STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

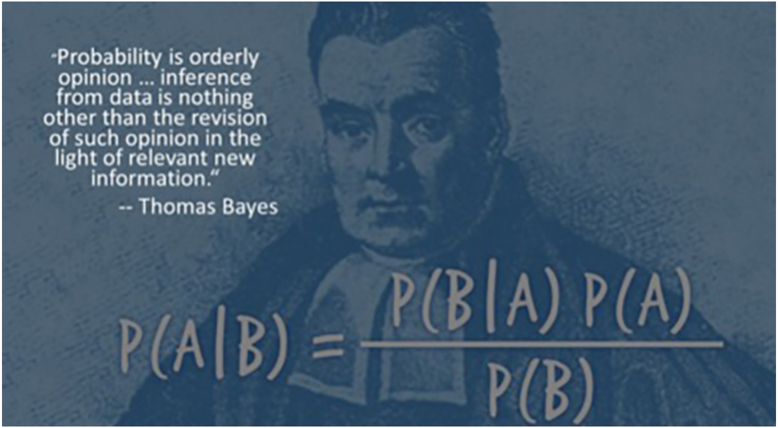
yinfeng@cuhk.edu.cn

April 27, 2020

Overview

1. Bayesian Inference
2. Bayesian Linear Regression
3. Appendix

The Heart of Bayesian Inference



Does anybody know a more general Bayes theorem?

The Heart of Bayesian Inference

Laplace further developed and popularized Bayesian inference:

Later Laplace acknowledges Bayes by

"Bayes a cherché directement la probabilité que les possibilités indiquées par des expériences déjà faites sont comprises dans les limites données et il y est parvenu d'une manière fine et très ingénieuse"

[Essai philosophique sur les probabilités, 1810]

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis/parameters as more evidence/information/data becomes available (often in sequence).

Bayesian Linear Regression: Model

A Bayesian multiple linear regression model is given by

$$y = \mathbf{x}^T \mathbf{w} + \varepsilon = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_k x_k + \varepsilon \tag{1}$$

where

- model parameters $w_j, j = 0, 1, 2, ..., k$ are unknown and assumed to be random with a prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_p)$;
- $x_j, j = 1, 2, ..., k$ are the inputs (deterministic and precisely known) and y is the output;
- ε is random error term, often assumed to be Gaussian i.i.d., namely $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Here, for simplicity, we assume σ^2 is known.

Here, we use $\mathbf{w}$, instead of $\boldsymbol{\beta}$, to differentiate between Bayesian linear regression model with the classic (Frequentist) linear regression model.

Bayesian Linear Regression: Two Major Tasks

1. Given the training data $\{\mathbf{X}, \mathbf{y}\}$, find out the posterior distribution $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$.
2. Most importantly, given a novel input $\mathbf{x}_*$, find out the posterior distribution of the predicted output $p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X})$.

Bayesian Linear Regression: Posterior $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$

Due to the Bayes rule:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}, \tag{2}$$

where

- $p(\mathbf{w})$ is the prior distribution of the model parameters, $\mathbf{w}$;
- $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ is the likelihood, given a $\mathbf{w}$;
- $p(\mathbf{y}|\mathbf{X})$ is the normalizing constant, also known as marginal likelihood, because $p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}$.

The posterior combines the likelihood and the prior, and captures everything we know about the model parameters.

Bayesian Linear Regression: Posterior $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$

From our assumptions, $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$ and $p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_n)$, the posterior is Gaussian distributed and can be derived as:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \sim \mathcal{N}(\bar{\mathbf{w}}, \boldsymbol{\Sigma}), \tag{3}$$

where

$$\bar{\mathbf{w}} = \sigma^{-2} \boldsymbol{\Sigma} \mathbf{X}^T \mathbf{y}, \tag{4}$$

$$\boldsymbol{\Sigma} = \left(\sigma^{-2} \mathbf{X}^T \mathbf{X} + \boldsymbol{\Sigma}_p^{-1} \right)^{-1}. \tag{5}$$

Note that the mean of the posterior distribution, $\bar{\mathbf{w}}$, is also its mode, which is also called the *maximum-a-posteriori* (MAP) estimate of $\mathbf{w}$.

The posterior distribution of $\mathbf{w}$ lays the foundation for deriving the posterior distribution of the output.

