Sandia National Laboratories

Exceptional service in the national interest

SCIENTIFIC MACHINE LEARNING AND TENSORFLOW TUTORIAL

Bayesian Inference

Ravi G Patel

Scientific Machine Learning Department

February 1 – 2, 2024

Numerical PDEs: Analysis, Algorithms, and Data Challenges

ICERM

Brown University

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-00806O

LIMITATIONS OF POINT ESTIMATES

- Without enough data, inverse problems are ill posed
  - E.g., fit $y = a_0 + a_1 x + a_2 x^2$ given $\{(x_0, y_0), (x_1, y_1)\}$

- Solution,

$$\min_{a_0, a_1, a_2} \sum_i (y_i - (a_0 + a_1 x_i + a_2 x_i^2))^2$$

$$= \min_{\boldsymbol{a}} ||\boldsymbol{y} - X\boldsymbol{a}||_2^2$$

$$\rightarrow X^T X \boldsymbol{a} = X^T \boldsymbol{y}$$

Design matrix, $X = \begin{bmatrix} 1 & x_0^1 & x_0^2 \\ 1 & x_1^1 & x_1^2 \end{bmatrix}$

- Since only 2 data points are given, $X^T X$ is singular

REGULARIZATION IS ADHOC

- Solve $X^T X\boldsymbol{a} = X^T\boldsymbol{y}$ with the Moore-Penrose pseudo-inverse
  - Solution with minimum $||\boldsymbol{a}||_2$

- Regularize the optimization problem

$$\min_{\boldsymbol{a}} ||\boldsymbol{y} - X\boldsymbol{a}||_2^2 + \alpha||\boldsymbol{a}||_2^2$$

$$\rightarrow \boldsymbol{a} = (X^T X + \alpha I)^{-1} X^T \boldsymbol{y}$$

  - How do you choose the regularization parameter?

- Range of reasonable fits. How do we quantify this uncertainty?

OVERVIEW OF PROBABILITY FOR CONTINUOUS RANDOM VARIABLES

- When a random variable (RV) is distributed by a probability density, $x \sim p$

  It has probability, $P(a \leq x \leq b) = \int_a^b p(x)dx$, of taking a value within the interval

- Joint probability, $p(x, y)$, and conditional probability, $p(x|y)$
  - Are related by, $p(x, y) = p(x|y)p(y)$

- Marginal distribution, $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$

- Bayes rule, $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$

- Independent and identically distributed (iid) random variables are independently distributed by the same distribution. Jointly, they are distributed as, $\prod_i p(x_i)$

BAYESIAN INFERENCE

- Instead of the least squares solution, find a distribution of parameters that fit the data
- Bayes rule,

$$p(\boldsymbol{a}|X, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|X, \boldsymbol{a})p(\boldsymbol{a})}{p(\boldsymbol{y}|X)}$$

Likelihood: $p(\boldsymbol{y}|X, \boldsymbol{a})$

Marginal likelihood: $p(\boldsymbol{y}|X) = \int p(\boldsymbol{y}|X, \boldsymbol{a})p(\boldsymbol{a})d\boldsymbol{a}$

Prior distribution: $p(\boldsymbol{a})$

Posterior distribution: $p(\boldsymbol{a}|X, \boldsymbol{y})$

High dimensional integral

- Usually intractable

- E.g., For the quadratic fit,

$$y_i - X_{ij}\alpha_j \sim N(0, \sigma_{likelihood}) \rightarrow p(\boldsymbol{y}|X, \boldsymbol{\alpha}) = \prod_i \frac{1}{\sigma_{likelihood}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - X_{ij}\alpha_j}{\sigma_{likelihood}}\right)^2\right)$$

$$\alpha_i \sim N(0, \sigma_{prior}) \rightarrow p(\boldsymbol{\alpha}) = \prod_k \frac{1}{\sigma_{prior}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\alpha_k}{\sigma_{prior}}\right)^2\right)$$

