Linear Regression

Advanced Statistical Inference

Simone Rossi

Maximum Likelihood Estimation

1. Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ with $\boldsymbol{x}_n \in \mathbb{R}^D$, assume the generative model $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ independently. Write down the log-likelihood $\ell(\boldsymbol{w}) = \log p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X})$ and find $\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}) = 0$ to derive the MLE $\boldsymbol{w}^*$.

2. Let $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ be the least squares estimator. The true data is generated as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. Prove that $\mathbb{E}[\hat{\boldsymbol{w}}] = \boldsymbol{w}^*$ by substituting the generative model into the estimator.

3. For a dataset with $\boldsymbol{X} \in \mathbb{R}^{3\times 2}$ and $\boldsymbol{y} \in \mathbb{R}^3$:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 4 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 5 \\ 8 \\ 11 \end{pmatrix}$$

   Compute the ridge regression solution $\boldsymbol{w}^*_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ for $\lambda = 1$. Use Cholesky decomposition to solve the system numerically.

4. Starting from the regularized loss $\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$, show that this is equivalent to the negative log posterior (up to constants):

$$-\log p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = -\log p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) - \log p(\boldsymbol{w})$$

   with Gaussian likelihood $\mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I})$ and prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \tau^2 \boldsymbol{I})$. What is λ in terms of σ^2 and τ^2?

Bayesian Linear Regression

1. Given:

   - Prior: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$

- Likelihood: $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma_y^2 \boldsymbol{I})$

Show that the posterior is Gaussian by computing the exponent of $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X})p(\boldsymbol{w})$ and identifying the posterior precision matrix $\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_y^2}\boldsymbol{X}^\top\boldsymbol{X} + \frac{1}{\sigma_w^2}\boldsymbol{I}$ and mean $\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\frac{1}{\sigma_y^2}\boldsymbol{X}^\top\boldsymbol{y}\right)$.

2. For a new input $\boldsymbol{x}_*$ and posterior $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the predictive distribution is obtained by marginalizing:

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{y}, \boldsymbol{X}) = \int \mathcal{N}(y_* \mid \boldsymbol{w}^\top\boldsymbol{x}_*, \sigma_y^2)\mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{w}$$

Show that this equals $\mathcal{N}(y_* \mid \boldsymbol{\mu}^\top\boldsymbol{x}_*, \boldsymbol{x}_*^\top\boldsymbol{\Sigma}\boldsymbol{x}_* + \sigma_y^2)$ using the property that the convolution of two Gaussians is Gaussian.

3. Consider 1D Bayesian linear regression with:

- Prior: $p(w) = \mathcal{N}(w \mid 0, 1)$
- Single observation: $(x, y) = (1, 2)$ with noise variance $\sigma_y^2 = 1$

Compute the posterior mean μ and variance σ^2 using the formulas $\sigma^2 = \left(\frac{1}{\sigma_y^2}x^2 + \frac{1}{\sigma_w^2}\right)^{-1}$ and $\mu = \sigma^2\frac{1}{\sigma_y^2}xy$. Then predict the distribution of $y_* = f(x_* = 2)$.

4. For a dataset with two observations $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (2, 3)$, and:

- Prior: $p(w) = \mathcal{N}(w \mid 0, 2)$
- Noise variance: $\sigma_y^2 = 0.5$

Construct the matrices $\boldsymbol{X}$, $\boldsymbol{y}$ and compute the posterior covariance $\boldsymbol{\Sigma} = \left(\frac{1}{\sigma_y^2}\boldsymbol{X}^\top\boldsymbol{X} + \frac{1}{2}\right)^{-1}$ and posterior mean. Make a prediction at $x_* = 3$.

Model Selection

1. Suppose two models $\mathcal{M}_1$ (linear) and $\mathcal{M}_2$ (polynomial degree 5) are fit to data. Model $\mathcal{M}_2$ always achieves a higher likelihood $p(\boldsymbol{y} \mid \hat{\boldsymbol{w}}_2, \boldsymbol{X}, \mathcal{M}_2) > p(\boldsymbol{y} \mid \hat{\boldsymbol{w}}_1, \boldsymbol{X}, \mathcal{M}_1)$ on the training set. However, Bayesian model selection chooses $\mathcal{M}_1$. Explain why the marginal likelihood $p(\boldsymbol{y} \mid \boldsymbol{X}, \mathcal{M})$ (which marginalizes over parameters) provides a better criterion than the marginal likelihood of the best-fit parameters.

2. Consider two models for the observed data $\boldsymbol{y}$:

- $\mathcal{M}_1$: Likelihood $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}, \mathcal{M}_1) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, 0.1^2\boldsymbol{I})$ with prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid 0, 100\boldsymbol{I})$

- $\mathcal{M}_2$: Likelihood $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}, \mathcal{M}_2) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, 1^2\boldsymbol{I})$ with prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid 0, 100\boldsymbol{I})$

Which model assigns higher marginal likelihood to a large-variance dataset? (Hint: the marginal likelihood for a Gaussian regression is $p(\boldsymbol{y} \mid \boldsymbol{X}, \mathcal{M}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{X}\boldsymbol{\Sigma}_p\boldsymbol{X}^\top + \sigma^2\boldsymbol{I})$ where $\boldsymbol{\Sigma}_p$ is the prior covariance.)