Gaussian Processes Regression (GPR) -for prediction

April 16, 2019. by Guorui Shen.

Reference : PRML, by M. Bishop. pp. 304 – 309

Conclusion

Given training examples : $(x_1, t_1), \cdots, (x_N, t_N)$,

for prediction purpose :

$$\hat{t}_{N+1} = k^T C_N^{-1} \cdot t \; \cancel{= k(x_1, x_{N+1}), \cdots, k(x_N, x_{N+1})}$$

$$= \left[k(x_1, x_{N+1}), \cdots, k(x_N, x_{N+1})\right] C_N^{-1} \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

where $k(\cdot)$ is kernel function, chosen by your own.

C_N has element : C_N is a square matrix

$$C_N(x_n, x_m) = k(x_n, x_m) + \beta^{-1}\delta_{nm}, \quad n = 1, \cdots, N, \; m = 1, \cdots, N$$

where β^{-1} is variance of noise :

$$t_n = \underset{\substack{\uparrow \\ \text{noise-free}}}{y_n} + \epsilon_n, \quad \epsilon_n \sim N(0, \beta^{-1})$$

Development of GPR

① Consider $y(x) = w^T \phi(x) = (\omega_1, \cdots, \omega_M) \cdot \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_M(x) \end{bmatrix}$

basis function, we don't need to give basis, since we're going to use kernel trick.

② assume $p(w) = N(w \mid 0, \alpha^{-1} I)$ α is hyperparameter.

③ Calculate $p(\vec{y})$: $p(\vec{y}) = N(\vec{y} \mid 0, K)$, proved as follow:
Given $x_1, \cdots, x_N$, by ①, we have

$$\vec{y} = \Phi w, \quad \Phi_{nk} = \cancel{\phi_k(x_n)}\, \phi_k(x_n), \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{bmatrix}$$

then $\vec{y}$ is Gaussian since $\vec{y}$ is a linear combination of Gaussian distributed variables given by elements of w,

$$E(\vec{y}) = \Phi E(w) = 0$$

$$\mathrm{cov}(\vec{y}) = E[\vec{y}\vec{y}^T] = \Phi E(ww^T) \Phi^T = \frac{1}{\alpha}\Phi\Phi^T = K$$

$$K_{nm} = k(x_n, x_m) = \frac{1}{\alpha}\phi(x_n)^T \phi(x_m)$$

④ Consider noise to model : $t_n = y_n + \epsilon_n$, $\epsilon_n \sim N(0, \beta^{-1})$

$$\Rightarrow p(t_n \mid y_n) = N(t_n \mid y_n, \beta^{-1})$$

Given y_n, y_n is now not a random variable.

$$\Rightarrow p(\vec{t} \mid \vec{y}) = N(\vec{t} \mid \vec{y}, \beta^{-1} I_N) \quad \vec{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

⑤ By ③ & ④, we have

$$p(\vec{t}) = \int p(\vec{t} \mid \vec{y}) p(\vec{y}) d\vec{y} = N(\vec{t} \mid 0, C) \quad (*)$$

where C has elements $C(x_n, x_m) = k(x_n, x_m) + \beta^{-1}\delta_{nm}$

⑥ let $\vec{t}_{N+1} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \\ t_{N+1} \end{bmatrix} = \begin{bmatrix} \vec{t} \\ t_{N+1} \end{bmatrix}$ $C_{N+1} = \begin{bmatrix} C_N & k \\ k^T & c \end{bmatrix}$

then $p(\vec{t}_{N+1}) \overset{⑤}{=} N(\vec{t}_{N+1} \mid 0, C_{N+1})$

⑦ Finally $p(t_{N+1} \mid \vec{t})$ is Gaussian and (**)

$$\boxed{m(x_{N+1}) = k^T C_N^{-1} \vec{t} \text{ was used to approximate } \hat{t}_{N+1}}$$

$$\sigma^2(x_{N+1}) = c - k^T C_N^{-1} k$$

Supplementary Information to prove (*) & (**)

⑤' Marginal ~~and~~ and Conditional Gaussians for (*)

If a marginal Gaussian distribution for $\vec{x}$ (vector) and a conditional Gaussian distribution for $\vec{y}$ given x are

$$p(x) = N(x \mid \mu, \Lambda^{-1})$$

$$p(\vec{y} \mid x) = N(\vec{y} \mid Ax + b, L^{-1})$$

then the marginal of $\vec{y}$ and the conditional distribution of x given $\vec{y}$ are

$$p(\vec{y}) = N(\vec{y} \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

$$p(x \mid \vec{y}) = N(x \mid \Sigma\{A^T L(\vec{y} - b) + \Lambda\mu\}, \Sigma)$$

where $\Sigma = (\Lambda + A^T L A)^{-1}$

⑦' Conditional Gaussian Distributions for (**)

let $x = \begin{bmatrix} x_a \\ x_b \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$

assume x is a vector with Gaussian Distribution

$$x \sim N(x \mid \mu, \Sigma)$$

then $p(x_a \mid x_b)$ is Gaussian and

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}.$$