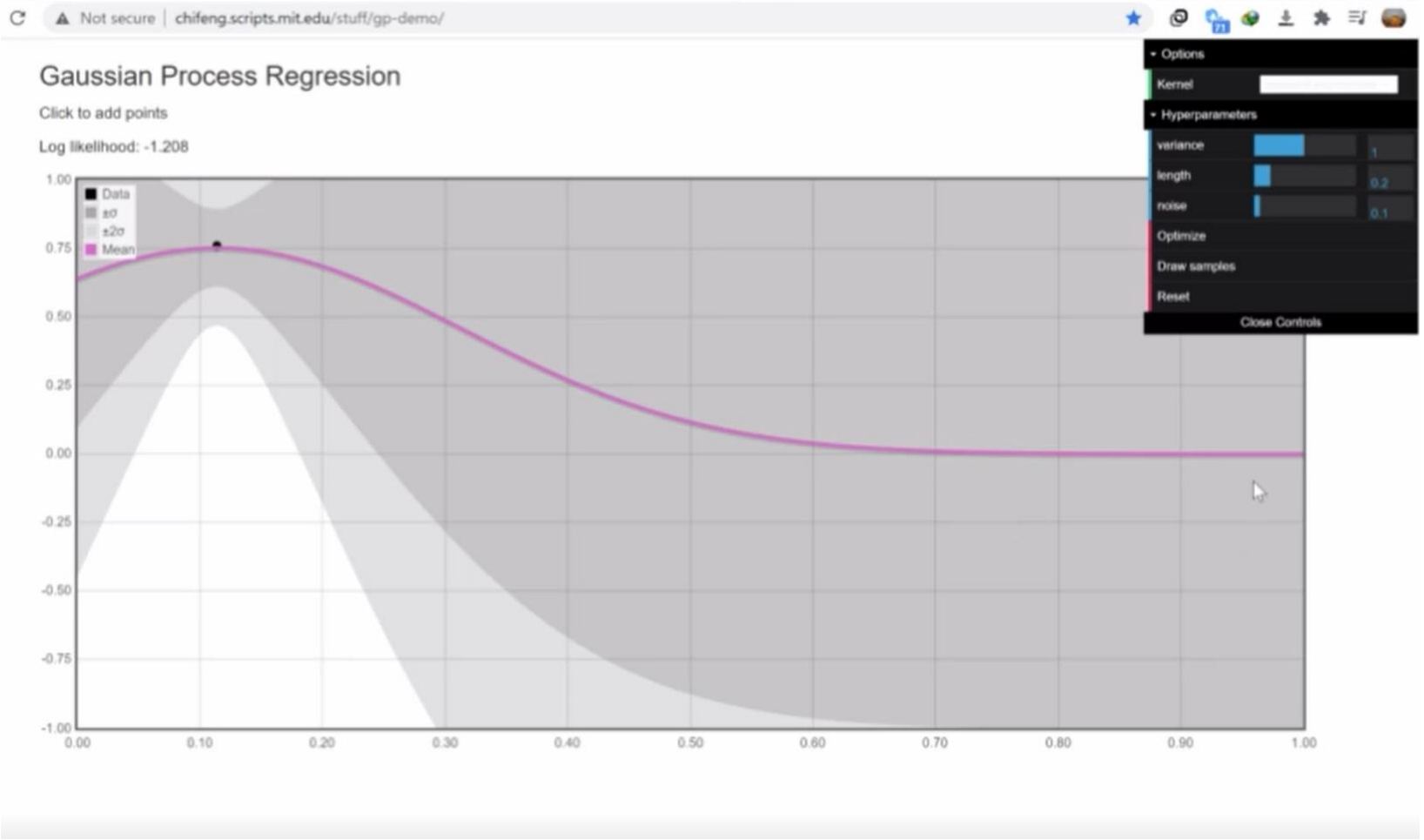
$$p(y' \mid x', D) = \int p(y' \mid x', f, D)\, p(f \mid D)\, df$$

