The evolution from MLE to MAP to Bayesian Learning

Zhe Li

January 13, 2015

1 The evolution from MLE to MAP to Bayesian

Based on the linear regression, we illustrate the evolution from Maximum Loglikelihood Estimation(MLE) to Maximum A Posterior (MAP) to Bayesian Learning (BL). Highlevel speaking, MLE is the parameter estimation without considering prior knowledge (information) of parameters or prior knowledge of parameters unavailable. MAP is the parameters estimation when encoded prior knowledge of parameters. Different from the idea of obtaining of single parameters in MLE or MAP, Bayesian Learing consider that parameters we intend to obtain is not single point and it is a distribution. We will give the detailed derivation of MLE, MAP and BL to show the difference among them. Given a set of training data $(\mathbf{x}_i, y_i), i = 1, \cdots n$, where $\mathbf{x}_i \in \mathbf{R}^d$ denotes the feature representation of the i^{th} example and y_i denotes its target output.

1.1 Maximum Loglikelihood Estimation

In the linear regression, we assume that

$$y = w^T \Phi(\mathbf{x}) + \epsilon \tag{1}$$

where $w \in \mathbf{R}^d$ and $\Phi(\mathbf{x})$ is the basis function. And ϵ is the Gaussian noise, that is $\epsilon \sim \mathcal{N}(0, \beta^{-1})$. We would like to find a w which has the maximux probability to generate $\{(\mathbf{x}_i, y_i), i = 1, \cdots n\}$,

$$p(y|\mathbf{x}; w) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, w) \tag{2}$$

where $p(y_i|\mathbf{x}_i; w)$

$$p(y_i|\mathbf{x}_i; w) = \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\{-\frac{\beta}{2}(y_i - w^T \Phi(\mathbf{x}_i))^2\} \tag{3}$$

Taking log on both sides of Eq. (**??**) and plugging Eq. (**??**), it gives

$$\begin{aligned}\log p(y|\mathbf{x};w) &= \sum_{i=1}^{n} \log p(y_i|x_i;w)\\ &= \sum_{i=1}^{n} \log \left\{ \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\{-\frac{\beta}{2}(y_i - w^T\Phi(\mathbf{x}_i))^2\} \right\}\end{aligned}$$

when one attempts to maximize the above equation, some constant terms can be ignored, that is

$$\begin{aligned}&\underset{w\in\mathbf{R}^d}{max} \sum_{i=1}^{n} \log p(y_i|x_i;w)\\ &\underset{w\in\mathbf{R}^d}{max} \sum_{i=1}^{n} \log \left\{ \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\{-\frac{\beta}{2}(y_i - w^T\Phi(\mathbf{x}_i))^2\} \right\}\\ &\equiv \underset{w\in\mathbf{R}^d}{max} \sum_{i=1}^{n} \left\{ \frac{1}{2}\log\beta - \frac{\beta}{2}(y_i - w^T\Phi(\mathbf{x}_i))^2 \right\}\\ &\equiv \underset{w\in\mathbf{R}^d}{max} \left\{ \sum_{i=1}^{n} \frac{1}{2}\log\beta - \sum_{i=1}^{n} \frac{\beta}{2}(y_i - w^T\Phi(\mathbf{x}_i))^2 \right\}\\ &\equiv \underset{w\in\mathbf{R}^d}{min} \sum_{i=1}^{n} \frac{\beta}{2}(y_i - w^T\Phi(\mathbf{x}_i))^2 \qquad \text{(Least Square)}\end{aligned}$$

The above shows that MLE with Gaussian noise is equvelent to Least Square. It is easy to obtain the closed form for the above optimization problem. Taking the gradient of objective function w.r.t w and setting it to zeros

$$\begin{aligned}&\sum_{i=1}^{n} \beta(y_i - w^T\Phi(\mathbf{x}_i))(-\Phi(\mathbf{x}_i)) = 0\\ &\sum_{i=1}^{n} w^T\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i) = \sum_{i=1}^{n} y_i\Phi(\mathbf{x}_i)\\ &\sum_{i=1}^{n} \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T w = \sum_{i=1}^{n} y_i\Phi(\mathbf{x}_i)\end{aligned}$$

