Prof. Dr. Volker Roth
volker.roth@unibas.ch

Vitali Nesterov
vitali.nesterov@unibas.ch

Department of Mathematics and Computer Science
Spiegelgasse 1
4051 Basel

Exercise 10

Due date: **Wednesday, May 13$^{\text{th}}$ 2020**

10.1: Gaussian processes

Recall the Bayesian linear regression, the amount of model parameters depends on the dimensionality of the input expansion, for instance, the amount of polynomial basis functions. A Gaussian process is a non-parametric model. The non-linearity of the model is achieved by using a Gaussian kernel function, sucht that further expansion of the input is not necessary. Assume D-dimensional data $(x_1, x_2, \ldots, x_D) \in \mathbb{R}^D$, such that all observations can be written in a matrix form with $\mathbf{X} \in \mathbb{R}^{N \times D}$ and the corresponding labels $\mathbf{y} \in \mathbb{R}^N$. Accordingly, unobserved data is modeled with $\mathbf{X}_* \in \mathbb{R}^{N_* \times D}$ and the predictions $\mathbf{f}_* \in \mathbb{R}^{N_*}$.

Definitions

- A **gaussian process** as a prior on a regression function is

$$f(\boldsymbol{x}) \sim GP\left(m(\boldsymbol{x}), \kappa\left(\boldsymbol{x}, \boldsymbol{x}'\right)\right),$$

  where $m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$ is the mean function and $\kappa\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{x}')$ is a kernel or a covariance function.

- The **noisy observations** are modeled as follows:

$$\boldsymbol{y} \sim \mathcal{N}\left(f(\boldsymbol{x}), \mathbf{K} + \sigma_y^2 \mathbf{I}_N\right),$$

  such that the fluctuation of the data is modeled with the kernel $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$ of the gaussian process and a diagonal covariance matrix $\sigma_y^2 \mathbf{I}_N$ for the observation noise.

- The **squared exponential kernel** is defined as follows:

$$\kappa\left(x, x'\right) = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}\left(x - x'\right)^2\right),$$

  where l is the horizontal and σ_f is the vertical variation factors.

- The **joint distribution** of a gaussian process is defined as follows:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^\top & \mathbf{K}_{**} \end{pmatrix}\right)$$

where the mean is a zero vector. The covariance matrix is partitioned with $\mathbf{K}_y \triangleq \mathbf{K} + \sigma_y^2 \mathbf{I}_N \in \mathbb{R}^{N\times N}$, $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{N\times N}$, $\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*) \in \mathbb{R}^{N\times N_*}$, and $\mathbf{K}_{**} = \kappa\left(\mathbf{X}_*, \mathbf{X}_*\right) \in \mathbb{R}^{N_*\times N_*}$.

- The **posterior predictive density** for noisy observations is given with:

$$P\left(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{y}\right) = \mathcal{N}\left(\mathbf{f}_* | \mathbf{K}_*^\top \mathbf{K}_y^{-1} \mathbf{y}, \mathbf{K}_{**} - \mathbf{K}_*^\top \mathbf{K}_y^{-1} \mathbf{K}_*\right).$$

Exercise

Derive the posterior predictive mean and covariance given the joint distribution $P(\mathbf{y}, \mathbf{f}_*)$. Use the Schur complement to invert a partitioned matrix. What is the advantage of a Gaussian kernel compared with a linear kernel?

10.2: Python implementation

In the above derived results, the posterior predictive density is given. Write a Python program for a regression task with noisy observations.

- Implement the squared exponential kernel function.
- Compute the posterior predictive parameters. Use the numpy package to compute an inverse of a matrix.
- Sample two functions from the posterior predictive density. Use a multivariate normal distribution.
- Plot the mean estimate, sampled regression functions, the uncertainty boundaries, and the observed data points.