03 Bayesian Linear Regression

July 1, 2021

Information: *Basic concepts and simple examples of Bayesian linear regression*

Written by: *Zihao Xu*

Last update date: *07.01.2021*

1 Maximum Likelihood Estimation

1.1 Motivation

In the chapter talking about ***Generalization and Regularization***, the concepts of parameter estimation, bias and variance are used to formally characterize notions of generalization, underfitting and overfitting. Here are some important remarks.

- View the parameter estimator $\hat{\boldsymbol{\theta}}$ as a **function** of the sampled training dataset

$$\hat{\boldsymbol{\theta}} = g\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(m)}\right)$$

- The datasets (training, testing and probably validation) are generated by a **i.i.d.** probability distribution over datasets called the **data-generating process** (i.i.d. assumptions can be applied to almost all the common tasks)
- Assume that the true parameter value $\boldsymbol{\theta}$ is fixed but unknown
- Since the **data** is drawn from a **random process**, any function of the data is random, which means the parameter estimator $\hat{\boldsymbol{\theta}}$ is a **random variable**

The concepts of **bias** and **variance** are used to measure the performance of a parameter estimator. However, **for obtaining a good estimator**, it's not a good idea to guess that some function might make a good estimator and then to analyze its bias and variance. This motivated some principles from which specific functions that are good estimators for different models can be derived.

1.2 Definition

- Consider a set of m examples $\mathcal{D} = \left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(m)}\right\}$ drawn independently from the true but unknown data-generating distribution $p_{\text{data}}\left(\mathbf{x}\right)$. Let $p_{\text{model}}\left(\mathbf{x}|\boldsymbol{\theta}\right)$ be a parametric family of probability distributions over the same space indexed by $\boldsymbol{\theta}$
  - That is to say, p_{model} maps any configuration $\mathbf{x}$ to a real number estimating the true probability $p_{\text{data}}(\mathbf{x})$

- Particularly, focus on the **likelihood** which is first introduced in the prerequisite chapter ***Probability Theory***

  $$\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)} \mid \boldsymbol{\theta} \sim p_{\text{model}}\left(\mathbf{x}^{(1:m)} \mid \boldsymbol{\theta}\right)$$

  As a fast review, the likelihood tells us how *plausible* it is to observe $\mathbf{x}^{(1:m)}$ if we know the model parameters are θ

- Since the examples are assumed to be drawn **independently**, the likelihood can be factorized

  $$p_{\text{model}}\left(\mathbf{x}^{(1:m)} \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{m} p_{\text{model}}\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$$

  Then **maximum likelihood** estimator for $\boldsymbol{\theta}$ is then defined as

  $$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$$

- While this simple production may lead to a lot of inconveniences such as **numerical underflow**, taking the **logarithm** of the likelihood does not change the location for maximum ($\arg\max_{\boldsymbol{\theta}}$) but does conveniently transform a product into a sum

  $$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\text{model}}\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$$

- Obviously, rescaling the likelihood does not change the location for maximum ($\arg\max_{\boldsymbol{\theta}}$), we can divide by m to obtain a version of the criterion that is expressed as an expectation with respect to the empirical distribution $\hat{p}_{\text{data}}$ defined by the training data

  $$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log p_{\text{model}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\right]$$

- The most common choice for the likelihood of a single measurement is to pick it to be **Gaussian**

1.3 KL divergence

- Maximum likelihood estimation can be viewed as minimizing the dissimilarity between the empirical distribution $\hat{p}_{\text{data}}$, defined by the training set and the model distribution, with the degree of dissimilarity between the two measure by the **KL divergence**

  $$D_{\text{KL}}\left(\hat{p}_{\text{data}} \| p_{\text{model}}\right) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log \hat{p}_{\text{data}}(\mathbf{x}) - \log p_{\text{model}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\right]$$

