TAM 598 Lecture 22:

Gaussian Process Regression

Announcements:

- HW 6 covers lectures 21-23; due on Fri May 2

where we are:

* we have a Gaussian Process prior (an "infinite variate" gaussian) for unknown function $f(\cdot)$, given by $p(f(\cdot))$
* we are going to collect some data and use Bayes' rule to update our prior and obtain the posterior

$$p(f(\cdot) \mid D) \propto p(D \mid f(\cdot))\, p(f(\cdot))$$

→ (we'll do this for an arbitrary n-variate Gaussian; the approach generalizes to the continuous infinite case)

$$p(f_1, f_2, \ldots, f_n \mid D)$$

Gaussian Process : a distribution over functions

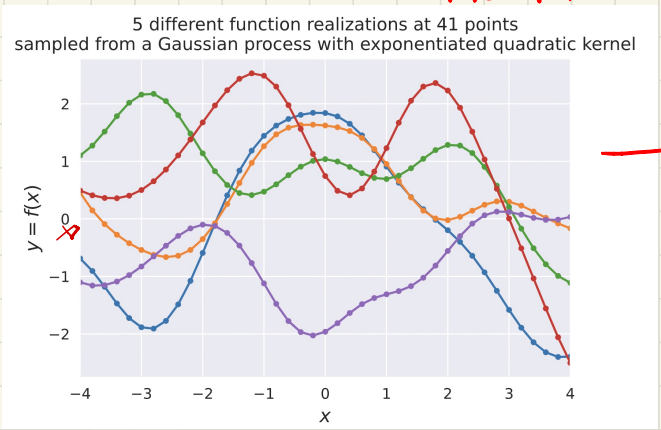
$$D = \{ (x_i, f_i), i = 1, \ldots, N \}$$

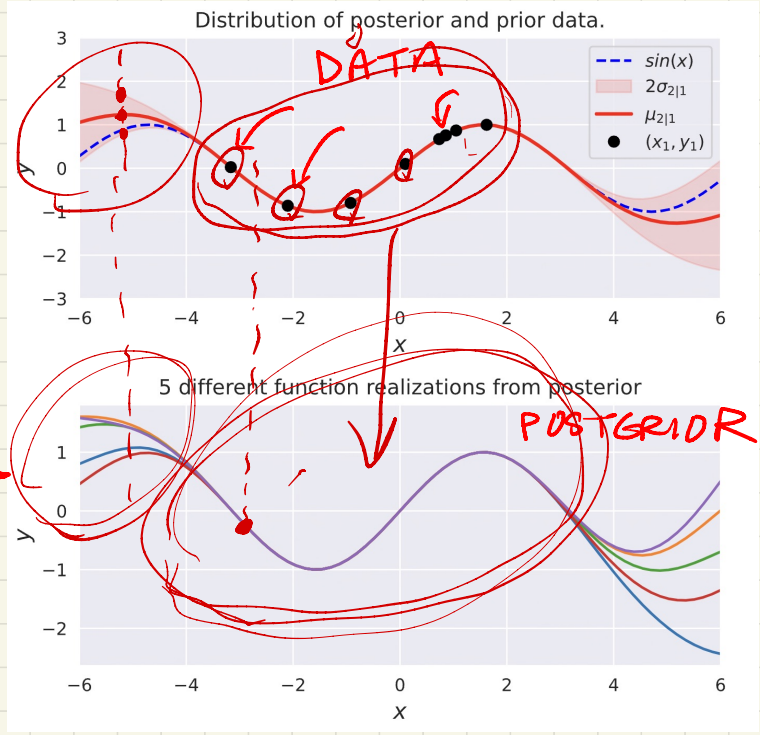
$$p(f|D) = \frac{p(D|f)\, p(f)}{p(D)}$$

prior

posterior

PRIOR





$f(\cdot) \sim GP(m(\cdot), k(\cdot,\cdot))$ so

for any points $\underline{x}_{1:n} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ the joint pdf of $\underline{f}_{1:n} = (f(\underline{x}_1), \ldots, f(\underline{x}_n))$ is the multivariate gaussian

