Machine Learning and Neural Networks (MATH3431) Epiphany term, 2023

Handout 8: Gaussian process regression

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce the ideas of learning machines by introducing data into high-dimensional feature spaces for accuracy gains; introduce the kernel trick, and kernel functions.

Reading list & references:

(1) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
- Ch. 6.4 Gaussian process

(2) Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning (Vol. 1, p. 159). Cambridge, MA: MIT press.
- Chapter 2, Regression (supplementary)

(3) Roustant, O., Ginsbourger, D., & Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of statistical software, 51, 1-55.
- Supplementary material related to the the implementation of GP in R computing environment.

1. Intro and motivation

Note 1. As motivation for the Gaussian process regression, we "Kernelize" the standard Bayesian normal linear regression int he machine learning framework.

Note 2. Consider the predictive rule $h(x) = \eta(x)$, and cast it in a linear form $\eta(x) = (\psi(x))^\top w$ where $\psi(x) = (\psi_1(x), ..., \psi_d(x))$ is a vector of basis functions mapping the input space $\mathcal{X}$ into a feature space $\mathcal{F}$. Assume there is available a set of observables $\{z_i = (x_i, y_i)\}_{i=1}^n$. We associate the learning problem with the Bayesian linear regression model

(1.1)
$$\begin{cases} y_i|\psi(x_i), w, \sigma^2 & \overset{\text{ind}}{\sim} \mathrm{N}\left(\eta(x_i), \sigma^2\right),\ i = 1, ..., n \\ \eta(\cdot) & = (\psi(\cdot))^\top w \\ w & \sim \mathrm{N}(\mu_0, V_0) \end{cases} \quad \text{equiv.} \quad \begin{cases} y|\eta, \sigma^2 \sim \mathrm{N}\left(\eta, I\sigma^2\right) & \text{(sampl. distr.)} \\ \eta = \Psi w & \text{(linear model restr.)} \\ w \sim \mathrm{N}(\mu_0, V_0) & \text{(prior)} \end{cases}$$

where $[\Psi]_{i,j} = \psi_j(x_i)$.

Note 3. The marginal likelihood is

$$f(y) = \mathrm{N}\left(y|\Psi^\top\mu_0,\ \Psi V_0 \Psi^\top + I\sigma^2\right) \tag{1.2}$$

where $\mathrm{N}(y|\mu, \Sigma)$ denotes the pdf of the Normal distribution with mean μ,and covariance matrix Σ.

Note 4. The predictive distribution of a new outcome y_* at a new input x_* given the observables $\{z_i = (x_i, y_i)\}_{i=1}^{n}$ is

$$f\left(y_* | x_*, \{(x_i, y_i)\}\right) = \mathrm{N}\left(\mu_*\left(x_*\right), \sigma_*^2\left(x_*\right)\right)$$

with

$$\mu_*\left(x_*\right) = \psi\left(x_*\right)^\top \mu_0 + \frac{1}{\sigma^2} \overbrace{\psi\left(x_*\right)^\top V \Psi}^{K(x_*,X)=} \left(\overbrace{\Psi^\top V \Psi}^{K(X,X)=} + \sigma^2 \right)^{-1} \left(\Psi^\top \mu_0 - y\right) \tag{1.3}$$

$$\sigma_*^2\left(x_*\right) = \underbrace{\psi\left(x_*\right)^\top V \psi\left(x_*\right)}_{=K(x_*,x_*)} - \underbrace{\psi\left(x_*\right)^\top V \Psi}_{=K(x_*,X)} \left(\underbrace{\Psi^\top V \Psi}_{=K(x,X)} + \sigma^2 \right)^{-1} \underbrace{\left(\psi\left(x_*\right)^\top V \Psi\right)^\top}_{=K(X,x_*)} \tag{1.4}$$

according to Proposition 38.

