How we can do non-linear regression with MVN theory

$y = \beta^T x + \varepsilon$. no longer Linear

LECTURE 9

Gaussian process regression

Gaussian process model.

Nonparametric models.

The idea behind a Gaussian process regression is to place a distribution over a space of functions say $\mathcal{H}$. Consider for example an rkhs $\mathcal{H}_K$ over which we want to do Bayesian inference. Assume a regression model with the standard noise assumption

$$\boxed{Y_i = f(X_i) + \varepsilon_i,} \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \quad f \in \mathcal{H}_K.$$

→ Hilbert space defined by a fcn. $K : X \times X \to \mathbb{R}$

If we knew how to place a prior over the function space we in theory could do Bayesian inference.

9.1. Gaussian process

A Gaussian process is a specification of probability distributions over functions $f(x)$, $f \in \mathcal{H}$ and $x \in \mathcal{X}$ parameterized ny a mean function μ and a covariance function $K(\cdot,\cdot)$. The idea can be informally stated as

covariance Kernel

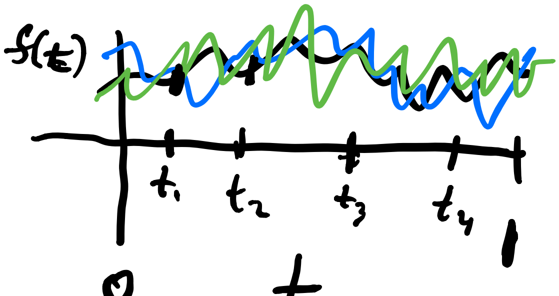
$$p(f) \propto \exp\Big(-\frac{1}{2}\|f\|^2_{\mathcal{H}_K}\Big), \quad p(f) \geq 0 \,\forall\, f \in \mathcal{H}, \int_{f\in\mathcal{H}} p(f)\,\mathrm{d}f = 1,$$

where we use the term informal because df is not well defined, it is not clear what the normalization constant is for $p(f)$ and what the space of functions $\mathcal{H}$ is not clear not is the relation of $\mathcal{H}$ to $\mathcal{H}_K$ stated clearly. Instead of making all the points clear we will develop Gaussian processes from an alternative perspective. There are many ways to define and think about a Gaussian process. A standard formulation is that a Gaussian process is an infinite version of a multivariate Gaussian distribution and has two parameters: a mean function μ corresponding to the mean vector and a positive definite covariance or kernel function K corresponding to a positive definite covariance matrix.

A common approach in defining an infinite dimensional object is by defining it's finite dimensional projections. This is the approach we will take with a Gaussian process. Consider $x_1, ..., x_n$ as a finite collection of points in $\mathcal{X}$. For a Gaussian process over functions $f \in \mathcal{H}$ the probability density of $\mathbf{f} = \{f(x_1), ..., f(x_n)\}^T$ is a multivariate normal with $\boldsymbol{\mu} = \{\mu(x_1), ..., \mu(x_n)\}$ and covariance $\boldsymbol{\Sigma}_{ij} = K(x_i, x_j)$

$$\mathrm{f} \sim \mathrm{N}\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\Big),$$

$$\mathbf{f}^T = \begin{bmatrix} f(t_1) \\ f(t_2) \\ f(t_3) \end{bmatrix} \sim N(\mu, \Sigma)$$



$[\, f(t_4) \,]$

$\mu(x) = \mathbb{E} f(x)$

$K(x_i, x_j) = \mathbb{E}[(f(x_i) - \mu(x_i))(f(x_j) - \mu(x_j))]$

$\mu(x) = 0$

$K(x_i, x_j) = \exp(-\tau \|x_i - x_j\|^2)$

where $\mu(x) = \mathbb{E}f(x)$ and $K(x_i, x_j) = \mathbb{E}\big[(f(x_i) - \mu(x_i))(f(x_j) - \mu(x_j))\big]$ and

$$f \sim \mathcal{GP}(\mu(\cdot), K(\cdot,\cdot)).$$

Definition. *A stochastic process over domain $\mathcal{X}$ with mean function μ and covariance kernel K is a Gaussian process if and only if for any $\{x_1, ..., x_n\} \in \mathcal{X}$ and $n \in \mathbb{N}$ the distribution of $\mathbf{f} = \{f(x_1), ..., f(x_n)\}^T$ is*

$$\mathbf{f} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim N\left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) \\ \vdots & \ddots & \vdots \\ K(x_1,x_n) & \cdots & K(x_n,x_n) \end{bmatrix} \right).$$

training responses

9.2. Gaussian process regression

Consider data $D = \{(x_i, y_i)\}_{i=1}^n$ drawn from the model

$$Y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2),$$

$\mu(\cdot) = 0$

we will place a prior on the space of functions using a Gaussian process

$$f \sim \mathcal{GP}(\mu(\cdot), K(\cdot,\cdot)).$$

We are also given some new variables or test data $T = \{x_i^*\}_{i=1}^m$ each of which would have a corresponding y_i^*.