  The term on the left is a function only of the data-generating process, not the model. This means when we train the model to minimize the KL divergence, we need only minimize

  $$-\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log p_{\text{model}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\right]$$

- Minimizing this KL divergence corresponds exactly to minimizing the cross-entropy between the distributions. By definition, any loss consisting a negative log-likelihood is a **cross-entropy** between the **empirical distribution** defined by the training set ($\hat{p}_{\text{data}}$), and the **probability distribution** defined by the model (p_{model})

1.4 Conditional Log-Likelihood

- To apply *MLE* to most **supervised learning** tasks of predicting $\mathbf{y}$ given $\mathbf{x}$, the maximum likelihood estimator is generalized to estimate a conditional probability $p_{\text{model}}\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}\right)$
- Consider a set of m examples $\mathcal{D} = \left\{\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right), \left(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}\right), \cdots, \left(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\right)\right\}$ drawn independently from the true but unknown data-generating distribution $p_{\text{data}}\left(\mathbf{x}, \mathbf{y}\right)$. Factorize the data-generating process

  $$p_{\text{data}}\left(\mathbf{x}, \mathbf{y}\right) = p_{\text{data}}\left(\mathbf{y} \mid \mathbf{x}\right) p_{\text{data}}\left(\mathbf{x}\right)$$

  Let $p_{\text{model}}\left(\mathbf{x}, \mathbf{y} |\, \boldsymbol{\theta}\right)$ be a parametric family of probability distributions over the same space indexed by $\boldsymbol{\theta}$. It also can be factorized

  $$p_{\text{model}}\left(\mathbf{x}, \mathbf{y} |\, \boldsymbol{\theta}\right) = p_{\text{model}}\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}\right) p_{\text{data}}\left(\mathbf{x}\right)$$

  Notice that the later part $p_{\text{data}}\left(\mathbf{x}\right)$ is **fixed and shared**, the maximum likelihood estimation is going to focus on

  $$p_{\text{model}}\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}\right)$$

  Under the **i.i.d.** assumption, it can be decomposed into

  $$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}\left(\mathbf{y}^{(i)} \,\middle|\, \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$

  Similarly, this optimization problem is usually converted into a minimization problem by the **negative logarithm** operation considering computation issues

  $$\boldsymbol{\theta}_{\text{ML}} = \arg\min_{\boldsymbol{\theta}} \left[-\sum_{i=1}^{m} \log\left[p_{\text{model}}\left(\mathbf{y}^{(i)} \,\middle|\, \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)\right]\right]$$

1.5 Least Squares as Maximum Likelihood

- Least squares minimizing the mean square error is **equal** to maximum likelihood estimation when the likelihood is assigned to be **Gaussian**
- Assume the model is $\hat{y} = f(\mathbf{x}; \boldsymbol{\theta})$ with the dataset

  $$\mathbf{X} = \begin{bmatrix}\mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(m)}\end{bmatrix}, \mathbf{y} = \begin{bmatrix}y^{(1)} & y^{(2)} & \cdots & y^{(m)}\end{bmatrix}$$

  The solution for $\boldsymbol{\theta}$ via least squares would be

  $$\boldsymbol{\theta}_{\text{LS}} = \arg\min_{\boldsymbol{\theta}} \left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2$$

- From the point of view of maximum likelihood estimation, think of the model as producing a conditional distribution $p_{\text{model}}\left(y \mid \mathbf{x}, \boldsymbol{\theta}\right)$ instead of producing a single prediction $\hat{y} = f(\mathbf{x}; \boldsymbol{\theta})$. Assign this likelihood of a single measurement to be Gaussian

  $$p_{\text{model}}\left(y^{(i)} \,\middle|\, \mathbf{x}^{(i)}, \boldsymbol{\theta}, \sigma\right) = \mathcal{N}\left(y^{(i)} \left| f(\mathbf{x}; \boldsymbol{\theta}), \sigma^2\right.\right)$$

  where σ models the **noise**. This correspond to the belief that the measurement is around the model prediction $f(\mathbf{x}; \boldsymbol{\theta})$ but it is contained with Gaussian noise of variance σ^2. For all the data, we have

  $$\begin{aligned} p_{\text{model}}\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma\right) &= \mathcal{N}\left(\mathbf{y} \left| f\left(\mathbf{X}; \boldsymbol{\theta}\right), \sigma^2 \mathbf{I}_m\right.\right) \\ &= (2\pi)^{-m/2} \sigma^{-m} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2\right) \end{aligned}$$