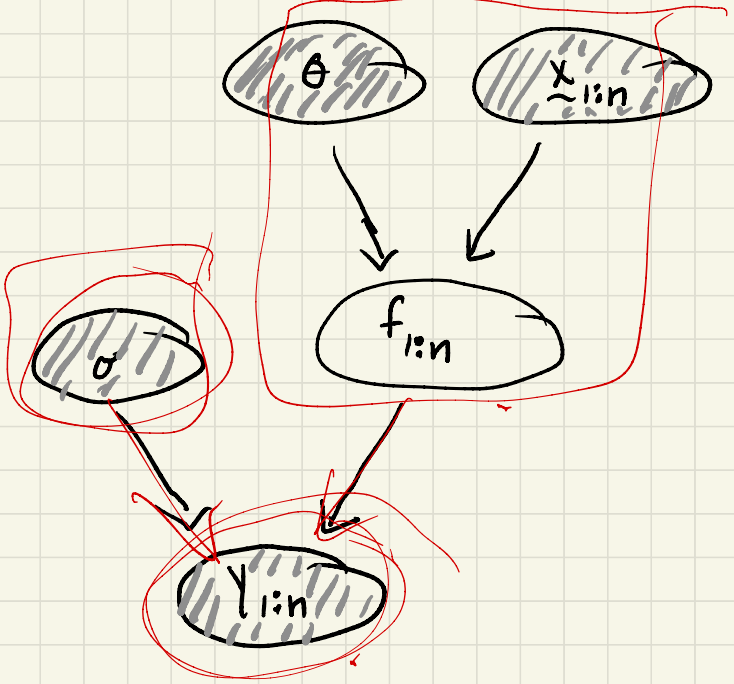
$$\underline{f}_{1:n} \mid \underline{x}_{1:n} \sim N\left(\underline{m}(\underline{x}_{1:n}), \underline{\underline{K}}(\underline{x}_{1:n}, \underline{x}_{1:n})\right)$$

we observe $D = (\underline{x}_{1:n}, y_{1:n})$

↳ noisy measure of $f(\underline{x}_{1:n})$, gaussian noise, variance σ^2

single observation $y_i \mid f(\underline{x}_i) \sim N(f(\underline{x}_i), \sigma^2)$

all observations: $y_{1:n} \mid f(\underline{x}_i) \sim N(\underline{f}_{1:n}, \sigma^2 \underline{\underline{I}}_n)$



- fixed hyperparameters θ and σ

- arbitrary collection of test points, densely cover input space

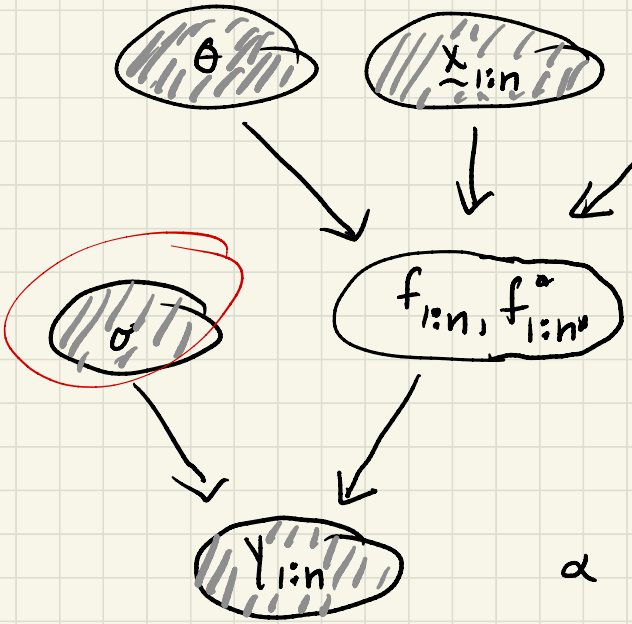
$$\underline{x}^*_{1:n^*} = (\underline{x}^*_1, \underline{x}^*_2, \ldots \underline{x}^*_{n^*})$$

and function values

$$\underline{f}^*_{1:n^*} = (f(\underline{x}^*_1), \ldots, f(\underline{x}^*_{n^*}))$$

consider the joint pdf of $\underline{f}$ and $\underline{f}^*$, a MVG

$$p(\underline{f}_{1:n}, \underline{f}^*_{1:n^*}) = N\left(\begin{pmatrix} f_{1:n} \\ f^*_{1:n^*} \end{pmatrix} \middle| \begin{pmatrix} \underline{m}(\underline{x}_{1:n}) \\ \underline{m}(\underline{x}^*_{1:n^*}) \end{pmatrix}, \begin{pmatrix} \underline{\underline{K}}(\underline{x}_{1:n}, \underline{x}_{1:n}) & \underline{\underline{K}}(\underline{x}_{1:n}, \underline{x}^*_{1:n^*}) \\ \underline{\underline{K}}(\underline{x}^*_{1:n^*}, \underline{x}_{1:n}) & \underline{\underline{K}}(\underline{x}^*_{1:n^*}, \underline{x}^*_{1:n^*}) \end{pmatrix}\right)$$



$$p(\underline{f}^*_{1:n^*} \mid \underline{x}^*_{1:n^*}, D) = p(\underline{f}^*_{1:n^*} \mid \underline{x}^*_{1:n^*}, \underline{x}_{1:n}, \underline{y}_{1:n})$$

$$= \int p(\underline{f}_{1:n}, \underline{f}^*_{1:n^*} \mid \underline{x}^*_{1:n^*}, \underline{x}_{1:n}, \underline{y}_{1:n}) \, d\underline{f}_{1:n}$$

