DTU

Parameter Estimation

Bayesian Linear Regression and EM

Bayesian Approach to Parameter Inference I

Bayes' Theorem

- The posterior probability $p(\boldsymbol{\theta}|\mathcal{X})$, which updates our belief about $\boldsymbol{\theta}$ after observing data $\mathcal{X}$, is given by:

$$p(\boldsymbol{\theta}|\mathcal{X}) = \frac{p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})}$$

- Here:
  - $p(\mathcal{X}|\boldsymbol{\theta})$ is the **likelihood** of observing $\mathcal{X}$ given the parameters $\boldsymbol{\theta}$.
  - $p(\boldsymbol{\theta})$ is the **prior probability** of $\boldsymbol{\theta}$.
  - $p(\mathcal{X})$, the **evidence**, acts as a normalization constant, ensuring that the posterior probabilities sum to one.

Regression: A Bayesian Perspective I

Linear Regression Model

- The regression model is expressed as:

$$y = \theta_0 + \sum_{k=1}^{K-1} \theta_k \phi_k(\boldsymbol{x}) + \eta$$

- Here, $\eta \in \mathbb{R}$ represents noise, $\boldsymbol{\theta} \in \mathbb{R}^k$ the parameters, and $\boldsymbol{\phi}(\boldsymbol{x})$ the basis functions.

Regression: A Bayesian Perspective II

Bayesian Treatment

- **Maximum Likelihood Estimator (MLE):** $\hat{\theta}_{\text{ML}} = \arg\max_{\theta} p(\mathcal{X}|\theta)$:

$$\hat{\theta}_{ML} = (\Phi^T \Sigma_\eta^{-1} \Phi)^{-1} \Phi^T \Sigma_\eta^{-1} y$$

- **Maximum A Posteriori (MAP)**: $\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{X})$

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- Assuming that both the prior and the likelihood to be Gaussian, the posterior mean and covariance are given by:

$$\mu_{\theta|y} = \theta_0 + (\Sigma_\theta^{-1} + \Phi^T \Sigma_\eta^{-1} \Phi)^{-1} \Phi^T \Sigma_\eta^{-1} (y - \Phi\theta_0)$$

$$\Sigma_{\theta|y} = (\Sigma_\theta^{-1} + \Phi^T \Sigma_\eta^{-1} \Phi)^{-1}$$

Regression: A Bayesian Perspective III

Posterior Predictive Distribution

Given a new input $\boldsymbol{x}$, the predictive distribution for the output y is given by:

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{X}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{X}) d\boldsymbol{\theta}$$

This integration, often not analytically tractable, involves approximation techniques such as Monte Carlo methods for evaluation.

Regression: A Bayesian Perspective IV

Implications and Extensions

There is a direct correspondence between Bayes theorem and common loss functions:

- **MSE**: Normal likelihood with variance $\sigma^2 I$ and a constant prior.
- **Ridge**: Normal likelihood with variance $\sigma^2 I$ and a Normal prior with variance $\sigma_p^2 I$.
- **LASSO**: Normal likelihood with variance $\sigma^2 I$ and a Laplace prior.

Exercise 9.1.4: Deriving Ridge Regression I

Log-Posterior Derivation

Starting from a normal prior for θ, we derive:

$$\ln p(\theta; 0, \sigma_\theta^2 I) = -\frac{K}{2}\ln(2\pi) - \frac{K}{2}\ln \sigma_\theta^2 - \frac{1}{2\sigma_\theta^2}\|\theta\|^2$$

Combining with the log-likelihood:

$$\theta_{\text{MAP}} = \arg\min_\theta \left\{ \frac{1}{2\sigma^2}\|y - f(X, \theta)\|^2 + \frac{1}{2\sigma_\theta^2}\|\theta\|^2 \right\}$$

Exercise 9.1.4: Deriving Ridge Regression II

Reparameterization

Reparameterize with $\sigma_\theta^2 = \frac{\sigma^2}{\lambda} \Rightarrow \lambda = \frac{\sigma^2}{\sigma_\theta^2}$, resulting in:

$$\theta_{\text{MAP}} = \arg\min_\theta \left\{ \|y - f(X, \theta)\|^2 + \lambda \|\theta\|^2 \right\}$$

Linear Regression and the EM Algorithm I

Simplified Regression Model

Assuming Gaussian models for likelihood and prior:

$$p(\theta|y) = N(\theta|\mu_{\theta|y}, \Sigma_{\theta|y})$$

with $\Sigma_\eta = \sigma_\eta^2 I$ and $\Sigma_\theta = \sigma_\theta^2 I$, and $\theta_0 = 0$.
We define the precision variables:

$$\alpha = \frac{1}{\sigma_\theta^2} \quad \text{and} \quad \beta = \frac{1}{\sigma_\eta^2}$$

Linear Regression and the EM Algorithm II

EM Algorithm

E-Step: Compute the Gaussian posterior characteristics:

$$\Sigma_{\theta|y}^{(j)} = \left(\alpha^{(j)} I + \beta^{(j)} \Phi^T \Phi\right)^{-1}$$
$$\boldsymbol{\mu}_{\theta|y}^{(j)} = \beta^{(j)} \Sigma_{\theta|y}^{(j)} \Phi^T \boldsymbol{y}$$

M-Step: Update α and β by optimizing:

$$\alpha^{(j+1)} = \frac{K}{\|\boldsymbol{\mu}_{\theta|\boldsymbol{y}}^{(j)}\|^2 + \text{trace}(\Sigma_{\theta|y}^{(j)})}$$
$$\beta^{(j+1)} = \frac{N}{\|y - \boldsymbol{\Phi}\boldsymbol{\mu}_{\theta|y}^{(j)}\|^2 + \text{trace}(\boldsymbol{\Phi}\Sigma_{\theta|y}^{(j)} A^T)}$$

Exercise 9.3: Bayesian linear regression on real data I

Assume parameters are known.