Note 5. In the prior part of (1.1), let's assume $\mu_0 = 0$ (arguably) denoting complete ignorance whether $\eta\left(\cdot\right)$ is positive or negative. By Kernel trick, in (1.3) and (1.4), the feature space always enters in the form inner products. In fact we can define a kernel $K\left(x, x'\right) = \left\langle L\psi\left(x\right), L\psi\left(x'\right)\right\rangle = \psi\left(x\right)^\top V \psi\left(x'\right)$ where L is such that $V = L^\top L$, in terms of Section 4 in Handout 7: Kernel methods. We can denote $K\left(x_*, X\right) = \psi\left(x_*\right)^\top V\Psi$, and $K\left(x_*, x_*\right) = \psi\left(x_*\right)^\top V \psi\left(x_*\right)$.

Note 6. No need to memorize the formulas in (1.2), (1.3), and 1.4. The material in Notes 2, 3, and 4 is given as a motivation for the Gaussian process regression.

2. The Gaussian process regression model

Definition 7. Gaussian process (GP) is a collection of random variables $\{f(x);\, x \in \mathcal{X}\}$, indexed by label x, where any finite collection of those variables has a multivariate normal distribution. It is fully specified by its mean and covariance functions. It is denoted as

$$f(\cdot) \sim \mathrm{GP}(\mu(\cdot), C(\cdot, \cdot))$$

with mean

$$\mu(x) := \mathrm{E}(f(x)),\, x \in \mathcal{X}$$

and covariance function

$$C(x, x') :=\mathrm{Cov}(f(x), f(x')),\, x, x' \in \mathcal{X}$$

Note 8. Essentially, GP is a distribution defined over functions.

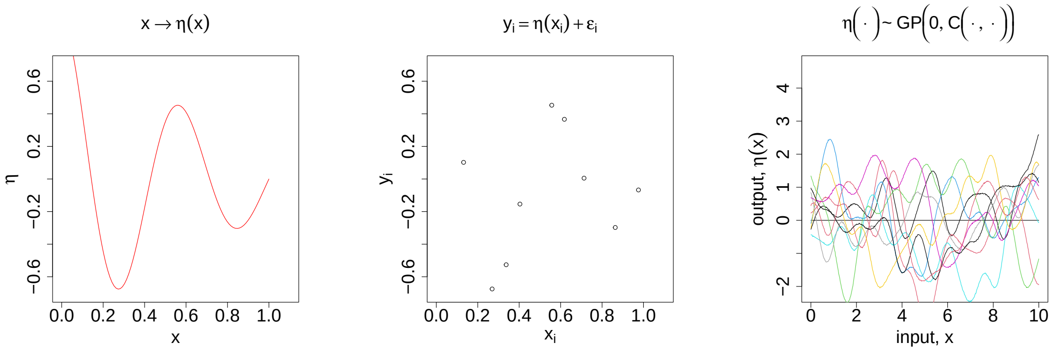
Note 9. Consider a function $\eta : \mathcal{X} \to \mathbb{R}$ with $\eta\left(x\right) = \left\langle\psi\left(x\right), w\right\rangle$ where $\psi\left(x\right)$ is a vector of known basis (feature) functions mapping from the input space $\mathcal{X}$ to the feature space $\mathcal{F}$, and $w \in \mathbb{R}^d$ is an unknown vector a priori following a normal distribution $w \sim \mathrm{N}\left(0, V\right)$, where the prior mean is set to zero denoting complete uncertainty about the sign of w's. Then the marginal $\eta\left(\cdot\right)$ follows a Normal distribution as a linear transformation of Normal variates with mean $\mathrm{E}\left(\eta\left(x\right)\right) = 0$ and covariance $\mathrm{Cov}\left(\eta\left(x\right), \eta\left(x'\right)\right) = \psi\left(x\right)^\top V \psi\left(x'\right)$ for any $x, x' \in \mathcal{X}$. Based on the Kernel trick and

Definition 7, we can equivalently specify $\eta(\cdot) \sim \text{GP}(0, C(\cdot,\cdot))$ for some kernel / covariance function $C(x, x') = \psi(x)^\top V \psi(x')$.

Note 10. We introduce the concept of Gaussian process regression in the machine learning framework below.

