TAM 598 Lecture 21:

Gaussian Process Regression

Announcements:

- HW 5 covers lectures 17-20; due on Fri Apr 18
- HW 6 covers lectures 21-23; due on Fri May 2

Gaussian Process Regression - fully non-parametric, Bayesian regression

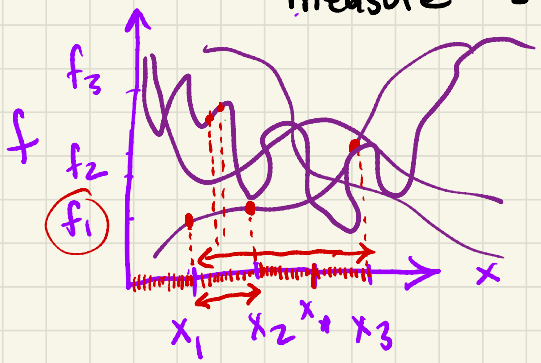
goal: given some data, learn a function $f(\cdot): \mathcal{X} \to \mathbb{R}$

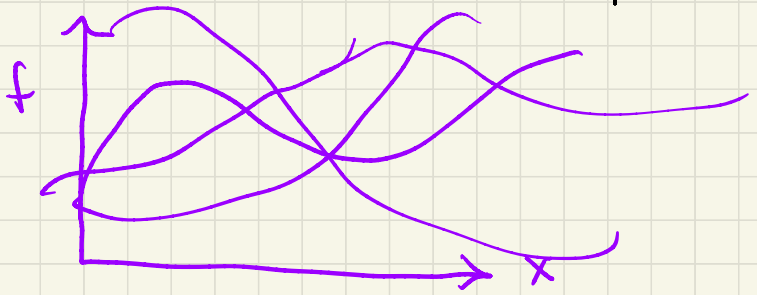
approach:

(1) before gathering any data, use your beliefs to generate a probability measure $p(f(\cdot))$ so that we can sample possible f's. This is your prior

(2) gather data D and model the likelihood of the data $p(D \mid f(\cdot))$

(3) use Bayes' rule to come up with your posterior probability measure over the space of functions $p(f(\cdot) \mid D)$





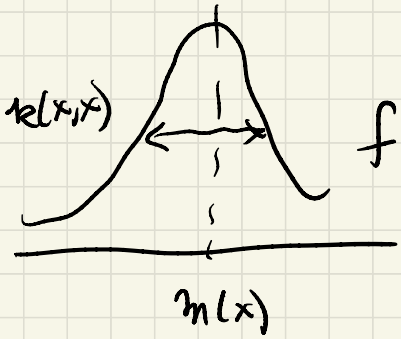
Stochastic process - collection of random variables $\{\underset{\sim}{x}_i\}$ for some i in set I.

eg $\underset{\sim}{X}_t = \underset{\sim}{X}(t)$ is a stochastic process, parametrized by time

usually discrete

gaussian process - a generalization of a multivariate gaussian to infinite dimensions, so a continuous stochastic process

"$f(\cdot)$ is a gaussian process" $\Longrightarrow$ f is a random variable that is a function



