Exercises
Advanced Machine Learning
Fall 2020

Series 3. Oct 19th 2020 (Gaussian Processes and Newton's Method)

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Teaching assistant: **Luca Corinzia**
luca.corinzia@inf.ethz.ch

Problem 1 (Gaussian Process Kernels):

The following exercises are related to the previous topic of Gaussian Processes.
Part 1) is dedicated to students new to kernels, however is a good practice to everyone.

1) Given valid kernels $k_1(\boldsymbol{x}, \boldsymbol{x}')$ and $k_2(\boldsymbol{x}, \boldsymbol{x}')$, show that the following new kernels are also valid:

   (a) $k(\boldsymbol{x}, \boldsymbol{x}') = ck_1(\boldsymbol{x}, \boldsymbol{x}')$, with constant $c > 0$;

   (b) $k(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})k_1(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}')$, with any function $f(\cdot)$;

   (c) $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}') + k_2(\boldsymbol{x}, \boldsymbol{x}')$;

   (d) $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}')k_2(\boldsymbol{x}, \boldsymbol{x}')$;

2) We investigate the influence of kernels and parameters on the prior distribution of functions. Decide which kernel is used in the following examples and match the parameters used to generate the resulting covariance matrices and the functions sampled from the corresponding prior distributions.

   Kernels:

   (a) RBF kernel: $\sigma^2 exp\left(-\frac{||t-t'||^2}{2l^2}\right)$

   (b) periodic kernel: $\sigma^2 exp\left(-\frac{2sin^2(\pi|t-t'|/p)}{l^2}\right)$

   Parameter settings:

   (a) $\sigma = 0.8$, $l = 0.5$, $(p = 0.5)$

   (b) $\sigma = 0.8$, $l = 2$, $(p = 0.5)$

   (c) $\sigma = 0.33$, $l = 0.5$, $(p = 0.5)$

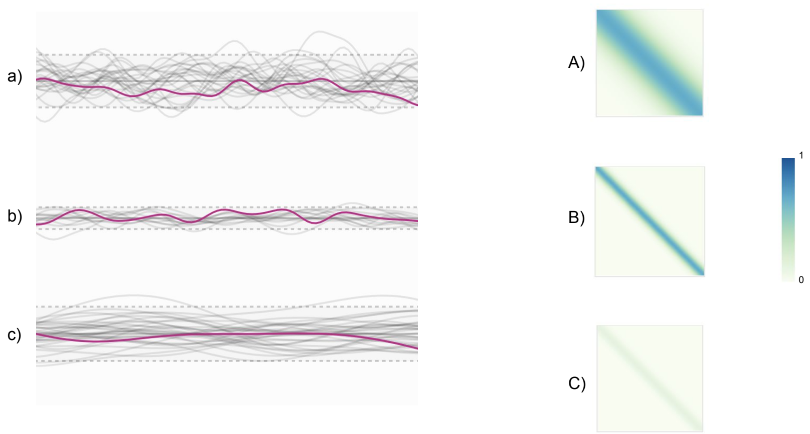

Figure 1: Samples from prior distributions and the corresponding covariance matrices.
Source: `https://www.jgoertler.com/visual-exploration-gaussian-processes/`.

3) We could see in 1) that combining kernels through addition results in a valid kernel. For the following examples from posterior distribution decide which kernels (RBF, periodic, linear) were combined.

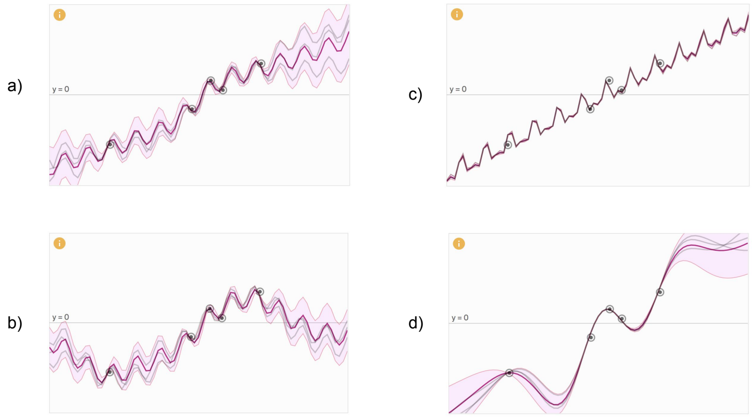

Figure 2: Samples from posterior distribution obtained using combinations of different kernels.
Source: `https://www.jgoertler.com/visual-exploration-gaussian-processes/`.

Problem 2 (Newton Method):

Newton's method was originally created to find a root $f(x^*) = 0$ of a given function $f(x)$ via the iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{1}$$

where $f'(x)$ denotes the derivative of f with respect to x. In optimization we are usually interested in finding the minimum of a function. This can be achieved using Newton's method to find a root of the first derivative $f'(x^*) = 0$, the optimization step reads then.

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}. \tag{2}$$

However, to keep the notation simple we will use equation (1) and focus on one dimensional functions $f : \mathbb{R} \mapsto \mathbb{R}$. In this exercise we will examine the convergence of Newton's method and some common pitfalls.

1. **Complete Failure:** Show that Newton's method will never converge for $f(x) = \sqrt[3]{x}$ and $x_0 \neq 0$ (we define the third root of negative numbers $x < 0$ as $\sqrt[3]{x} = -\sqrt[3]{|x|}$). Why does it fail?
2. **Convergence for simple roots:** A sequence x_n converges with order m towards x^*, if there exists a constant C, such that $|x_{n+1} - x^*| \leq C|x_n - x^*|^m$ (as $n \to \infty$). Show that, if Newton's method converges, it will converge quadratic ($m = 2$) for a smooth function f with one root $f(x^*) = 0$ and $f'(x^*) \neq 0$ in a region around x^*.

   Hint: Use the Taylor expansion of f to second order around the point x_n.
3. **Convergence for higher order roots:** A root $f(x^*) = 0$ has order k if all derivatives $f^{(i)}(x^*) = 0$ vanish for $i < k$ and $f^{(k)}(x^*) \neq 0$.

(a) Show that Newton's method converges linear ($m = 1$) for a smooth function f with one root $f(x^*) = 0$ of order $k > 1$ in a region around the point x^*.

Hint: Write f as $f(x) = (x - x^*)^k g(x)$ with $g(x^*) \neq 0$ and use the Taylor approximation.

(b) How should we adapt equation (1) to achieve quadratic convergence?

(c) What implications does this have for an optimization problem where we want to minimize a function f by finding a root of the derivative f'?

Problem 3 (Gradient Descent):

In gradient descent, we iteratively estimate $\min_w f(w)$ by computing a sequence of estimates $w^{(0)}, \ldots, w^{(k)}, \ldots$. Each estimate $w^{(k+1)}$ is obtained from the previous by adding an update η_k:

$$w^{(k+1)} \leftarrow w^{(k)} + \eta_k.$$

Assume that f's Hessian is invertible when evaluated at any point. Using a second-order Taylor approximation, demonstrate that the optimal update is

$$\eta_k^* = -H_f^{-1}(w^{(k)}) \nabla f(w^{(k)}).$$