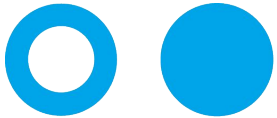


Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Gaussian Process

Manoj Kolpe Lingappa

Introduction

- In parametric model p**(y|X, θ)** to explain the data and find optimal value of **θ** using MLE.
- Posterior can be obtained using the formula p**(θ|X, y)**
- Gaussian processes is a non parametric method.
- Gaussian process can be used to infer a distribution over function directly.
- Gaussian process define a prior over function and once some new data points are observed, a posterior can be inferred.
- Inference of continuous function values in this context is known as GP regression.

Gaussian process

- A Gaussian process is a random process where any points $x \in Rd$ is assigned a random variable f(x) where the joint distribution of a finite number of these variables p(f(x1),…,f(xN)) is itself Gaussian:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K})$$

- f = (f(x1), f(x2), …….., f(xn)), μ = (m(x1), ……., m(xn)) and Kij = k(xi, xj) is the kernel, it is common to use m(x) = 0
- Gaussian process is a distribution over function and shape of distribution is defined by K.

Introduction

- Let's consider a noise free function **f** for input **X.**
- A GP prior can be converted into a GP posterior p(f∗|X∗,X,f) which can be used to make prediction f∗ at new inputs X∗
- By definition of marginalization, the joint distribution of observed values f and predictions f∗ is again a Gaussian which can be partitioned into

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$

- where K∗=κ(X,X∗) and K∗∗=κ(X∗,X∗)
- With N training data and N∗ new input data K is a N×N matrix , K∗ a N×N∗ matrix and K∗∗ a N∗×N∗ matrix.
- The predicted distribution is given by,

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) &= \mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{y} \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_* \end{aligned}$$

Introduction

- where Ky=K+σ2yI
- To additionally include noise ϵ into predictions y∗ we have to add σ2y to the diagonal of Σ∗

$$p(\mathbf{y}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{y}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_* + \sigma_y^2\mathbf{I})$$

- Let's take a RBF kernel

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)\right)$$

RBF kernels

The **squared exponential kernel** (SE kernel) or **Gaussian kernel** is defined by

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}')\right)$$

If $\boldsymbol{\Sigma}$ is diagonal, this can be written as

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\sum_{j=1}^{D}\frac{1}{\sigma_j^2}(x_j - x_j')^2\right)$$

$$k(x, x*) = exp(-\frac{||x - x*||^2}{2\sigma^2})$$

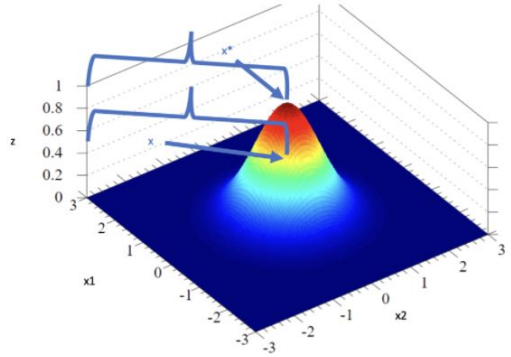
If you take the function apart, you'll see that part of the function is a measure of the squared distance between x and x*

$$||x - x*||^2$$

$$k(x, x*) = exp(-\frac{\sqrt{2}}{2 \times 2^2}) = 0.84$$

For example, our vectors x and x* would approximately look like this on our Gaussian kernel.

$$x* = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \; x = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$



First, we have to calculate the Euclidian distance between the two vectors.

$$||x - x*||^2 = \sqrt{(-1-0)^2 + (-1-0)^2} = \sqrt{2}$$

Then we plug this result into the Gaussian kernel, assuming that the standard deviation σ equals 1.

$$k(x, x*) = exp(-\frac{\sqrt{2}}{2 \times 1^2}) = 0.49$$

The similarity is pretty close to 0.5