COMP 9501: Machine Learning Spring 2019

Assignment 3

:

1 Problem 1

In a typical regression problem we have labels $\mathbf{y} = y_1, \cdots, y_n$ of the function at points $\mathbf{X} = \{x_1, \cdots, x_n\}$. Here $y_i = y(x_i) \in \mathcal{R}$, $x_i = [x_i^{(1)}, \cdots, x_i^{(d)}] \in \mathcal{R}^d$. We are interested in estimating the value of the function $y_* = y(x_*)$ at a new point x_*. Alternatively, we may also be interested in the gradient and integral of of the function at x_*, GPs provide a nice framework for obtaining an estimate for the gradient and integral, as well as quantifying the uncertainty.

1.1 Gradients using Gaussian Processes

Recall the gradient at x_* is a d-vector containing the d partial derivatives. Denote the partial derivatives of y as $g_i(x) = \frac{\partial y(x)}{\partial x^{(i)}}$, where $g_i : \mathcal{R}^d \to R$, $i = 1, \cdots, d$. The gradient is $\mathbf{g}(x) = [g_1(x), \cdots, g_d(x)]^T$. Write $g_{i*} = g_i(x_*)$ and $\mathbf{g}_* = g(x_*)$.

We will model our observation as a Gaussian Process $y(x)$ over $\mathcal{R}^d$ with zero mean and auto covariance K. We know that if a function y is sampled from a GP, the distribution of $y_1, \cdots, y_n, y_*$ at $x_1, \cdots, x_n, x_*$ follow a Gaussian distribution. It can also be shown that the gradient $\mathbf{g}_*$ at x_* is also a Gaussian.

Further, the gradient $\mathbf{g}_*$ and the function values $y_1, \cdots, y_n$ are jointly Gaussian, i.e.

$$[\mathbf{y}, \mathbf{g}_*] = [y_1, \cdots, y_n, g_1(x_*), \cdots, g_d(x_*)] \in \mathcal{R}^{n+d}$$

is also a Gaussian. This Gaussian has zero mean. The covariance K_i between $y(x_1)$ and $g_i(x_2)$ and the covariance K_{ij} between $g_i(x_1)$ and $g_i(x_2)$ are given respectively by

$$K_i(x_1, x_2) = \frac{\partial K(x_1, x_2)}{\partial x_2^{(i)}}$$
$$K_{ij}(x_1, x_2) = \frac{\partial^2 K(x_1, x_2)}{\partial x_1^{(i)} \partial x_2^{(j)}}$$

a) Let $\mathbf{K} \in \mathcal{R}^{n \times n}$, $\mathbf{J} \in \mathcal{R}^{n \times d}$, and $\mathbf{B} \in \mathcal{R}^{d \times d}$ such that

$$\mathbf{K}_{ij} = K(x_i, x_j), \quad \mathbf{J}_{ij} = \frac{\partial K(x_i, x_*)}{\partial x_*^{(j)}}, \quad \mathbf{B}_{ij} = \frac{\partial^2 K(x_*, x_*)}{\partial x_*^{(i)} \partial x_*^{(j)}}.$$

Write the (Gaussian) prior over $[\mathbf{y}, \mathbf{g}_*]$ in terms of $\mathbf{K}, \mathbf{J}, \mathbf{B}$ if we don't have any observation yet.

b) Derive the posterior for $\mathbf{g}_*$ given that $\mathbf{y}$ was observed.

c) Verify that the posterior mean of $\mathbf{g}_*$ is the same as the gradient of the posterior mean for $y(x_*)$.

1.2 Integration using Gaussian Processes

Now we want to estimate the value of $\int_a^b y(t)dt$, i.e., the integral of a continuous 1-dimensional (i.e. $d = 1$ function $y(x)$ over the interval $[a, b]$ from labels $\mathbf{y} = y_1, \cdots, y_n$ at points $\mathbf{X} = \{x_1, \cdots, x_n\}$ where $y_i = y(x)$ and $x_i \in \mathcal{R}$.

We define a function $Y(x) = \int_0^x y(t)dt$, and according to the Newton Leibniz formula, there is $\int_a^b y(t)dt = Y(b) - Y(a)$.

a) We treat Y as a GP with kernel K_Y, then the joint distribution of y and Y can also be described by Gaussians. Using the properties described in Problem 1, outline a procedure to obtain this joint distribution $(\mathbf{y}, \mathbf{Y})$, where $\mathbf{Y} = [Y(b), Y(a)]^T$.

Hint: you may want to define $K_y(x_1, x_2) = \frac{\partial^2 K_Y(x_1,x_2)}{\partial x_1 \partial x_2}$ and $K_{yY}(x_1, x_2) = \frac{\partial K_Y(x_1,x_2)}{\partial x_2}$.

b) Derive the posterior $\mathbf{Y}|X, \mathbf{y}, a, b$, which is also a Gaussian distribution.

c) Derive the posterior distribution of the integral $\int_a^b y(t)dt$.

2 Problem 2

By now you may be used to minimizing problems with respect to squared error loss. But there are many other loss functions. In this problem, we will investigate the regression problem using a different loss function.

2.1 Quantile loss

(a) Let's define the following loss

$$\rho_\tau(z) = z \cdot (\tau - I(z < 0)) = \begin{cases} z \cdot (\tau - 1) & \text{if } z < 0 \\ z \cdot \tau & \text{if } z \geq 0, \end{cases}$$

where $\tau \in (0, 1)$ is called the τ-th quantile, and $I(z < 0)$ is the indicator function, i.e. 1 if $z < 0$ and 0 otherwise.

Show that

$$\arg\min_w \sum_i \rho_\tau(y_i - w) = y_\tau,$$

where y_τ is an observation sitting at the τ-th top percentile of the observations (specifically, this means that y_τ is at least exactly τ percent of the observations).

Hint: split the problem into positive and negative parts.

(b) When $\tau = 0.5$, this loss function has a well-known name in statistics. What is that?

2.2 Quantile regression

(a) Let $\{x_i\}_{i=1,\cdots,N}$ be points in $\mathcal{R}^K$ with outputs $\{y_i\}_{i=1,\cdots,N}$ in $\mathcal{R}$. Let $X = (x_1, \cdots, x_N)$. We define the regression quantile as

$$\hat{\beta}(\tau) = \arg\min_{\beta \in \mathcal{R}^K} \sum_{i=1}^{N} \rho_\tau(y_i - x_i^T \beta).$$

Prove that the solution of this problem is equivalent to the solution of the following linear program:

$$\begin{aligned} \arg \min_{\beta \in \mathcal{R}^K, u, v \in \mathcal{R}^N} & u^T \mathbf{1}\tau + v^T \mathbf{1}(1-\tau) \\ \text{subject to } & X^T\beta - y + u - v = 0, \\ & u \geq 0 \\ & v \geq 0 \end{aligned}$$

Hint: split the problem into positive and negative parts.

(b) Show that the dual of the above linear program is:

$$\begin{aligned} \max_z \; & y^T z \\ \text{subject to } & Xz = (1-\tau)X\mathbf{1} \\ & z \in [0,1]^n \end{aligned}$$

(c) What does the value of z_i in the dual problem tell us about $y_i - x_i^T\beta$ in the primal? Specifically, using the KKT conditions, if $z_i = 0$, what can you say about $y_i - x_i^T\beta$? What if $z = 1$? What if $z \in (0, 1)$?