PROPERTIES OF MULTIVARIATE NORMAL DISTRIBUTIONS

- A multivariate normal is parameterized by a mean and positive definite covariance,

$$y \sim MvN(\mu, \Sigma)$$

- Sampling a MvN,

$$y_i = \mu + L z_i$$
$$z_i \sim MvN(0, I), \quad L = \text{Choleskey}(\Sigma)$$

- Marginal distribution,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim MvN\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}\right) \to \begin{matrix} y_1 \sim MvN(\mu_1, \Sigma_{11}) \\ y_2 \sim MvN(\mu_2, \Sigma_{22}) \end{matrix}$$

- Conditional distribution

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim MvN\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}\right) \to \begin{matrix} y_1 | y_2 \sim MvN(\mu_{1|2}, \Sigma_{1|2}) \\ \mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2) \\ \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \end{matrix}$$

- Linear transformations of MvN RV's are also MvN RV's

$$y \sim MvN(\mu, \Sigma) \to Ly \sim MvN(L\mu, L\Sigma L^T)$$

CONJUGACY AND BAYESIAN INFERENCE

- Occasionally, a likelihood and prior are *conjugate* and the posterior has a closed form expression
- For Gaussian likelihoods and priors are conjugate and the posterior is also Gaussian
- For the quadratic regression problem,

$$p(\boldsymbol{a}|X, y) = MvN(\mu, \Omega)$$

$$\mu = \left[X^T X + \frac{\sigma^2_{likelihood}}{\sigma^2_{prior}} I\right]^{-1} X^T y$$

$$\Omega = \left[\frac{1}{\sigma^2_{likelihood}} X^T X + \frac{1}{\sigma^2_{prior}} I\right]^{-1}$$

- Applies to general linear models with the same form for the likelihood and prior

CONNECTION TO REGULARIZATION IN REGRESSION

- Regularized optimization problem

$$\min_{\boldsymbol{a}} ||\boldsymbol{y} - X\boldsymbol{a}||_2^2 + \alpha||\boldsymbol{a}||_2^2$$

$$\rightarrow \boldsymbol{a} = (X^T X + \alpha I)^{-1} X^T \boldsymbol{y}$$

- Mean of the posterior

$$p(\boldsymbol{\alpha}|X, y) = MvN(\mu, \Omega)$$

$$\mu = \left(X^T X + \frac{\sigma^2_{likelihood}}{\sigma^2_{prior}} I \right)^{-1} X^T y$$

- The *maximum a posterior* (MAP) estimate is equivalent to the solution of the regularized least squares problem (doesn't depend on marginal distribution)

$$\max_{\boldsymbol{\alpha}} p(\boldsymbol{y}|X, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}} - \log p(\boldsymbol{y}|X, \boldsymbol{\alpha}) p(\boldsymbol{\alpha})$$

  - For quadratic fit,

$$\min_{\boldsymbol{\alpha}} \frac{1}{\sigma^2_{likelihood}} ||\boldsymbol{y} - X\boldsymbol{\alpha}||_2^2 + \frac{1}{\sigma^2_{prior}} ||\boldsymbol{\alpha}||_2^2$$

- The *maximum likelihood estimate* (MLE) is equivalent to the least squares solution
- Equivalent relationship for $\text{Laplace Distribution} \rightarrow \ell_1$ regularization
- Bayesian formulation provides intuition for regularization

POSTERIOR PREDICTIVE DISTRIBUTION

- How does uncertainty in the parameters propagate to uncertainty in predictions?
- Posterior predictive distribution,

$$p(y_t|X, y, x_t) = \int p(y_t|\boldsymbol{a}, x_t)p(\boldsymbol{a}|X, \boldsymbol{y})d\boldsymbol{a}$$

- For the general linear model,

$$p(y_t|X, y, x_t) = MvN(\nu, \Sigma_y)$$

$$\nu = x_t\boldsymbol{a}, \quad \Sigma_y = \frac{1}{\sigma_{likelihood}}I + x_t\Omega x_t^t$$

- In general, linear transformations of Gaussian RV’s produce Gaussian RV’s