$$p(y \mid x) = \int p(y \mid f, x)\, p(f)\, df$$

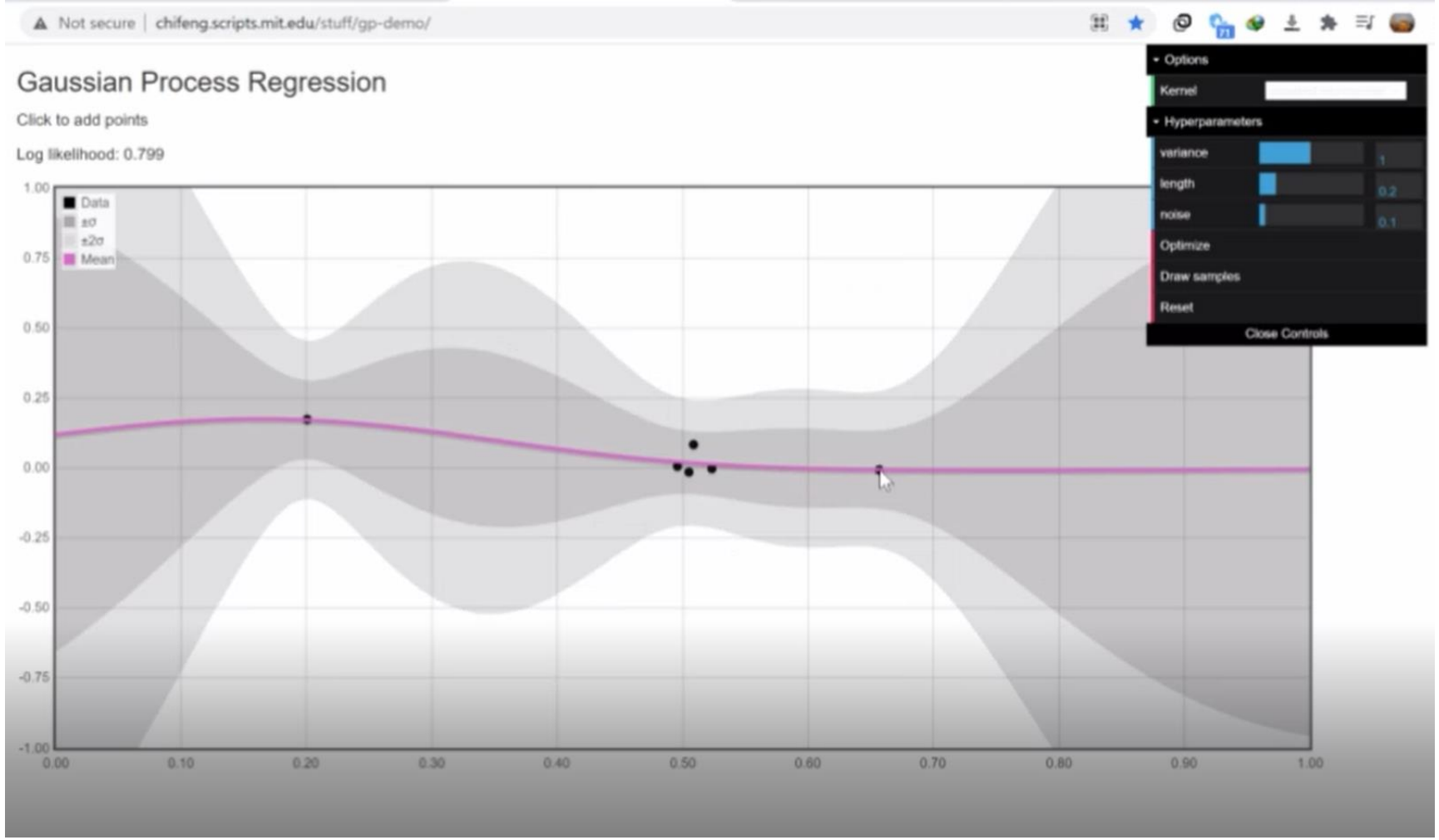
Demo: Gaussian Process: http://chifeng.scripts.mit.edu/stuff/gp-demo/