Note 11. Consider the predictive rule $h(x) = \eta(x)$, and assume that $\eta : \mathcal{X} \to \mathbb{R}$ with unknown image (possibly up to a set of properties, we will discuss this later) and $\mathcal{X} \subseteq \mathbb{R}^d$.

Example 12. [1]Figure 2.1a shows a function $\eta(x)$ we pretend that we do not know but we wish to recover. To recover $\eta(x)$, we collect training data set as in Figure 2.1b.



(A) True $\eta(x)$

(B) Training Examples $\{z_i = (x_i, y_i)\}$

(C) Realizations of a prior Gaussian process $\eta(\cdot) \sim \text{GP}(0, C(\cdot,\cdot))$

FIGURE 2.1. A toy example

Note 13. For training purposes, assume there is available a set of observables $\{z_i = (x_i, y_i)\}_{i=1}^n$ whose sampling distribution is such that

$$\begin{aligned} & y_i = \eta(x_i) + \epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} \text{N}\left(0, \sigma^2\right),\ i = 1, ..., n \\ & \text{or equivalently} \\ & y_i | \eta(\cdot), \sigma^2 \overset{\text{iid}}{\sim} \text{N}\left(\eta(x_i), \sigma^2\right),\ i = 1, ..., n \end{aligned} \tag{2.1}$$

for some unknown $\sigma^2 > 0$. (2.1) can result by considering, a quadratic loss $\ell(h, z = (x, y)) = \frac{1}{\sigma^2}(h(x) - y)^2$, and sampling distribution with pdf

$$\text{pr}(y | \{x_i, y_i\}) \propto \exp\left(-\sum_{i=1}^n \ell(h(x_i), (x_i, y_i))\right).$$

[1] R code for the plots/running example is available from `https://github.com/georgios-stats/Machine_Learning_and_Neural_Networks_III_Epiphany_2023/tree/main/Lecture_handouts/code/08.Gaussian_process_regression/plots.R`

As $\eta(x)$ is assumed to be unknown, according to the Bayesian paradigm and by taking advantage of Note 9, we assign a GP prior on $\eta(\cdot)$

$$\eta(\cdot) \sim \mathrm{GP}\left(\mu(\cdot|\beta), C(\cdot,\cdot|\phi)\right) \tag{2.2}$$

where μ is parametrized by unknown β (e.g. $\mu(x|\beta) = x^{\top}\beta$), and C is parametrized by unknown ϕ (e.g. $C(x,x'|\phi) = \exp\left(-\frac{1}{2\phi}\|x-x'\|_2^2\right)$; radial/Gaussian kernel). Summing up, the Bayesian model

$$\begin{cases} y_i|\eta(\cdot) & \overset{\text{iid}}{\sim} \mathrm{N}\left(\eta(x_i), \sigma^2\right),\ i = 1,...,n \\ \eta(\cdot) & \sim \mathrm{GP}\left(\mu(\cdot|\beta), C(\cdot,\cdot|\phi)\right) \end{cases}$$

up to some unknown tuning parameters σ^2, β , and $\phi > 0$ which are suppressed from the conditioning to easy the notation.

Note 14. Consider $\eta_* = \eta(X_*)$ where $X_* = (x_{*,1}, x_{*,2}, ..., x_{*,m})^{\top}$ is a vector of new inputs of any length $m > 0$. The joint distribution of $(\eta_*, y)^{\top}$ is

$$\begin{pmatrix} \eta_* \\ y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu(X_*) \\ \mu(X) \end{pmatrix}, \begin{pmatrix} C(X_*,X_*) + I\sigma^2 & C(X_*,X) \\ C(X,X_*) & C(X,X) + I\sigma^2 \end{pmatrix}\right) \tag{2.3}$$

where $C(X,X_*)$ is a Gram matrix over X and X_* such as $[C(X,X_*)]_{i,j} = C(x_i, x_{*,j})$.

