Lecture 12

Gaussian Process Regression

We have discussed a variety of techniques for function approximation. Primarily, we have focused on the regression task, and we have done this by restricting the class of functions considered (e.g., radial basis functions, trigonometric functions, etc.). In this chapter, we will review Gaussian Process Regression, which tries to give a prior probability to every possible function. Higher probabilities are given to functions that are considered to be more likely.

12.1. Review of Gaussian Mixture Models

Consider

$$f(x_i) = \sum_j w_j k(-||x_i - c_j||) \tag{12.1}$$

where $k(-||x_i - c_j||) = \exp(-\beta_j||x_i - c_i||)$ or $\exp(-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j))$. We must decide the number of basis functions, and μ_j, Σ_j, w_j for each basis function.

12.2. Stochastic Processes

We can think of stochastic processes as a distribution over functions. Consider

$$x_{k+1} = x_k + w_k \tag{12.2}$$

where $w_k \sim \mathcal{N}(0, \nu)$.

For each sample of these stochastic dynamics, we have a different function $x[k]$.

12.3. Gaussian Processes

Let us define a Gaussian process to be a collection of random variables, any finite number of which have a joint Gaussian distribution.

For a function $f(x)$, a Gaussian process is defined as

$$m(x) = \mathbb{E}[f(x)] \tag{12.3}$$
$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{12.4}$$

The Gaussian process is typically written as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')). \tag{12.5}$$

Here the random variables represent the value of the function at location x. For a finite set of the domain x, where $X = \{x_1 ... x_n\}$, by definition the distribution is given as

$$f(X) \sim \mathcal{N}(m(X), k(X, X)) \tag{12.6}$$

with mean vector $\mu = m(X)$ and covariance $\Sigma = k(X, X)$. The covariance function is often chosen to be an exponentiated quadratic given as

$$k(x, x') = \exp\left(-\frac{1}{2\sigma^2}||x - x'||^2\right) \tag{12.7}$$

Also, we have a useful property, the marginalization property which states that if we have

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma). \tag{12.8}$$

we also have

$$y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \tag{12.9}$$

where Σ_{11} is a sub-matrix of Σ.

If $X \sim \mathcal{N}(\mu, \Sigma)$ and we have

$$Y = BX + b \implies \quad Y \sim \mathcal{N}(B\mu + b, B\Sigma B^T). \tag{12.10}$$

Let

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{12.11}$$

and let $B = \begin{bmatrix} I & 0 \end{bmatrix}$ and $b = 0$, then $X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$.

12.4. Prediction using Gaussian Processes

We would like to make predictions about future data points from past data points. For instance, assume you would like to evaluate y_2 where $y_2 = f(X_2)$ and X_2 represent new sample points. To do this, form the distribution $p(y_2|y_1, X_1, X_2)$. Since y_1 and y_2 come from the same multivariate distribution, we can write:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \tag{12.12}$$

where

$$\mu_1 = m(X_1) \tag{12.13}$$
$$\mu_2 = m(X_2) \tag{12.14}$$
$$\Sigma_{11} = k(X_1, X_1) \tag{12.15}$$
$$\Sigma_{22} = k(X_2, X_2) \tag{12.16}$$
$$\Sigma_{12} = k(X_1, X_2) = \Sigma_{21}^T \tag{12.17}$$

The conditional distribution is then given as

$$p(y_2|y_1, X_1, X_2) = \mathcal{N}(\mu_{2|1}, \Sigma_{2|1}) \tag{12.18}$$

where

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1) \tag{12.19}$$
$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{12.20}$$

We also must account for the case where the predictions come from noisy measurements, i.e.

$$f(X_1) = y_1 + \epsilon \tag{12.21}$$

If we model ϵ as i.i.d. Gaussian noise with variance σ, then we can write

$$\Sigma_{11} = k(X_1, X_1) + \sigma_\epsilon^2 I \tag{12.22}$$

12.5. Computing Hyperparameters

Both the covariance function and the mean function depend on parameters θ, often via a nonlinear relationship.

What we would like to do is

$$\theta^* = \arg\max_\theta p(y|X,\theta) \tag{12.23}$$

For a Gaussian process, the marginal likelihood is a Gaussian distribution, given as

$$p(y|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) \tag{12.24}$$

If $\mu_\theta = m_\theta(X)$ and $\Sigma_\theta = k_\theta(X,X)$, we can write

$$p(y|X,\theta) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_\theta|}} \exp\left(-\frac{1}{2}(y-\mu_\theta)^T\Sigma_\theta^{-1}(y-\mu_\theta)\right) \tag{12.25}$$

Take the log to compute the log likelihood.

$$\log p(y|X,\theta) = -\frac{1}{2}(y-\mu_\theta)^T\Sigma_\theta^{-1}(y-\mu_\theta) - \frac{1}{2}\log|\Sigma_\theta| - \frac{d}{2}\log 2\pi \tag{12.26}$$

Take the gradient and apply gradient descent.

Bibliography

[1] Miao Liu, Girish Chowdhary, Bruno Castra Da Silva, Shih-Yuan Liu, and Jonathan P How. Gaussian processes for learning and control: A tutorial with examples. *IEEE Control Systems Magazine*, 38(5):53–86, 2018.

Lecture 13

Koopman Operators

Consider the discrete time dynamical system

$$x_{k+1} = F(x_k) \tag{13.1}$$

where $x_k \in M$. The observed state of the system $y_k \in \mathbb{C}$. Define the observation

$$y_k = g(x_k) \tag{13.2}$$

Here $g \in \mathbb{G} : M \to \mathbb{C}$ and $\mathbb{G}$ is a function space. The Koopman operator is given as $\mathcal{K} : \mathbb{G} \to \mathbb{G}$ and defined as

$$[Kg](x) = g(F(x)) \tag{13.3}$$

We can also write

$$[Kg](x_k) = g(x_{k+1}) \tag{13.4}$$

To approximate the Koopman operator, we can write

$$y_k = g(x_k) = \Psi(x_k) \tag{13.5}$$

where

$$\Psi(x) = \begin{bmatrix}\psi_1(x), \psi_2(x), ...\psi_N(x)\end{bmatrix} \tag{13.6}$$

$$\Psi(x_{k+1}) = \Psi(x_k)K + r(x_k) \tag{13.7}$$

Here $K \in \mathbb{C}^{N \times N}$ and $r(x_k)$ is the residual error.

Assume some system trajectory $X = [x_1, x_2..x_P]$, where P is the number of data points.

Solve the least-squares problem to get

$$K = G^{\dagger}A \tag{13.8}$$

where

$$G = \frac{1}{P}\sum_{p=1}^{P-1} \Psi(x_p)^T \Psi(x_p) \tag{13.9}$$

and

$$A = \frac{1}{P}\sum_{p=1}^{P-1} \Psi(x_p)^T \Psi(x_{p+1}) \tag{13.10}$$

To approximate a dynamical system, we can write

$$\Psi(x) = [x^T, \psi_{n+1}(x)...\psi_N(x)] \tag{13.11}$$

where the approximate dynamics are given as

$$x_{k+1} \approx \hat{K}^T \Psi(x_k)^T \tag{13.12}$$

where $\hat{K}^T \in \mathbb{R}^{n \times N}$ is the first n columns of K.

Bibliography

[1] Ian Abraham, Gerardo De La Torre, and Todd D Murphey. Model-based control using koopman operators. *arXiv preprint arXiv:1709.01568*, 2017.