Then we have the maximum likelihood estimation to be

$$\begin{aligned}
\boldsymbol{\theta}_{\mathrm{ML}} &= \arg\min_{\boldsymbol{\theta}} \left[-\log\left[p_{\mathrm{model}}\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma\right)\right]\right] \\
&= \arg\min_{\boldsymbol{\theta}} \left[\frac{m}{2}\log\left(2\pi\right) + m\log(\sigma) + \frac{1}{2\sigma^2}\left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2\right] \\
&= \arg\min_{\boldsymbol{\theta}} \left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2 \\
&= \boldsymbol{\theta}_{\mathrm{LS}}
\end{aligned}$$

Maximizing the likelihood with respect to $\boldsymbol{\theta}$ yields the same estimate as minimizing the squared error.

- The two criteria have **different values** but the **same location of the optimum**, which justifies the use of the LS as a maximum likelihood estimation procedure.
- Notice that σ is also a parameter to be optimized, maximize the likelihood with respect to σ

$$\begin{aligned}
\sigma_{\mathrm{ML}} &= \arg\min_{\sigma} \left[-\log\left[p_{\mathrm{model}}\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma\right)\right]\right] \\
&= \arg\min_{\sigma} \left[\frac{m}{2}\log\left(2\pi\right) + m\log(\sigma) + \frac{1}{2\sigma^2}\left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2\right] \\
&= \arg\min_{\sigma} \left[m\log(\sigma) + \frac{1}{2\sigma^2}\left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2\right]
\end{aligned}$$

It can be easily solved by setting the derivative with respect to σ to zero

$$\begin{aligned}
m\frac{1}{\sigma_{\mathrm{ML}}} - \frac{1}{\sigma_{\mathrm{ML}}^3}\left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2 &= 0 \\
m\sigma_{\mathrm{ML}}^2 - \left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2 &= 0 \\
\sigma_{\mathrm{ML}}^2 &= \frac{1}{m}\left\|\mathbf{y} - f\left(\mathbf{X}; \boldsymbol{\theta}\right)\right\|_2^2
\end{aligned}$$

- With the maximum likelihood estimation $\boldsymbol{\theta}_{\mathrm{ML}}, \sigma_{\mathrm{ML}}$, we can make **predictions** about y at a new point $\mathbf{x}$

$$p\left(y \left|\mathbf{x}, \boldsymbol{\theta}_{\mathrm{ML}}, \sigma_{\mathrm{ML}}\right.\right) = \mathcal{N}\left(y \left| f(\mathbf{x}; \boldsymbol{\theta}_{\mathrm{ML}}), \sigma_{\mathrm{ML}}^2\right.\right)$$

1.6 Properties of Maximum Likelihood

- Maximum likelihood estimator can be shown to be the **best** estimator asymptotically as the number of examples $m \to \infty$, in terms of its rate of convergence as m increases
- Under appropriate conditions, the maximum likelihood estimator has the property of **consistency**
  - The true distribution p_{data} must lie within the model family p_{model}
  - The true distribution p_{data} must correspond to exactly one value of $\boldsymbol{\theta}$
- The **statistical efficiency**, meaning that one consistent estimator may obtain lower generalization error for a fixed number of samples m, of the maximum likelihood estimator is very high among consistent estimators

1.7 Example: Linear Regression

[ ]:

```
```