Note 15. The conditional distribution of $\eta_* = \eta(X_*)$ given the training sample $\{z_i = (x_i, y_i)\}$, as results from 2.3 (Proposition 38),

$$\eta_*|y \sim \mathrm{N}\left(\mu_*(X_*), C_*(X_*,X_*)\right)$$

is a normal distribution, with mean

$$\mu_*(X_*) = \mathrm{E}(\eta_*|y) = \mu(X_*) + C(X_*,X)\left(C(X,X) + I\sigma^2\right)^{-1}(y - \mu(X)) \tag{2.4}$$

at X_* and with covariance function

$$C_*(X_*,X_*) = \mathrm{Cov}(\eta_*|y) = C(X_*,X)\left(C(X,X) + I\sigma^2\right)^{-1}\left(C(X_*,X)\right)^{\top} \tag{2.5}$$

Note 16. Because X_* is of any finite length, and the derivations in Note 15, by definition of GP, the predictive distribution of $\eta(\cdot)$ given the data $\{z_i = (x_i, y_i)\}$ is the Gaussian process

$$\eta(\cdot) \sim \mathrm{GP}\left(\mu_*(\cdot), C_*(\cdot,\cdot)\right) \tag{2.6}$$

with mean function and covariance function

$$\mu_*(x_*) = \mu(x_*) + C(x_*,X)\left(C(X,X) + I\sigma^2\right)^{-1}(y - \mu(X)) \tag{2.7}$$

$$C_*\left(x_*, x_*'\right) = C(x_*,X)\left(C(X,X) + I\sigma^2\right)^{-1}C\left(X, x_*'\right) \tag{2.8}$$

for any points $x_*, x_*' \in \mathcal{X}$. If I consider $X_* = (x_*, x_x')^{\top}$, (2.7) results as the first block of 2.4, and (2.8) results as the top off-diagonal block of 2.5.

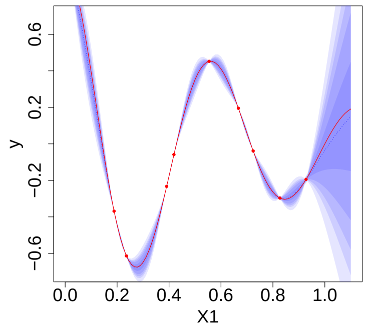


Figure 2.2. Predictive GP

Note 17. Note that the posterior expected rule at $x_{*}\in\mathcal{X}$ is

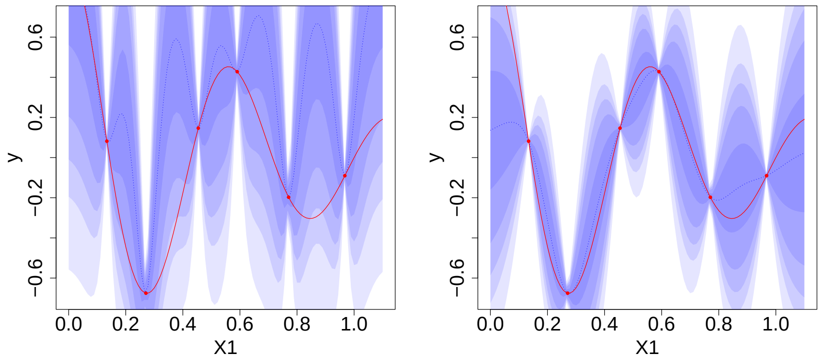
$$\mathrm{E}\left(h\left(x_{*}\right)|y\right)=\mathrm{E}\left(\eta\left(x_{*}\right)|y\right)=\mu\left(x_{*}\right)+C\left(x_{*},X\right)\left(C\left(X,X\right)+I\sigma^{2}\right)^{-1}\left(y-\mu\left(X\right)\right)\tag{2.9}$$
$$=\sum_{i=1}^{n}\alpha_{i}C\left(x_{i},x_{*}\right)$$

where $\alpha=\mu\left(x_{*}\right)+\left(y-\mu\left(X\right)\right)\left(C\left(X,X+I\sigma^{2}\right)\right)^{-1}$. This is in accordance to the Representation theorem (Theorem 21 in Handout 7 Kernel methods) with reference to the Bayesian linear regression (Note 2).