$$f(\cdot) \sim GP\left(m(\cdot), k(\cdot,\cdot)\right)$$

$m(\cdot)$: $\cdot$ = x

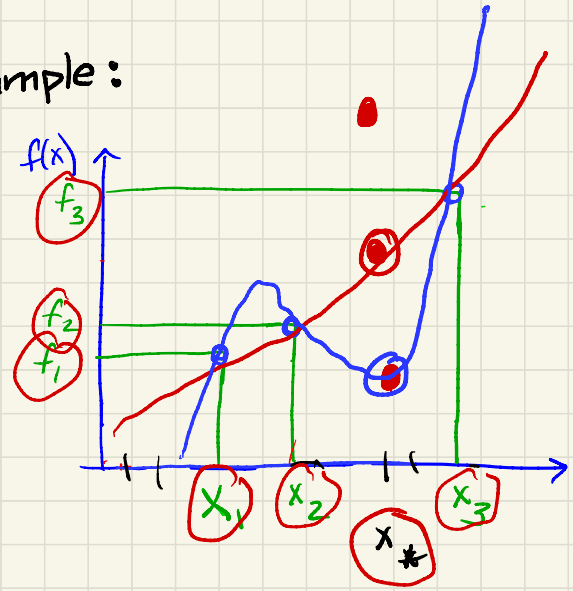
$k(\cdot,\cdot)$: $\cdot$ = x, x

difference w/ multivariate Gaussian

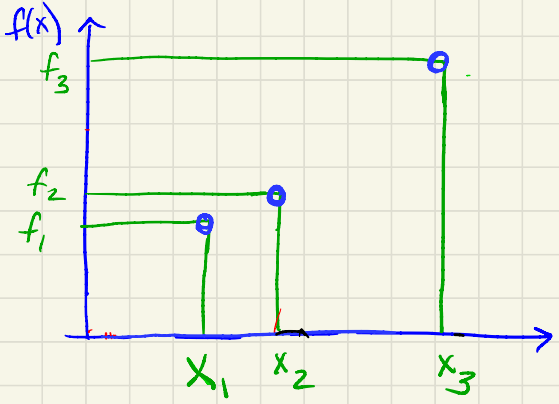
$f(\cdot) \sim GP(m(\cdot), k(\cdot,\cdot))$

- $f(\cdot)$: random function, not a random vector
- $m(\cdot)$: mean function, not mean vector
- $k(\cdot,\cdot)$: covariance function, not cov. matrix

example:



Say we know $f(x_1) = f_1$, $f(x_2) = f_2$, $f(x_3) = f_3$. We use a GP to predict f at any other point $x = x^*$



Approach:

① imagine that f_1, f_2, f_3 are components of a 3D random vector, generated by a 3-variate gaussian.

R.V.

$$\underset{\sim}{f} = \begin{pmatrix} f(\underset{\sim}{x}_1) \\ f(\underset{\sim}{x}_2) \\ f(\underset{\sim}{x}_3) \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}$$

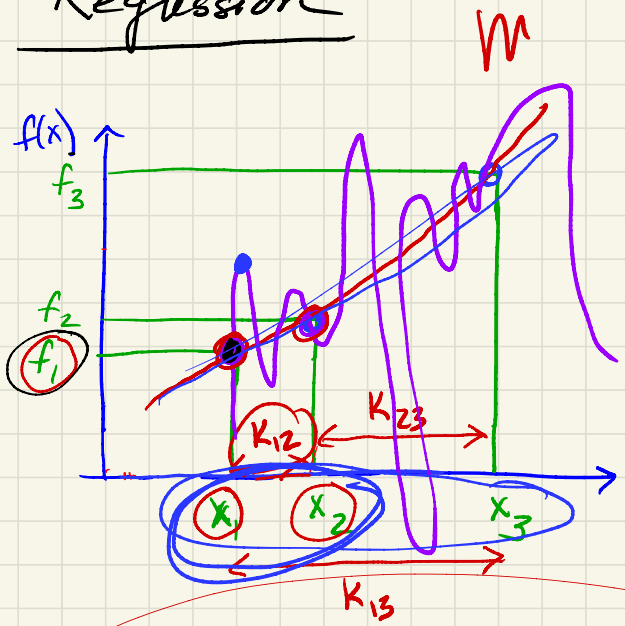
② Then

$$\underset{\sim}{f} \,\Big|\, \underset{\sim}{x}_{1:3}, m(\cdot), k(\cdot,\cdot) \sim N\left(\underset{\sim}{m}(\underset{\sim}{x}_{1:3}), k(\underset{\sim}{x}_{1:3}, \underset{\sim}{x}_{1:3}) \right)$$

where $\underset{\sim}{m}(\underset{\sim}{x}_{1:3}) = \begin{pmatrix} m(\underset{\sim}{x}_1) \\ m(\underset{\sim}{x}_2) \\ m(\underset{\sim}{x}_3) \end{pmatrix}$ and $k(\underset{\sim}{x}_{1:3}, \underset{\sim}{x}_{1:3}) =$

$$\begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & k(x_1,x_3) \\ k(x_2,x_1) & k(x_2,x_2) & k(x_2,x_3) \\ k(x_3,x_1) & k(x_3,x_2) & k(x_3,x_3) \end{pmatrix}$$