test values or values we want to predict on

We now provide some notation

training covariates

$$\mathbf{X} = \begin{bmatrix} -x_1- \\ \vdots \\ -x_n- \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} -x_1^*- \\ \vdots \\ -x_m^*- \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{Y}^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_m^* \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\varepsilon}^* = \begin{bmatrix} \varepsilon_1^* \\ \vdots \\ \varepsilon_m^* \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad \mathbf{f}^* = \begin{bmatrix} f(x_1^*) \\ \vdots \\ f(x_m^*) \end{bmatrix}.$$

predicted response

Our ultimate objective will be to specify the predictive distribution on $\mathbf{Y}^*$ which we know will be multivariate normal

$$\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{X} \sim \mathrm{N}(\mu^*, \Sigma^*).$$

Now first observe

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}^* \end{bmatrix} \Big| \mathbf{X}^*, \mathbf{X} = \begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}^* \end{bmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma^2 \mathbf{I} & K(\mathbf{X},\mathbf{X}^*) \\ K(\mathbf{X}^*,\mathbf{X}) & K(\mathbf{X}^*,\mathbf{X}^*) + \sigma^2 \mathbf{I} \end{bmatrix} \right),$$

where $K(\mathbf{X},\mathbf{X})$ is the $n \times n$ matrix with $\mathbf{K}_{ij} = K(x_i, x_j)$ and $K(\mathbf{X}^*,\mathbf{X}^*)$ is the $m \times m$ matrix with $\mathbf{K}_{ij}^* = K(x_i^*, x_j^*)$.

To get to the predictive distribution on $\mathbf{Y}^*$ we write the conditional $\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{X}$. Given the above multivariate normal distribution we simply condition on all the other variables to get the mean and covariance for the normal distribution for the posterior predictive density:

$$\begin{aligned} \mu^* &= K(\mathbf{X}^*,\mathbf{X})(K(\mathbf{X},\mathbf{X}) + \sigma^2 \mathbf{I})^{-1}\,\mathbf{Y} \\ \Sigma^* &= K(\mathbf{X}^*,\mathbf{X}^*) + \sigma^2 \mathbf{I} - K(\mathbf{X}^*,\mathbf{X})(K(\mathbf{X},\mathbf{X}) + \sigma^2 \mathbf{I})^{-1} K(\mathbf{X},\mathbf{X}^*). \end{aligned}$$

The beauty of Gaussian process regression is that we can place priors over functions using a kernel and evaluating the variance of the function values at a finite number of points, all just based on properties of the multivariate normal

$Y^* | X^*, X, Y \sim N(\mu^*, \Sigma^*)$

Extra credit:

is $K(x_i, x_j) = x_i^T x_j$

Show if [illegible]
then GP regression reduces
to standard Bayesian linear
regression with a normal
prior

distribution. This is a very powerful non-linear prediction tool. There is a strong relation between the kernels, rkhs and Gaussian processes. There are also some subtle differences. The main difference comes from what is called the Kalianpur $0 - 1$ law

Theorem (Kallianpur 1970). *If $Z \sim \mathcal{G}P(\mu, K)$ is a Gaussian process with covariance kernel K and mean $\mu \in \mathcal{H}_K$ and $\mathcal{H}_K$ is infinite dimensional then*

$$\mathbf{P}(Z \in \mathcal{H}_K) = 0.$$

The point of the above theorem is that if we specify a kernel K and ensure the mean of the Gaussian process is in the rkhs $\mathcal{H}_K$ corresponding to the kernel K, draws from this Gaussian process will not be in the rkhs. What one can formally show is that if one takes any of the random functions, call them g then the following is true for all g

$$\int_{\mathcal{X}} g(u) K(x, u) \, \mathrm{d}u \in \mathcal{H}_K.$$