Example 18. Figure 2.2 presents the predictive distribution from a trained GP regression

3. Training (via Empirical Bayes)

Note 19. Recall that the mean and covariance functions in (2.6) depend on tunable parameters σ^{2}, ϕ, and β. When the number of training examples is small, the behavior of (2.6) is sensitive to these hyperparameters. In Figure 3.1, there are two instances of GP regression given 6 examples where the tunable parameters are different.



(A) Careless choice of ϕ Here $\sigma^2 = 0$, and $\beta = 0$.

(B) Wise choice of ϕ Here $\sigma^2 = 0$, and $\beta = 0$.

FIGURE 3.1. Sensitivity in the tunable parameters ϕ. Here $\sigma^2 = 0$, and $\beta = 0$.

Note 20. The marginal likelihood $f\left(y|\sigma^2,\phi,\beta\right)$ of y given the known parameters σ^2,ϕ,β results from (2.3) as

$$y|\sigma^2,\phi,\beta \sim \mathrm{N}\left(\mu\left(X|\beta\right), C\left(X,X|\phi\right)+I\sigma^2\right).$$

Note 21. Let $\mu_\beta = \mu\left(X|\beta\right)$ and $C_\phi = C\left(X,X|\phi\right)$. To learn the unknown hyper-parameters $\theta = \left(\sigma^2,\phi,\beta\right)$ according to Empirical Bayes procedure as well as according to classical methods using as error function the marginal likelihood, we need to minimize

$$\left(\hat{\sigma}^2,\hat{\phi},\hat{\beta}\right) = \arg\min_{\sigma^2,\phi,\beta}\left(-2\log\left(\mathrm{N}\left(y\,|\,\mu_\beta\,,\,C_\phi+I\sigma^2\right)\right)\right) \tag{3.1}$$

$$= \arg\min_{\sigma^2,\phi,\beta}\left(\underbrace{\log\left(\left|C_\phi+I\sigma^2\right|\right)+\left(y-\mu_\beta\right)^\top\left(C_\phi+I\sigma^2\right)^{-1}\left(y-\mu_\beta\right)}_{\hat{R}(\sigma^2,\phi,\beta)}\right).$$

Note 22. (3.1) can be solved via GD (Algorithm 1 Handout 2 Gradient descent), or the Stochastic Gradient (Handout 3: Stochastic gradient descent) as the required gradient ca be easily computed as

No need t
memorized

$$\frac{\mathrm{d}R}{\mathrm{d}\beta_j} = \left(C_\phi+I\sigma^2\right)^{-1}\left(y-\mu_\beta\right)\frac{\mathrm{d}\mu_\beta}{\mathrm{d}\beta_j}$$

$$\frac{\mathrm{d}R}{\mathrm{d}\phi_j} = \mathrm{tr}\left(\left(C_\phi+I\sigma^2\right)^{-1}\left[\frac{\partial C_\phi}{\partial\phi_j}\right]\right)+\left(y-\mu\right)^\top\left(C_\phi+I\sigma^2\right)^{-1}\left[\frac{\partial C_\phi}{\partial\phi_j}\right]\left(C_\phi+I\sigma^2\right)^{-1}\left(y-\mu\right)$$

$$\frac{\mathrm{d}R}{\mathrm{d}\sigma^2} = \mathrm{tr}\left(\left(C_\phi+I\sigma^2\right)^{-1}\right)+\left(y-\mu\right)^\top\left(C_\phi+I\sigma^2\right)^{-1}\left(C_\phi+I\sigma^2\right)^{-1}\left(y-\mu\right)$$

4. Examples of Examples of Covariance Functions

Note 23. The covariance function $C : \mathcal{X}\times\mathcal{X}\rightarrow[0,\infty)$ with $C\left(x,x'\right)$ describes how much two random variables x,x' change together.

Note 24. Covariance function is a functional parameter of the GP prior (2.2). Different covarince functions represent different properties, hence they impose different prior info in the GP regression model; they are crusial parameters.

Note 25. Any positive definite Kernel as described in Section 4 in Handout 7 Kernel methods can be used as a covarince function. Consequently kernel construction approached and theories introduced can be use for the covariance functions as well.