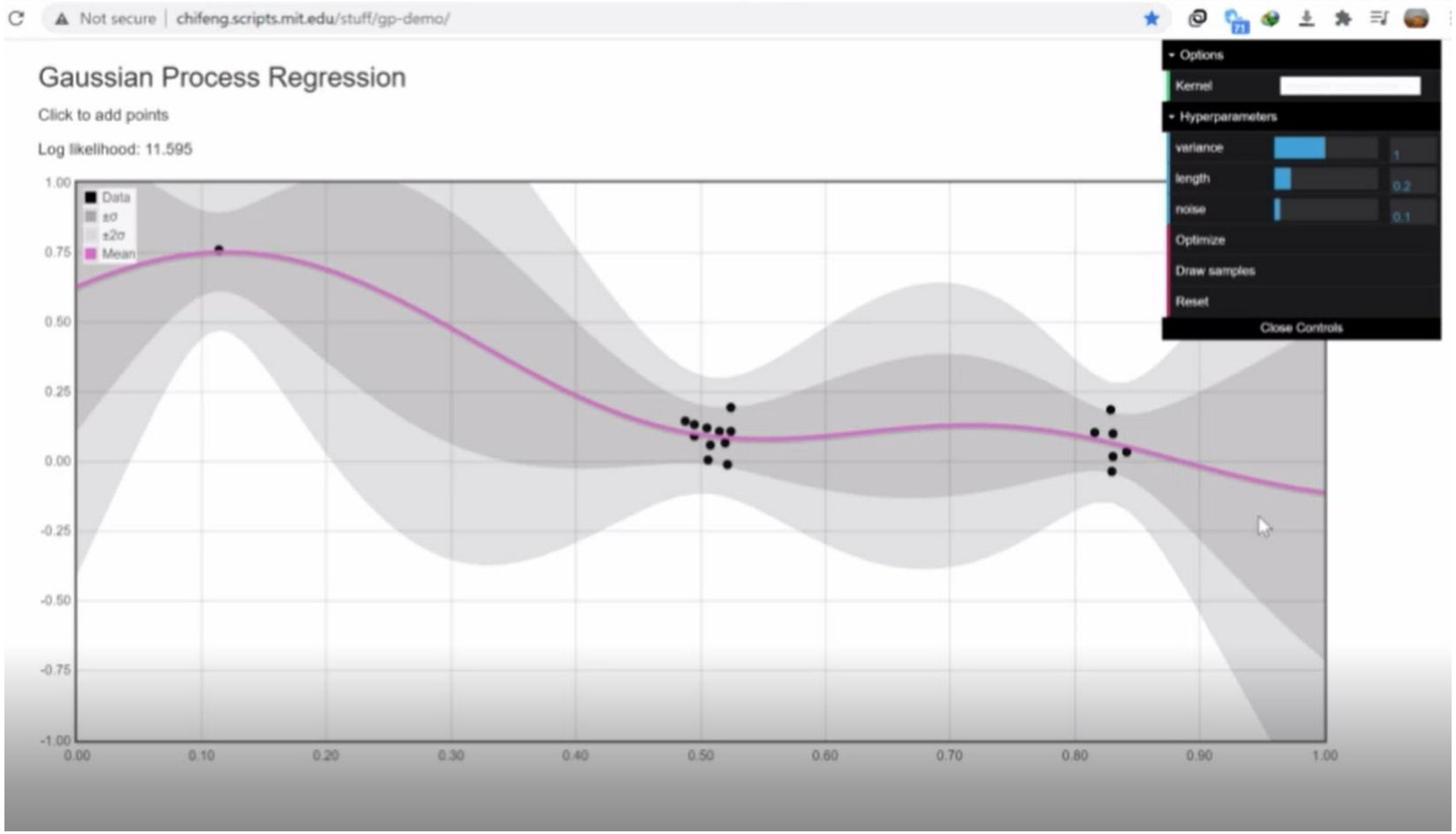
Demo: Gaussian Process: http://chifeng.scripts.mit.edu/stuff/gp-demo/



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



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



2. How do you use GP for linear regression?

x_i input y_i output

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Basis function $\phi_l(x_i)$

$$y_i = \sum_{l=1}^{K} \beta_l \phi_l(x_i) + \epsilon_i$$

← linear reg. with K Basis fn.

$\beta_l \sim N(\cdot | 0, \lambda_l)$

$\epsilon_i \sim N(\cdot | 0, \delta^2)$

$E[y_i] = 0$

$$\mathrm{Cov}(Y_i, Y_j) \triangleq C_{ij}$$

$$\triangleq \sum_{\ell} \lambda_\ell \phi_\ell(x_i)\phi_\ell(x_j) + \delta_{ij}\delta^2$$

$[Y_i]$ is a GP : cov. fn. $c(x_i, x_j) \triangleq C_{ij}$

Can we use GP for classification? Yes

How? for Binary classification

$Y \in \{-1, +1\}$