HSE 2021: Mathematical Methods for Data Analysis. Assignment 6: optional

May 31, 2021

Disclaimer

- This is an optional homework, which contains of **4** theoretical problems, 2.5 point each.
- We encourage you to use LaTeXto write the solution. Overleaf is a nice online editor, if you don't want to install it locally. Hand-written solutions will be also accepted, but only if you provide high quality **scans** in the form of a single pdf file. Please, make sure that TAs can read what you've submitted, otherwise, the submission will not be graded.
- You have **10 days** to complete the assignment. We recommend you to start early. No late submissions will be accepted.
- Please, give as much details in your derivation as possible.

Problem 1. Intro to Bayesian ML. [2.5 points]

Consider a univariate Gaussian likelihood:

$$p(x|\mu,\tau) = \mathcal{N}(x|\mu,\tau^{-1}) = \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^2\tau\right). \tag{1}$$

Let's define the following prior for the parameters (μ,τ):

$$p(\mu,\tau) = \mathcal{N}(\mu|\mu_0,(\beta\tau)^{-1}) \cdot \mathrm{Gamma}(\tau;a,b) \tag{2}$$

$$= \left(\frac{\beta\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mu-\mu_0)^2\beta\tau\right) \cdot \frac{b^a\tau^{a-1}\exp(-b\tau)}{\Gamma(a)}. \tag{3}$$

The task

Find the posterior distribution of (μ,τ) after observing N i.i.d. samples $X = (x_1,\ldots,x_N)$ from the $p(x|\mu,\tau)$.

Solution

`YOUR SOLUTION HERE`

Problem 2. Gaussian Processes. [2.5 points]

Assume, that the function $y(x)$, $x \in \mathbb{R}^d$, is a realization of a Gaussian Process with the kernel $K(a, b) = \exp(-\gamma \|a - b\|_2^2)$):

$$y(x) \sim GP\big(0; K(x, x)\big). \tag{4}$$

Namely, for a given x, y has a Gaussian distribution $\mathcal{N}(y|0, K(x, x))$

Suppose two datasets were observed: **noiseless** and **noisy**:

$$D_0 = \{x_n, y(x_n)\}_{n=1}^{N}, \tag{5}$$

$$D_1 = \{x'_m, y(x'_m) + \varepsilon_m\}_{m=1}^{M}, \tag{6}$$

where ε_m are i.i.d. Gaussian: $\varepsilon_m \sim \mathcal{N}(\varepsilon_m|0, \sigma^2)$.

The task

Derive the conditional distribution for a new point $y^* = y(x^*)$, given observed data: $p(y^* \big| D_0, D_1)$.

Hint

You can find useful properties of the Gaussian distribution for this task in the Matrix Cookbook

Solution

`YOUR SOLUTION HERE`

Problem 3. Boosting. [2.5 points]

In this task you will be working with gradient boosting algorithm. Let's firstly recap the notation and the algorithm itself.

$$b_m(x) := \text{the best base model from the family of the algorithms } \mathcal{A} \tag{7}$$
$$\gamma_m(x) := \text{scale or weight of the new model} \tag{8}$$
$$a_M(x) = \sum_{m=0}^{M} \gamma_m b_m(x) := \text{the final composite model} \tag{9}$$

Consider a loss function $L(y, z)$ for the target y and prediction z, and let $\{x_n, y_n\}_{n=1}^N$ be the train dataset with N observations for a regression task. Then gradient boosting algorithm is the following:

1. Initialize $a_0(x) = \hat{z}$ with the constant prediction $\hat{z} = \arg\min\limits_{z \in \mathbb{R}} \sum_{n=1}^N L(y_n, z)$
2. For m from `1` to `M` do:

   Solve the current subproblem $G_m(b, \gamma) = \sum_{n=1}^N L\big(y_n, a_{m-1}(x_n) + \gamma b(x_n)\big) \to \min\limits_{b,\gamma}$, using the following method:

   - Compute the residuals
   $$s_n = -\frac{\partial}{\partial z} L(y_n, z)\Big|_{z=a_{m-1}(x_n)}, n = 1, \dots, N. \tag{10}$$
   - Train the next base algorithm
   $$b_m(x) = \arg\min_{b \in \mathcal{A}} \sum_{n=1}^N \big(b(x_n) - s_n\big)^2. \tag{11}$$
   - Find its weight
   $$\gamma_m = \arg\min_{\gamma} G_m(b_m, \gamma). \tag{12}$$
   - Update the mixture
   $$a_m(x) = a_{m-1}(x) + \gamma_m b_m(x). \tag{13}$$
3. Return $a_M(x) = a_0(x) + \sum_{m=1}^M \gamma_m b_m(x)$.

Finally, the task

Consider Poisson loss, namely $L(y, z) = -yz + \exp z$.

- Derive formula for the residuals at a step m
- Derive first-order conditions for γ at a step m

Solution

`YOUR SOLUTION HERE`

Problem 4. Variational AutoEncoder. [2.5 points]

We observe a dataset $\{x_1, \dots, x_N\}$, in other words, we consider an empirical distribution over x: $p_e(x) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(x)$. We want to infer a latent representation z for a point x from the dataset. Thus, we consider the following generative model with parameters θ:

$$z \sim p(z), \quad x \sim p_\theta(x|z). \tag{14}$$

We choose our generative model to be a linear and assume the presence of the normal noise:

$$p_\theta(x|z) = \mathcal{N}(x|W_p z + \mu_p, \Lambda_p^{-1}), \theta := \{W_p, \mu_p, \Lambda_p^{-1}\}. \tag{15}$$

We want to infer parameters from data as an MLE solution:

$$\theta^* = \arg\max_\theta \mathbb{E}_x \log \int p_\theta(x|z)p(z)dz. \tag{16}$$

Also, we would like to have the ability to find the latent representation z for a new datapoint x. Thus, we will use variational approach to solve the optimization problem:

$$\max_\theta \mathbb{E}_x \log \int p_\theta(x|z)p(z)dz \geq \max_{\theta,\phi} \mathbb{E}_x \int q_\phi(z|x) \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} dz. \tag{17}$$

Since generative process is linear, we would like to use similar structure for the inference:

$$q_\phi(z|x) = \mathcal{N}(z|W_q x + \mu_q, \Lambda_q^{-1}), \phi := \{W_q, \mu_q, \Lambda_q^{-1}\}. \tag{18}$$

Finally, note that taking expectations w.r.t empirical distribution is the same as averaging, which gives us the following objective:

$$\mathcal{L} = \mathbb{E}_x \int q_\phi(z|x) \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} dz = \frac{1}{N}\sum_{n=1}^{N} \int q_\phi(z|x_n) \log \frac{p_\theta(x_n|z)p(z)}{q_\phi(z|x_n)} dz. \tag{19}$$

Finally, the task

- Use first-order conditions (FOC) to find: W_p, μ_p, given W_q, μ_q, Λ_q using objective (19). Note that in the final formula W_p may depend on μ_p and vice versa.
- Is it enough to check the FOC for μ_p? Check the convexity over μ_p.

Solution

```
YOUR SOLUTION HERE
```