Definition 26. Stationary covariance function is called a covariance function $C:\mathcal{X}\times\mathcal{X}\rightarrow[0,\infty)$ whose image can be written as $C\left(x,x'\right)=C\left(\|x-x'\|\right)$ namely, the dependence between any pair of input points x , x' is a function of their distance and only.

Note 27. a one-dimensional Gaussian process one way to understand the characteristic length-scale of the process (if this exists) is in terms of the number of upcrossings of a level u. The expected number of upcrossings $\mathrm{E}\left(N_u\right)$ of the level u on the unit interval by a zero-mean, stationary, is

$$\mathrm{E}\left(N_u\right)=\frac{1}{2\pi}\sqrt{\frac{-K''\left(0\right)}{K'\left(0\right)}}\exp\left(-\frac{u^2}{2K\left(0\right)}\right)$$

Note 28. Popular covariance functions are

Gaussian covariance function: given as

$$C\left(r\right)=\exp\left(-\frac{1}{2\phi^2}r^2\right)$$

- It is infinitely differentiable, which means that the GP is very smooth.
- The parameter ϕ is called lengthscale.
- The number of upcrossing at level u is $\mathrm{E}\left(N_u\right)=\left(2\phi^2\right)^{-1}$ meaning that smaller ϕ represents more upcrossings, hence represents smaller scale dependences

Exponential Covariance Function: given as

$$C\left(r\right)=\exp\left(-\frac{1}{\phi}\left|r\right|\right)$$

- It is not differentiable at $r=0$

Matern Class of Covariance Functions: given as

$$C_{\nu}\left(r\right)=\frac{2^{1-\nu}}{\Gamma\left(\nu\right)}\left(\frac{\sqrt{2\nu}r}{\phi}\right)^{\nu}B_{\nu}\left(\frac{\sqrt{2\nu}r}{\phi}\right)$$

where $B_{\nu}\left(\cdot\right)$ is a modified Bessel functions (description of Bessel functions is out of the scope). Matern covariance function gives the Exponential one for $\nu=1/2$, and the exponential one for $\nu\rightarrow\infty$.

Note 29. Anisotropic versions of these isotropic covariance functions can be created anisotropy by setting $r\left(x,x'\right)=\left(x-x'\right)^{\top}M\left(x-x'\right)$ for some positive semi-definite matrix M. If M is diagonal this implements the use of different length-scales on different dimension of inputs. Off-diagonal elements of M implement cross-dimensional dependencies in the inputs.

5. Practical maters

Note 30. One may consider some low degree polynomial form $\mu(x|\beta) = \sum_{j=0}^{p} x^j \beta_j$, however in this case $\mu(\cdot|\beta)$ and $C(\cdot,\cdot|\phi)$ may compete as $C(\cdot,\cdot|\phi)$ can express such behaviors according to Note 9. For this reason, the usual specification of $\mu(\cdot|\beta)$ in (2.2) is $\mu(x|\beta) = 0$ implying a priori complete uncertainty about the sign of $\eta(x)$ at each x.

Note 31. Thinking as a statistician, one may decompose (2.2) as

$$\eta(\cdot) = \mu(\cdot|\beta) + \xi(\cdot|\phi)$$

where $\mu(\cdot|\beta)$ is modeled as a low degree polynomial (e.g., 2nd degree) representing large scale dependences (see polynomial regression in Term 1), and $\xi(\cdot|\phi) \sim \text{GP}(0, C(\cdot,\cdot|\phi))$ representing lower scale dependence by using an appropriate kernel. Seeing the big picture, (2.1) is re-states as $y_i = \mu(x_i|\beta) + \xi(x_i|\phi) + \epsilon_i$ where ϵ_i represents noise (or so short scale dependence that can be considered as noise in the model).

6. Practice, implementation, and code

Below is some practical examples on the implementation of Gaussian process regression in R programming environment by using the R packages: DiceKriging and DiceOptim.

- Roustant, O., Ginsbourger, D., & Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of statistical software, 51, 1-55.

