Introduction to Machine Learning
Homework 1

Lu Hongyu Student ID: 2023202269

March 13, 2025

Question 1

Answer

The polynomial regression model can be expressed as:

$$y_n = \mathbf{x}_n^T \theta + \epsilon_n$$

where $\mathbf{x}_n = (1, x_n, x_n^2, \ldots, x_n^{D-1})^T \in \mathbb{R}^D$ represents the feature vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_D)^T \in \mathbb{R}^D$ is the parameter vector, and $\epsilon_n \sim \mathcal{N}(0, x_n^2)$ is the noise term with variance dependent on x_n. The probability density of y_n given x_n and θ is:

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

The likelihood function for the entire dataset is:

$$L(\theta) = \prod_{n=1}^{N} p(y_n|x_n, \theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi x_n^2}} \exp\left(-\frac{(y_n - \mathbf{x}_n^T\theta)^2}{2x_n^2}\right)$$

Maximizing the log-likelihood is equivalent to minimizing the following cost function:

$$J(\theta) = \frac{1}{2}(\mathbf{y} - X\theta)^T W (\mathbf{y} - X\theta)$$

where: - $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$ is the vector of observations, - $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$ is the design matrix, - $W = \text{diag}\left(\frac{1}{x_1^2}, \frac{1}{x_2^2}, \ldots, \frac{1}{x_N^2}\right) \in \mathbb{R}^{N \times N}$ is a diagonal weight matrix.

To find the maximum likelihood estimate (MLE), compute the derivative of $J(\theta)$ with respect to θ and set it to zero:

$$\frac{\partial J}{\partial \theta} = -X^T W (\mathbf{y} - X\theta) = 0$$

Solving this, the closed-form solution for the MLE is:

$$\hat{\theta}_{\text{MLE}} = (X^T W X)^{-1} X^T W \mathbf{y}$$

Question 2

Answer

For maximum a posteriori (MAP) estimation with a Laplace prior, Bayes' theorem gives:

$$p(\theta|\mathbf{y}, X) \propto p(\mathbf{y}|X, \theta)p(\theta|X)$$

Maximizing the posterior is equivalent to maximizing the log-posterior:

$$\log p(\mathbf{y}|X, \theta) + \log p(\theta)$$

Given the likelihood from Question 1 and a Laplace prior $p(\theta) \propto \exp\left(-\sum_{d=1}^{D} \frac{|\theta_d|}{b_d}\right)$, this becomes:

$$\log p(\mathbf{y}|X, \theta) + \log p(\theta) = -\sum_{n=1}^{N} \frac{(y_n - \mathbf{x}_n^T\theta)^2}{2x_n^2} - \sum_{d=1}^{D} \frac{|\theta_d|}{b_d}$$

Thus, the MAP estimate is obtained by minimizing:

$$J_{\text{MAP}}(\theta) = \frac{1}{2}(\mathbf{y} - X\theta)^T W(\mathbf{y} - X\theta) + \sum_{d=1}^{D} \frac{|\theta_d|}{b_d}$$

This objective function includes an L1 regularization term due to the Laplace prior, making it non-differentiable at zero. I use the soft thresholding iterative method to solve it. In the t-th iteration, for each $d = 1, \ldots, D$:

$$\begin{aligned}
\theta_d^{(t+1)} &= \arg\min_{\theta_d} \frac{1}{2}\left[\left(\mathbf{y} - \sum_{i\neq d} \mathbf{x}_i\theta_i^{(t)}\right) - \mathbf{x}_d\theta_d\right]^T W\left[\left(\mathbf{y} - \sum_{i\neq d} \mathbf{x}_i\theta_i^{(t)}\right) - \mathbf{x}_d\theta_d\right] + \frac{|\theta_d|}{b_d} \\
&= S\left(\frac{\mathbf{x}_d^T W\left(\mathbf{y} - \sum_{i\neq d} \mathbf{x}_i\theta_i^{(t)}\right)}{\mathbf{x}_d^T W\mathbf{x}_d}, \frac{1}{b_d\mathbf{x}_d^T W\mathbf{x}_d}\right)
\end{aligned}$$