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

Draft Handout 7: Kernel methods

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 (supplamentary)

1. Intro and motivation

Note 1. Consider the predictive rule $h(x) = \eta(x)$. 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 2. 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 3. 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

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

with

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

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

according to Proposition 22.

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

Note 5. No need to memorize the formulas in (1.2), (1.3), and 1.4. The material in Notes 1, 2, and 3 is given as a motivation for the Gaussian process regression.

2. The Gaussian process regression model

Definition 6. 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 denotes 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 7. Essentially, GP is a distribution defined over functions.

Note 8. Consider a function $\eta : \mathcal{X} \to \mathbb{R}$ with $\eta(x) = \langle \psi(x), w\rangle$ where $\psi(x)$ is a vector of known bases (feature) functions mapping from the input space tot he feature space $\mathcal{F}$, and $w \in \mathbb{R}^d$ is an unknown vector a priori following a normal distribution $w \sim \mathrm{N}(0,V)$, where the prior mean is set to zero denoting complete uncertainty about the sign of w's. Then the marginal $\eta(\cdot)$ follows a Normal distribution as a linear transformation of Normal variates with mean $\mathrm{E}(\eta(x)) = 0$ and covariance $\mathrm{Cov}(\eta(x), \eta(x')) = \psi(x)^\top V\psi(x')$ for any $x,x' \in \mathcal{X}$. Based on the Kernel trick and Definition 6, we can equivalently specify $\eta(\cdot) \sim \mathrm{GP}(0, C(\cdot,\cdot))$ for some kernel / covariance function $C(x,x') = \psi(x)^\top V\psi(x')$.

Note 9. We introduce the concept of Gaussian process regression in the machine learning framework below.

Note 10. 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 $\mathcal{X} \subseteq \mathbb{R}^d$.

Note 11. 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

$$y_i = \eta(x_i) + \epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma^2\right),\ i = 1, ..., n \tag{2.1}$$

$$y_i | \eta(\cdot), \sigma^2 \overset{\text{iid}}{\sim} \mathrm{N}\left(\eta(x_i), \sigma^2\right),\ i = 1, ..., n$$

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

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

As we $\eta(x)$ is assumed to be unknown, according to the Bayesian paradigm and the observation in Note 8, we assign a GP prior on $\eta(\cdot)$

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

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

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

where the unknown tuning parameters σ^2, β , and $\phi > 0$ are suppressed from the conditioning.

Note 12. Consider $\eta_* = \eta(X_*)$ and $\eta_{**} = \eta(X_{**})$ where X_*, X_{**} are vectors of any length of new inputs. The joint distribution of $(\eta_*, \eta_{**}, y)^\top$ is

$$\begin{pmatrix} \eta_* \\ \eta_{**} \\ y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C(X_*, X_*) + \sigma^2 & C^\top(X_{**}, X_*) & C^\top(X, X_*) \\ C(X_{**}, X_*) & C(X_{**}, X_{**}) + \sigma^2 & C^\top(X, X_{**}) \\ C(X, X_*) & C(X, X_{**}) & C(X, X) + \sigma^2 \end{pmatrix}\right) \tag{2.2}$$

where C is a Gram matrix such as $[C(X, X)]_{i,j} = C(x_i, x_j)$.

Note 13. Consider X_*, X_{**} as vectors of any length of new inputs. The conditional distribution of $\eta_* = \eta(X_*)$ given the training sample $\{z_i = (x_i, y_i)\}$, aka $\eta(X_*)|y$ as results from 2.2 (Proposition 22) is a normal distribution, with mean

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

at X_* and with covariance function

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

Note 14. Given that the derivations in Note 13 is given X_*, X_{**} input vectors of any length, we can say that the predictive distribution of $\eta(\cdot)$ given the data $\{z_i = (x_i, y_i)\}$ is the Gaussian process

$$\eta(\cdot) \sim \mathrm{GP}\left(\mu_*(\cdot), C_*(\cdot,\cdot)\right) \tag{2.3}$$

with mean function and covariance function

$$\begin{aligned} \mu_*(x_*) &= \mu(x_*) + C(x_*, X)\left(C(X,X)\right)^{-1}(y - \mu(X)) \\ C_*(x_*, x_{**}) &= C(x_*, X)\left(C(X,X)\right)^{-1} C(X, x_{**}) \end{aligned}$$

at any new points x_*, and x_{**}.

Note 15. Note that

$$\begin{aligned} \mathrm{E}\left(h(x_*)\,|y\right) &= \mu(x_*) + C(x_*, X)\left(C(X,X)\right)^{-1}(y - \mu(X)) \\ &= \sum_{i=1}^{n} \alpha_i C(x_i, x_*) \end{aligned}$$

3. Training

Recall that the mean and covariance functions in (2.3) depend on tunable parameters σ^2, ϕ, and β. When the number of training examples is small, the behavior of (2.3) is sensitive to these hyperparameters.

Appendix A. Multivariate Normal distribution $x|\mu, \Sigma \sim \mathrm{N}_d(\mu, \Sigma)$[1]

Definition 16. 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 17. We denote the d-dimensional Normal distribution with mean μ and covariance matrix $\Sigma \geq 0$ as $\mathrm{N}_d(\mu, \Sigma)$.

Notation 18. The d-dimensional standardized Normal distribution is $\mathrm{N}_d(0, I)$.

Proposition 19. *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 20. *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 21. *Any sub-vector of a vector with multivariate Normal distribution has a multivariate Normal distribution.*

Proposition 22. *[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 23. *The density function of the d-dimensional Normal distribution with mean* μ *and covariance matrix* Σ*, when* Σ *is symmetric positive definite matrix* ($\Sigma > 0$)*, exists and it is equal to*

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

[1] 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.