The examples below were created for the undergraduate programme SURF 2016 at Purdue University (July 8, 2016) however they are suitable to the course.

Toy example
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Numerical_example.ipynb`

Realistic example: The Piston Simulation function model in 2D
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_2D.ipynb`

Practice Catalytic Reaction 5D
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_CatalyticReaction_5D_solution.ipynb`

Practice Piston 7D
`https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_Piston_7D_solution.ipynb`

Practice Robot Arm 8D

https://github.com/georgios-stats/Intro_GPR_SURF_2016/blob/master/Practice_Robot_Arm_8D_solution.ipynbb

Appendix A. Multivariate Normal distribution $x|\mu,\Sigma \sim \mathrm{N}_d(\mu,\Sigma)$[2]

Definition 32. A d-dimensional random variable $x \in \mathbb{R}^d$ is said to have a multivariate Normal (Gaussian) distribution, if for every d-dimensional fixed vector $\alpha \in \mathbb{R}^d$, the random variable $\alpha^\top x$ has a univariate Normal (Gaussian) distribution.

Definition 33. We denote the d-dimensional Normal distribution with mean μ and covariance matrix $\Sigma \geq 0$ as $\mathrm{N}_d(\mu,\Sigma)$.

Notation 34. The d-dimensional standardized Normal distribution is $\mathrm{N}_d(0,I)$.

Proposition 35. *Let random variable $x \sim N_d(\mu,\Sigma)$,* fixed vector $c \in \mathbb{R}^q$ and fixed matrix $A \in \mathbb{R}^q \times \mathbb{R}^d$. The random vector $y = c + Ax$ *has distribution* $y \sim N_q\left(c + A\mu, A\Sigma A^\top\right)$.

Proposition 36. *Let a d-dimensional random vector $x \sim N_{(any)}(\mu,\Sigma)$.*

(1) *Let $y = Ax$ and $z = Bx$, where $A \in \mathbb{R}^{q\times d}$ and $B \in \mathbb{R}^{k\times d}$: The vectors $y = Ax$ and $z = Bx$ are independent if and only if $A\Sigma B^\top = 0$.*

(2) *Let $x = (x_1,...,x_d)^\top$: The $x_1,...,x_d$ are mutually independent if and only if the corresponding off diagonal parts of the Σ are zero.*

Proposition 37. *Any sub-vector of a vector with multivariate Normal distribution has a multivariate Normal distribution.*

Proposition 38. *[Marginalization & conditioning] Let $x \sim \mathrm{N}_d(\mu,\Sigma)$.* Consider partition such that

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \qquad \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix},$$

where $x_1 \in \mathbb{R}^{d_1}$, and $x_2 \in \mathbb{R}^{d_2}$ Then:

(1) For the marginal, it is $x_1 \sim \mathrm{N}_{d_1}(\mu_1,\Sigma_1)$.

(2) For the conditional, if $\Sigma_1 > 0$, it is

$$x_2|x_1 \sim \mathrm{N}_{d_2}(\mu_{2|1},\Sigma_{2|1})$$

where

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_1^{-1}(x_1 - \mu_1) \text{ and } \Sigma_{2|1} = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top \tag{A.1}$$

Proposition 39. *The density function of the d-dimensional Normal distribution with mean μ and covariance matrix Σ, when $\underline{\Sigma \text{ is symmetric positive definite matrix } (\Sigma > 0})$, exists and it is equal to*

$$f(x) = (2\pi)^{-\frac{d}{2}}\det(\Sigma)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right) \tag{A.2}$$

[2] More detailed material about the Multivariate Normal distribution can be found in the can be found in “Handout 2: Revision in mixture of probability distributions” of the module “Bayesian Statistics III/IV (MATH3341/4031)” Michaelmas term, 2021 available from `https://github.com/georgios-stats/Bayesian_Statistics_Michaelmas_2021/blob/main/Lecture_handouts/02_Revision_in_mixture_of_probability_distributions.pdf`. The material in this section is just a sub-set of the statements in the referenced handout.