Regression



→ x's are given

→ model f's as if pulled from a multivariate gaussian

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

$$\sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0.2 \\ 0.7 & 1 & 0.5 \\ 0.2 & 0.5 & 1 \end{bmatrix} \right)$$

$$K_{ij} = \exp\left(-\frac{1}{\lambda_i} \| x_i - x_j \|^2 \right)$$

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

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

How to interpret the mean $m(\cdot)$?

$GP(m(\cdot), k(\cdot,\cdot))$

- For any point $\underset{\sim}{x}$, $m(\underset{\sim}{x})$ should give us the expected value $\mathbb{E}[f(\underset{\sim}{x})]$
- before we measure any data, options for $m(\underset{\sim}{x})$ are

prior

- ⊙ $m(\underset{\sim}{x}) = 0$ or $m(\underset{\sim}{x}) =$ constant
- ⊙ $m(\underset{\sim}{x}) = c_0 + \sum_{i=1}^{d} c_i \underset{\sim}{x}_i$ linear
- ⊙ use basis functions

$$m(\underset{\sim}{x}) = \sum_{i=1}^{m} c_i \phi_i(\underset{\sim}{x})$$

- ⊙ generalized polynomial chaos (gPC): use d polynomial basis functions up to degree p: $m(\underset{\sim}{x}) = \sum_{i=1}^{d} c_i \phi_i(\underset{\sim}{x})$ with $\int \phi_i(x)\,\phi_j(x)\,d\mu(x) = \delta_{ij}$

How to interpret the covariance?

- diagonal terms $k(\underset{\sim}{x}, \underset{\sim}{x})$ are the variance of the R.V. $f(\underset{\sim}{x})$

$$\mathbb{V}[f(\underset{\sim}{x})] = \mathbb{E}\left[(f(\underset{\sim}{x}) - m(\underset{\sim}{x}))^2\right]$$

so that w/ 95% confidence, $f(\underset{\sim}{x})$ falls in $m(\underset{\sim}{x}) \pm 2\sqrt{k(\underset{\sim}{x}, \underset{\sim}{x})}$

- off-diagonal terms $k(\underset{\sim}{x}, \underset{\sim}{x}')$ measure the correlation of $f(\underset{\sim}{x})$ and $f(\underset{\sim}{x}')$

$$k(\underset{\sim}{x}, \underset{\sim}{x}') = \mathbb{C}[f(\underset{\sim}{x}), f(\underset{\sim}{x}')]$$

$$= \mathbb{E}\left[(f(\underset{\sim}{x}) - m(\underset{\sim}{x}))(f(\underset{\sim}{x}') - m(\underset{\sim}{x}'))\right]$$

is a measure of similarity

properties of covariance function

1) $k(\underset{\sim}{x}, \underset{\sim}{x}) > 0$ since it is a variance

2) $k(\underset{\sim}{x}, \underset{\sim}{x}')$ becomes smaller as the distance between $\underset{\sim}{x}$, $\underset{\sim}{x}'$ grows

3) for any choice of specific points $\underset{\sim}{x}_{1:n}$, the covariance matrix K needs to be positive definite

- covariance functions govern smoothness
- they can be designed to model invariances, eg if $k(T\underset{\sim}{x}, \underset{\sim}{x}') = k(\underset{\sim}{x}, T\underset{\sim}{x}') = k(\underset{\sim}{x}, \underset{\sim}{x}')$ then the GP is also invariant w/r/t T

Gaussian Process : a distribution over functions

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

$$m(x) = \mathbb{E}[f(x)]$$

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

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

$N(m, \sigma^2)$

$= m + \sigma N(0, 1)$

$N_1 = m + L N(0, 1)$

$N(m_1, K)$

$N_1 = m_1 + L N(0, 1)$

$K = L L^T$