$$\propto \int \underbrace{p(\underline{y}_{1:n} \mid \underline{f}_{1:n})}_{\text{gaussian}} \underbrace{p(\underline{f}_{1:n}, \underline{f}^*_{1:n^*} \mid \underline{x}^*_{1:n^*}, \underline{x}_{1:n})}_{\text{gaussian}} df_{1:n}$$

this result is a new Gaussian

$$p(\underline{f}^*_{1:n^*} \mid \underline{x}^*_{1:n^*}, D) = \mathcal{N}\left(\underline{f}^*_{1:n^*} \mid \underline{m}_n(\underline{x}^*_{1:n}), \underline{\underline{K}}_n(\underline{x}^*_{1:n^*}, \underline{x}^*_{1:n^*})\right)$$

the result is a new Gaussian

prior: $\underline{m}(x)$, $\underline{k}(x)$

$$p(\underline{f}^*_{1:n^*} \mid \underline{x}^*_{1:n^*}, D) = \mathcal{N}\left(\underline{f}^*_{1:n^*} \mid \underline{m}_n(\underline{x}^*_{1:n^*}), \underline{K}_n(\underline{x}^*_{1:n^*}, \underline{x}^*_{1:n^*})\right)$$

with posterior mean

$$\underline{m}_n(\underline{x}) = \underline{m}(x) + \underline{k}(x, \underline{x}_{1:n})\left(\underline{\underline{K}}(\underline{x}_{1:n}, \underline{x}_{1:n}) + \sigma^2 \underline{\underline{I}}_n\right)^{-1}\left(\underline{y}_{1:n} - \underline{m}(\underline{x}_{1:n})\right)$$

and posterior covariance

$$k_n(x, x') = \underline{k}(x, x') - \underline{k}(x, \underline{x}_{1:n})\left(\underline{\underline{K}}(\underline{x}_{1:n}, \underline{x}_{1:n}) + \sigma^2 \underline{\underline{I}}_n\right)^{-1} \underline{k}^T(x, \underline{x}_{1:n})$$

where $\underline{k}(x, \underline{x}_{1:n}) = (k(x, \underline{x}_1), k(x, \underline{x}_2), \dots, k(x, \underline{x}_n))$ is a cross-covariance.

since the test points are arbitrary $\Rightarrow$

$$f(\cdot) \mid D \sim GP(m_n(\cdot), k_n(\cdot, \cdot))$$

point predictive distribution – to predict $f(\cdot)$ at a single point; have your test points $\underline{x}^*_{1:n^*}$ just be a single point $\underline{x}^*$. Then

$$p(f(\underline{x}^*) \mid D) = N\left(f(\underline{x}^*) \,\middle|\, \underline{m}_n(\underline{x}^*), \underline{k}_n(\underline{x}^*, \underline{x}^*)\right)$$

this is a **predictive variance**

but the predicted **measurement** outcome y^* at $\underline{x}^*$ is

$$p(y^* \mid x^*, D) =$$

$$= N\left(f(\underline{x}^*) \,\middle|\, \underline{m}_n(\underline{x}^*), \underline{k}_n(\underline{x}^*, \underline{x}^*) + \sigma^2\right)$$

need to include **noise variance**

Hyperparameter Optimization

- covariance function: θ
- measurement variance: σ^2

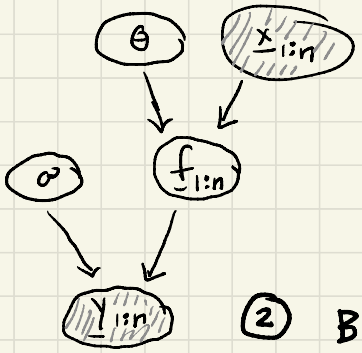
If we don't know them:

- start with priors $p(\theta)$, $p(\sigma)$
- Bayes Rule

$$p(\theta, \sigma \mid D) \propto p(D \mid \theta, \sigma)\, p(\theta)\, p(\sigma)$$

$$= \int p(\underline{y}_{1:n} \mid \underline{f}_{1:n}, \sigma)\, p(\underline{f}_{1:n} \mid \underline{x}_{1:n}, \theta)\, d\underline{f}_{1:n}\, p(\theta) p(\sigma)$$

- Maximum a posteriori estimation to evaluate



① joint pdf:

② Bayes Rule:

③ marginalize out the unobserved variable $\underline{f}_{1:n}$

Maximum a Posteriori (MAP) Estimate of Hyperparameters

$$p(\theta, \sigma | D) \approx \delta(\theta - \theta^*) \, \delta(\sigma - \sigma^*)$$

where θ^*, σ^* maximize $\log p(\theta, \sigma | D)$

so $\log p(\theta, \sigma | D) =$