Introduction to Machine Learning
Homework 2

Lu Hongyu   Student ID: 2023202269

Date of Submission

Question 1

Answer

1. Derive $p(\theta|X, y)$

The likelihood function is:

$$p(y_n|x_n, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n - x_n^T\theta)^2}{2\sigma^2}\right)$$

The prior distribution is:

$$p(\theta) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\theta - \mu)^T\Sigma^{-1}(\theta - \mu)\right)$$

Using Bayes' theorem, the posterior distribution is:

$$p(\theta|X, y) \propto p(\theta) \prod_{n=1}^{N} p(y_n|x_n, \theta)$$

Substituting the likelihood and prior, we get:

$$p(\theta|X, y) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^T\Sigma^{-1}(\theta - \mu)\right) \prod_{n=1}^{N} \exp\left(-\frac{(y_n - x_n^T\theta)^2}{2\sigma^2}\right)$$

2. Log-Posterior

$$\begin{aligned}\log p(\theta|X, y) &\propto -\frac{1}{2}(\theta - \mu)^T\Sigma^{-1}(\theta - \mu) - \sum_{n=1}^{N} \frac{(y_n - x_n^T\theta)^2}{2\sigma^2} \\ &\propto -(\theta - \mu)^T\Sigma^{-1}(\theta - \mu) - \frac{1}{\sigma^2}(\boldsymbol{y} - X^T\theta)^T(\boldsymbol{y} - X^T\theta)\end{aligned}$$

3. Closed-Form Solution for θ

$$\frac{\partial}{\partial\theta} \log p(\theta|X, y) = -\Sigma^{-1}(\theta - \mu) + \frac{1}{\sigma^2}X^T(\boldsymbol{y} - X^T\theta) = 0$$

Therefore:

$$\hat{\theta} = (\Sigma^{-1} + \frac{1}{\sigma^2}X^TX)^{-1}\left(\Sigma^{-1}\mu + \frac{1}{\sigma^2}X^Ty\right)$$

Question 2

Answer

1.Derivation from (22) to (24):

Given (22) (23):

$$f^* = \arg\min_{f\in\mathcal{H}_K} \sum_{n_1=1}^{N}(y_{n_1} - \sum_{n_2=1}^{N}\alpha_{n_2}K(x_{n_1}, x_{n_2}))^2 + \lambda\sum_{n_1=1}^{N}\sum_{n_2=1}^{N}\alpha_{n_1}\alpha_{n_2}K(x_{n_1}, x_{n_2}) \tag{22}$$

Therefore:

$$\alpha^* = \arg\min_{\alpha\in\mathbb{R}^N} \|y - K\alpha\|_2^2 + \lambda\alpha^T K\alpha \tag{24}$$

where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]$ is the vector of coefficients and $K = [K(x_n, x_n')] \in \mathbb{R}^{N\times N}$ is the Gram matrix.

2. Chain Rule Derivation for the RBF Kernel

$$\begin{aligned} L &= \|y - K\alpha\|_2^2 + \lambda\alpha^T K\alpha \\ &= y^T y - 2y^T K\alpha + \alpha^T K^T K\alpha + \lambda\alpha^T K\alpha \end{aligned}$$

The RBF kernel is given by:

$$K(x_n, x_n') = \exp\left(-\frac{\|x_n - x_n'\|_2^2}{h}\right)$$

Therefore:

$$\frac{\partial K}{\partial h} = [\frac{\partial k(x_i, x_j)}{\partial h}] = [\frac{\|x_i - x_j\|_2^2}{h^2}\cdot k(x_i, x_j)]$$

Let$K' = \frac{\partial K}{\partial h}$ So:

$$\begin{aligned} \frac{\partial L}{\partial h} &= \frac{\partial(-2y^T K\alpha + \alpha^T K^T K\alpha + \lambda\alpha^T K\alpha)}{\partial h} \\ &= -2y^T\frac{\partial K}{\partial h}\alpha + 2\alpha^T K^T\frac{\partial K}{\partial h}\alpha + \lambda\alpha^T\frac{\partial K}{\partial h}\alpha \\ &= -2y^T K'\alpha + 2\alpha^T K^T K'\alpha + \lambda\alpha^T K'\alpha \end{aligned}$$

Question 3

Answer

1. Sparse solution

To encourage a sparse solution for α, I choose $R(\alpha)$ as the L1 norm:

$$R(\alpha) = \|\alpha\|_1$$

The objective becomes:

$$\min_{\alpha\in\mathbb{R}^N} \|y - K\alpha\|_2^2 + \lambda\|\alpha\|_1$$

2. Optimization Algorithm

Iterative soft-thresholding:

- **Initialization:** Start with $\alpha^{(0)} = 0$.
- **Iteration:** For each α_i, update the value of α_i while keeping other coefficients fixed. The update rule is:

$$\alpha_i^{(t+1)} = \text{SoftThreshold}_{\frac{\lambda}{\|K_i\|_2^2}} \left(\frac{K_i^T (y - K_{-i}\alpha_{-i}^t)}{\|K_i\|_2^2} \right)$$

- **Convergence:** Repeat until convergence.