Optimization Algorithms
Weekly Exercises 10

Marc Toussaint
Learning & Intelligent Systems Lab, TU Berlin
Marchstr. 23, 10587 Berlin, Germany

Winter 2024/25

1 Gaussian Process Regression

In the lecture we mentioned Gaussian Processes (GP) as a basic approach to formulate a distribution $P(f|D)$ over continuous functions $f : \mathbb{R}^d \to \mathbb{R}$, given data D. Slide 9 of the lecture summarizes the essential equations; the standard reference for GPs is Rasmussen & Williams (2006) [pdf link]. In this exercise you learn about them by implementing a minimalistic case:
You are given a $D = \{(x_i, y_i)\}_{i=1}^n$. In this exercise, we assume $x \in \mathbb{R}$ (1-dimensional) and we just have $n = 2$ data points $(x_1 = 0, y_1 = 0)$ and $(x_2 = 1, y_2 = 1)$. Then compute the following:

a) Compute the kernel matrix $K \in \mathbb{R}^{n\times n}$ with entries

$$K_{ij} = k(x_i, x_j) \ , \quad k(x, x') = a \ \exp(-\frac{1}{2}\|x - x'\|^2/\ell^2) \ .$$

We choose $a = 1, \ell = 1$, and $k(x, x')$ is called squared exponential covariance function.

[This matrix desribes how correlated the observations at all data points x_i are.]

b) Trivially also prepare the data vector $Y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$.

c) Write a method, that for any new $x \in \mathbb{R}$ computes a vector $\kappa(x) \in \mathbb{R}^2$, a prediction $\mu(x) \in \mathbb{R}$, and a variance $\sigma^2(x) \in \mathbb{R}$ as follows

$$\kappa(x) \in \mathbb{R}^n \text{ with entries } \kappa_i(x) = k(x, x_i) \tag{1}$$
$$\mu(x) = \kappa(x)^\top \ (K + \sigma_0^2\mathbf{I})^{-1} \ Y \tag{2}$$
$$\sigma^2(x) = k(x, x) - \kappa(x)^\top \ (K + \sigma_0^2\mathbf{I})^{-1} \ \kappa(x) \ , \tag{3}$$

where $\mathbf{I}$ is the identity matrix and we choose observation noise $\sigma_0 = 0.1$.

[The vector $\kappa(x)$ describes how correlated a new observation at x should be with observations at all data points x_i. The prediction $\mu(x)$ and variance $\sigma^2(x)$ can be derived as the conditional marginal of a joint Gaussian distribution.]

d) Now sample $x \in [-2, 2]$ on a fine grid, compute $\mu(x)$ and $\sigma^2(x)$ for each x, and use this to plot the functions $\mu(x)$, $\mu(x) + \sqrt{\sigma^2(x)}$, and $\mu(x) - \sqrt{\sigma^2(x)}$ for the interval $x \in [-2, 2]$.

How does this change for $\sigma = 0$? How does this change for $\ell = 0.1$? How does this change with more observed points (e.g., sample them from the prediction, then consider them observed data points)?

2 Global Optimization in high dimensions?

Assume you have a GP prior $P(f)$ over functions $[0, 1]^n \to \mathbb{R}$ and search a global minimum in the bounded space $[0, 1]^n \subset \mathbb{R}^n$. We have a squared exponential kernel with length scale (kernel width) $\ell = 0.1$, i.e.,

$$k(x, x') = \exp(-\frac{\|x - x'\|^2}{2 \cdot 0.1^2}).$$

For simplicity, let us assume that all observations (whereever we query) turn out zero and we collect data $D = \{(x_i, y_i)\}_{i=1}^T$ with $y_i = 0$.

Estimate the number T of points you need to query to achieve some certainty that no function value of the true f is larger than 1. For instance, determine a T and a querying scheme that defines all x_i, so that $\forall x: \; P(f(x) > 1) \leq 0.0227$. (The last number is the probability that a random number from the standard normal distribution $z \sim \mathcal{N}(0, 1)$ is larger than 2. See https://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg)

Note: The following paper summarizes results on the Euclidean distance with increasing space dimensionality:

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001, January). On the surprising behavior of distance metrics in high dimensional space. In International conference on database theory (pp. 420-434). Springer, Berlin, Heidelberg.

E.g., stated overly briefly, with $n \to \infty$ the ratio of distances to a nearest and furthest random point converges to 1.