Regression Assignment — Theory Summary

Sahasra Oleti - 24bB4536

Overview

This assignment explores regression using basis function models with lognormal noise. We derive the appropriate loss function, construct flexible approximators using basis functions, and use decision theory to determine the optimal predictor. Finally, we train the model using gradient descent and examine the bias–variance trade-off.

Part 1: Setup

Q1: MLE Under Lognormal Noise

We assume that the target values are generated according to

$$t = f(x) \cdot \epsilon, \qquad \epsilon \sim \text{Lognormal}(0, \sigma^2),$$

which is equivalent to

$$\ln t = \ln f(x) + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2).$$

Thus the conditional likelihood is

$$p(t_n|x_n) = \frac{1}{t_n \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln t_n - \ln f(x_n))^2}{2\sigma^2} \right).$$

For N samples, the negative log-likelihood is

$$-\log p(\mathbf{t}|\mathbf{x}) = \sum_{n=1}^{N} \left[\ln t_n + \frac{(\ln t_n - \ln f(x_n))^2}{2\sigma^2} \right] + \text{const}.$$

Ignoring constants yields the loss

$$\boxed{\mathcal{L}(f) = \sum_{n=1}^{N} (\ln t_n - \ln f(x_n))^2}$$

Conclusion: Because the noise is lognormal, the appropriate objective is squared error in log-space, rather than standard squared error.

Q2: Choosing Basis Functions

To approximate the unknown function $f(x)$, we express it in terms of basis functions:

$$f(x) = \mathbf{w}^\top \boldsymbol{\phi}(x).$$

A standard 1D choice is the polynomial basis

$$\boldsymbol{\phi}(x) = [1, x, x^2, \dots, x^n]^\top.$$

The degree n controls the model complexity:

- Small n: high bias (underfitting)
- Large n: high variance (overfitting)

Conclusion: We adopt

$$\boxed{\boldsymbol{\phi}(x) = [1, x, x^2, \dots, x^n]^\top}$$

as a flexible basis for nonlinear regression.

Q3: Conditional Mean via Decision Theory

For any chosen function $f(x)$, the expected loss under squared error is

$$E[L] = \int (t - f(x))^2 p(t|x)\, dt.$$

Taking the derivative w.r.t $f(x)$ and setting to zero:

$$\frac{\partial}{\partial f(x)} \int (t - f(x))^2 p(t|x)\, dt = 0.$$

Evaluating the derivative gives

$$-2 \int (t - f(x))\, p(t|x)\, dt = 0.$$

Thus,

$$f(x) = \int t\, p(t|x)\, dt = \mathbb{E}[t|x].$$

Conclusion: Under L_2 loss, the optimal predictor is the conditional mean:

$$\boxed{f(x) = \mathbb{E}[t|x]}$$

Part 2: Training and Model Fitting

We aim to learn weights $\mathbf{w}$ that best approximate $f(x)$.

With polynomial basis functions, predictions are

$$y_n = \mathbf{w}^\top \boldsymbol{\phi}(x_n).$$

Using the log-normal loss derived earlier, the empirical loss becomes

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} \big(\ln t_n - \ln(\mathbf{w}^\top \boldsymbol{\phi}(x_n))\big)^2 + \lambda \|\mathbf{w}\|^2,$$

where λ is a regularization parameter controlling overfitting.

Weights are optimized using stochastic gradient descent (SGD), performing iterative updates

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{L}.$$

Conclusion: Part 2 matches weights to data by minimizing the log-square loss using SGD with optional regularization.

Part 3: Bias–Variance Trade-Off

Bias–variance decomposition explains model performance:

$$\mathbb{E}[(y - f(x))^2] = \underbrace{(\mathbb{E}[y] - f(x))^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(y - \mathbb{E}[y])^2]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible noise}}.$$

- Increasing basis complexity n reduces bias but increases variance.
- Regularization (ridge penalty $\lambda \|\mathbf{w}\|^2$) counteracts variance.
- The optimal λ minimizes

$$\boxed{\text{Bias}^2 + \text{Variance}}$$

Conclusion: We evaluate multiple λ values and empirically estimate the trade-off by computing the average model fit across multiple noisy samples.

Summary

This assignment demonstrates how noise assumptions lead to custom loss functions, how basis expansions yield nonlinear function approximators, and how model complexity and regularization affect performance through bias–variance trade-offs.