Bayesian Linear Regression: Posterior $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$

What is the connection between the posterior mean $\bar{\mathbf{w}}$ with the ordinary (Frequentist) least-squares and the regularized ridge regression?

Note that:

$$\bar{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + \sigma^2\boldsymbol{\Sigma}_p^{-1}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{6}$$

Bayesian Linear Regression: Posterior $p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X})$

To make predictions for a test case we average over all possible parameter values, weighted by their posterior probability. This is in contrast to non-Bayesian schemes, where a single parameter is typically chosen by some criterion.

The posterior distribution of the output is given by

$$
\begin{aligned}
p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}) &= \int p(y_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|\mathbf{y}, \mathbf{X}) d\mathbf{w} & (7)\\
&\sim \mathcal{N}(\sigma^{-2}\mathbf{x}_*^T \mathbf{\Sigma} \mathbf{X}^T \mathbf{y}, \mathbf{x}_*^T \mathbf{\Sigma} \mathbf{x}_* + \sigma^2), & (8)
\end{aligned}
$$

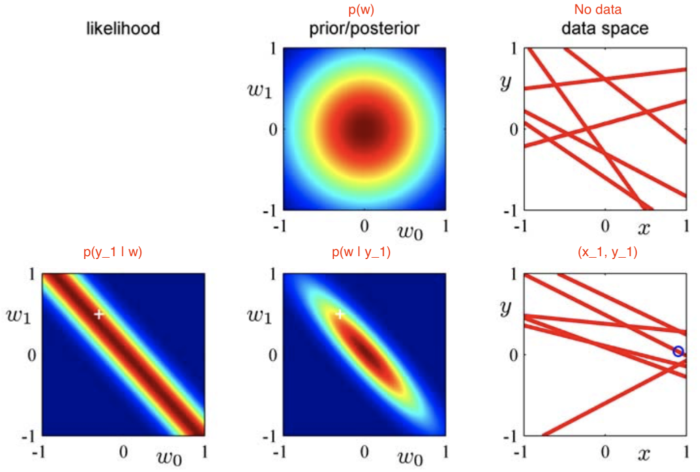
which is again Gaussian.

Bayesian Linear Regression: Example

About the data:

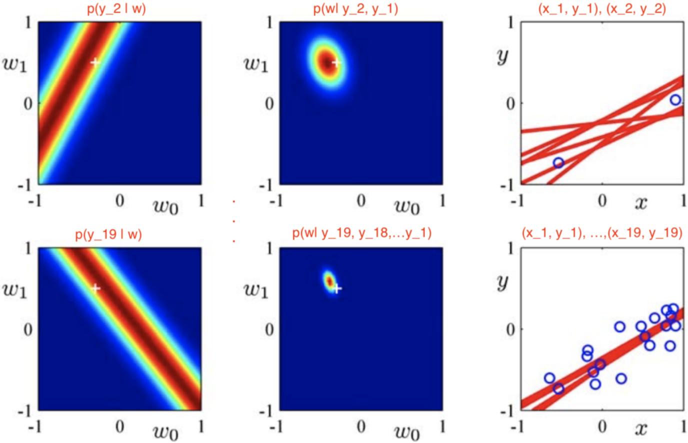
- We generate synthetic data from the function $f(x, a) = a_0 + a_1 x$ with the true parameter values $a_0 = -0.3$ and $a_1 = 0.5$ by first choosing values of x_i from the uniform distribution $\mathcal{U}(-1, 1)$.
- Add independent Gaussian noise with $\sigma = 0.2$ to obtain the output values y_i.
- Assume a Bayesian linear regression model with known σ.
- "Blue dots" represent data (x_i, y_i), "white plus" represents the true parameter.
- "Red lines" represent the **w** values in terms of regression line.

Bayesian Linear Regression: Example I



Bayesian simple linear regression with $y = w_0 + w_1 x + \varepsilon$.

