CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Gaussian Processes I

Lecturer: Michael I. Jordan *Scribes: Bruno Sinopoli*

Definition Given an index set $\mathcal{X}$, a collection of random variables $Z(x)$, with $x \in \mathcal{X}$ is a Gaussian Process if, for every finite set $\{x_1, \ldots, x_n\}$, $\{Z(x_i)\}$ has a multivariate Gaussian distribution, with mean $\mu \in \mathcal{R}^n$, and covariance $K \in \mathcal{R}^{n \times n}$.

Often $\mathcal{X} = \mathcal{R}$.

Bayesian Linear Regression For simplicity we shall cover the scalar case.

$Y_i = \theta^T \phi(x_i) + \xi_i$, where $\xi_i \sim N(0, \sigma^2)$, and $\theta \sim N(0, \Sigma)$, ϕ is an arbitrary transformation of x. Regardless of the transformation of x the expression is still linear in the parameter θ. The log likelihood:

$$\log p(\theta | y_1, \ldots, y_N) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \theta^T \phi(x_n))^2 + \frac{1}{2} \theta^T \Sigma^{-1} \theta$$

is the squared error. The MAP estimate is the θ that minimizes the expression above. Let's now define $\beta \stackrel{\triangle}{=} \frac{1}{\sigma^2}$, and the design matrix $\Phi \stackrel{\triangle}{=} \{\phi_j(x_i)\}_{ij}$. In matrix notation, finding the MAP estimate corresponds to minimizing the following objective function:

$$J = \frac{\beta}{2} (y - \Phi\theta)^T (y - \Phi\theta) + \frac{1}{2} \theta^T \Sigma^{-1} \theta$$

and the MAP estimate is:

$$\hat{\theta}_{MAP} = (\Sigma^{-1} + \beta \Phi^T \Phi)^{-1} \Phi^T y$$

As we can read from the solution, the MAP treatment amounts to incorporating a Σ^{-1} in the normal equations.

Let's define $A \stackrel{\triangle}{=} (\Sigma^{-1} + \beta \Phi^T \Phi)$:

$$\hat{\theta}_{MAP} = \beta A^{-1} \Phi^T y.$$

The prediction at x_* is

$$\hat{y}(x_*) = \hat{\theta}_{MAP}^T \phi(x_*) = (\beta \phi^T(x_*) A^{-1} \Phi^T) y$$

The prediction is a linear combination of the observations y.

Let's now define:

$$Z(x) \stackrel{\triangle}{=} \theta^T \phi(x) = \sum_i \theta_i \phi_i(x)$$

If θ is Gaussian, Z(x) is also Gaussian with the following mean and variance:

$$E[Z(x)] = E[\theta^T \phi(x)] = \phi^T(x) E[\theta] = 0$$

$$E[Z(x)Z(x')] = E[(\phi^T(x)\theta)(\theta^T \phi(x'))] = \phi^T(x) E[\theta\theta^T] \phi(x') = \phi^T(x) \Sigma \phi(x')$$

Given the set $\{x_i \in \mathcal{X} : i = 1, \ldots, m\}$ the random variables $\{Z(x_i)\}$ are jointly Gaussian with mean $\mu_m = 0$ and covariance $K_{m\times n} = (\phi(x_i)\Sigma\phi(x_j))_{ij}$.

Prediction for General Gaussian Processes Consider the multivariate Gaussian Random Variables $(Z_1, \ldots, Z_N, Z_*)$ with zero mean and covariance:

$$K_+ = \begin{pmatrix} K & k \\ k^T & k_* \end{pmatrix}$$

where $K \in \mathcal{R}^{\mathcal{N}\times\mathcal{N}}$ is the covariance matrix for $\{Z_i\}_{i=1,\ldots,N}$, $k_* \in \mathcal{R}$ is the variance of Z_* and $k \in \mathcal{R}^{\mathcal{N}}$ the correlation vector between $\{Z_i\}$ and Z_*. The prediction for Z_* is:

$$E[Z_*|z_1, \ldots, z_N] = k^T K^{-1} z$$

where $z = [z_1, \ldots, z_N]^T$. The variance is:

$$Var(Z_*|z) = k_* - k^T K^{-1} k$$

Linear Regression as a particular case of Gaussian Processes:

$$x \xrightarrow{\phi} \phi(x) \xrightarrow{\theta^T} Z(x) \xrightarrow{noise} Y(x)$$

Given $(y_1, \ldots, y_N)$ as data, we want to predict $z(x_*) = z_*$. First define $z_+ = (z_1, \cdots, z_N, z_*)$. The augmented design matrix is:

$$\Phi_+ \triangleq \begin{pmatrix} \Phi \\ \phi_1(x_*), \ldots, \phi_N(x_*) \end{pmatrix} = \begin{pmatrix} \Phi \\ \Phi_* \end{pmatrix}$$

$$Z_+ \sim N(0, \Phi_+\Sigma\Phi_+^T)$$

with

$$\Phi_+\Sigma\Phi_+^T = \begin{pmatrix} \Phi\Sigma\Phi^T & \Phi\Sigma\Phi_*^T \\ \Phi_*\Sigma\Phi^T & \Phi_*\Sigma\Phi_*^T \end{pmatrix}$$

The joint distribution $(Y_1, \ldots, Y_N, Z_*)$ is Gaussian with zero mean and covariance:

$$\begin{pmatrix} \Phi\Sigma\Phi^T + \sigma^2 I & \Phi\Sigma\Phi_*^T \\ \Phi_*\Sigma\Phi^T & \Phi_*\Sigma\Phi_*^T \end{pmatrix}$$

The prediction is:

$$E[Z_*|y_1, \ldots, y_N] = \Phi_*^T\Sigma\Phi^T(\Phi\Sigma\Phi^T + \sigma^2 I)^{-1} y$$

$$Var(Z_*|y_1, \ldots, y_N) = \Phi_*^T\Sigma\Phi_* - \Phi_*^T\Sigma\Phi^T(\Phi\Sigma\Phi^T + \sigma^2 I)^{-1}\Phi\Sigma\Phi_*$$