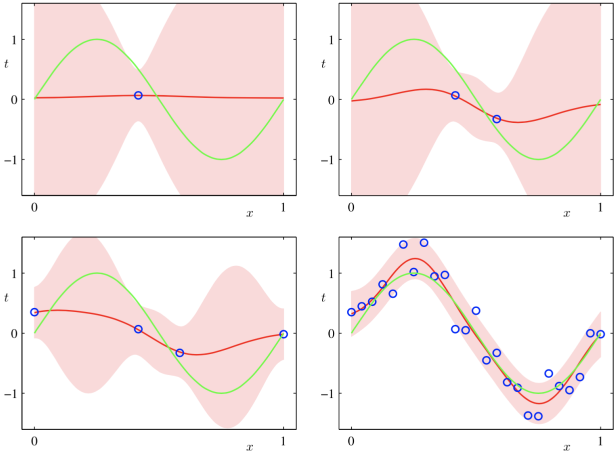
Bayesian Linear Regression: Example I



Bayesian simple linear regression with $y = w_0 + w_1 x + \varepsilon$.

Bayesian Linear Regression: Example II

Posterior prediction of the sinusoidal data.



Bayesian Vs. Frequentist

- Needs to select a prior distribution rather "subjectively"
- Inference relies on both the prior distribution and the likelihood
- More complicated model and heavier computational complexity due to the high dimensional integration over the parameters
- Provides a posterior distribution over the desired parameters/hypothesis/prediction instead of a point estimate

Gauss, 1777-1855, German

Bayes, 1701-1761, English

Reference

1. C. Rasmussen, Gaussian Process for Machine Learning, MIT press, 2006.
2. C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

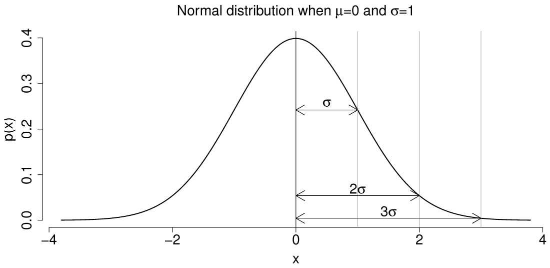
Appendix: Gaussian Distribution–Univariate Case

The probability density function (pdf) is

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]. \tag{9}$$

The mean and variance are

$$\mathbb{E}(x) = \mu, \quad var(x) = \mathbb{E}\left[(x - \mathbb{E}(x))^2\right] = \sigma^2. \tag{10}$$



Appendix: Gaussian Distribution–Multivariate Case

Let $\mathbf{x} \in \mathbb{R}^{d_x}$ be a multivariate Gaussian distribution with the probability density function (pdf) is

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma_{xx}) = \frac{1}{(2\pi)^{d_x/2} \det(\Sigma_{xx})^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma_{xx}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right] \quad (11)$$

The mean and covariance matrix are

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}, \quad Cov(\mathbf{x}) = \mathbb{E}\left[(\mathbf{x}-\mathbb{E}(\mathbf{x}))(\mathbf{x}-\mathbb{E}(\mathbf{x}))^T\right] = \Sigma_{xx}. \quad (12)$$

Appendix: Linear Gaussian System

If the two random variables $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$ have Gaussian distributions:

$$p(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_{xx}) \tag{13}$$

and

$$p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \Sigma_{yy}) \tag{14}$$

then the joint distribution is

$$p(\mathbf{x}, \mathbf{y}) \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \mathbf{A}\boldsymbol{\mu}_x + \mathbf{b} \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xx}\mathbf{A}^T \\ \mathbf{A}\Sigma_{xx} & \mathbf{A}\Sigma_{xx}\mathbf{A}^T + \Sigma_{yy} \end{bmatrix}\right). \tag{15}$$

Here, $\mathbf{A}$ is a known constant matrix of size $d_y \times d_x$.

Appendix: Conditional Gaussian Distribution

If the two random variables $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$ are jointly Gaussian with the following joint distribution:

$$p(\mathbf{x}, \mathbf{y}) \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right), \tag{16}$$

then it is easy to derive the following conditional probabilities:

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{x|y}, \Sigma_{x|y}), \qquad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{y|x}, \Sigma_{y|x}), \tag{17}$$

where

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \qquad \boldsymbol{\mu}_{y|x} = \boldsymbol{\mu}_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x), \tag{18}$$

and

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}, \qquad \Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{19}$$

Following the previous slide where the joint Gaussian distribution $p(\mathbf{x}, \mathbf{y})$ was defined.

The marginal distributions out of it are still Gaussian, i.e.,

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y})\mathrm{d}\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_x, \Sigma_{xx}\right), \tag{20}$$

and

$$p(\mathbf{y}) = \int p(\mathbf{x}, \boldsymbol{y})\mathrm{d}\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}_y, \Sigma_{yy}\right). \tag{21}$$