For simplicity, if we denote matrix Φ as

$$\Phi = \begin{bmatrix} \Phi(x_1)^T \\ \Phi(x_2)^T \\ \vdots \\ \Phi(x_n)^T \end{bmatrix}$$

we can write the above equation in the matrix form, which is

$$\Phi^T \Phi w = \Phi^T y \tag{4}$$

So the solution w is

$$w = (\Phi^T \Phi)^{-1} \Phi^T y \tag{5}$$

1.2 Maximum A Posterior

With the consideration of prior knowledge of parameter w, we would like to maximize the probability of w given the dataset $D = \{(\mathbf{x}_i, y_i), i = 1, \cdots n\}$, the posterior of w is

$$p(w|D) \propto p(D|w)p(w) \tag{6}$$

Where $p(w)$ is the prior distribution of w. Consider $p(w) \sim \mathcal{N}(0, \alpha^{-1} I)$, where matrix I is $d \times d$ identity matrix. Thus, $p(w|D)$ is

$$\begin{aligned}
\log p(w|D) &\propto \log p(D|w)p(w) \\
&= \log p(D|W) + \log p(w) \\
&= \sum_{i=1}^{n} \log \left\{ \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\{-\frac{\beta}{2}(y_i - w^T \Phi(\mathbf{x}_i))^2\} \right\} + \log \left\{ \frac{\alpha^{d/2}}{(2\pi)^{d/2}} \exp(-\frac{\alpha}{2}||w||^2) \right\}
\end{aligned}$$

Maximizing the $\log p(w|D)$ and ignoring the constant terms

$$\begin{aligned}
&\underset{w \in \mathbf{R}^d}{max} \sum_{i=1}^{n} \log \left\{ \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\{-\frac{\beta}{2}(y_i - w^T \Phi(\mathbf{x}_i))^2\} \right\} + \log \left\{ \frac{\alpha^{d/2}}{(2\pi)^{d/2}} \exp(-\frac{\alpha}{2}||w||^2) \right\} \\
&\equiv \underset{w \in \mathbf{R}^d}{max} \sum_{i=1}^{n} \left\{ \frac{1}{2} \log \beta - \frac{\beta}{2}(y_i - w^T \Phi(\mathbf{x}_i))^2 \right\} + \frac{d}{2} \log(\alpha) - \frac{\alpha}{2}||w||^2 \\
&\equiv \underset{w \in \mathbf{R}^d}{max} \left\{ \sum_{i=1}^{n} \frac{1}{2} \log \beta - \sum_{i=1}^{n} \frac{\beta}{2}(y_i - w^T \Phi(\mathbf{x}_i))^2 + \frac{d}{2} \log(\alpha) - \frac{\alpha}{2}||w||^2 \right\} \\
&\equiv \underset{w \in \mathbf{R}^d}{min} \sum_{i=1}^{n} \frac{\beta}{2}(y_i - w^T \Phi(\mathbf{x}_i))^2 + \frac{\alpha}{2}||w||^2 \\
&\equiv \underset{w \in \mathbf{R}^d}{min} \sum_{i=1}^{n} (y_i - w^T \Phi(\mathbf{x}_i))^2 + \frac{\alpha}{\beta}||w||^2 \quad \text{(Ridge Regression Let } \lambda = \frac{\alpha}{\beta})
\end{aligned}$$

Similar to MLE, the closed form for ridge regression also can be obtained. Taking gradient of objective function w.r.t w and set it to zeros, the solution w is

$$w^* = (\Phi^T \Phi + \alpha I)^{-1} \Phi^T y \tag{7}$$

Here, it is necessary to mension that if the prior distribution of w is not normal distribution, it will result in the different regression model. For example, if the prior distribution of w is Lapacian distribution, it leads to Lasso model.

Take a detour to discuss the dual form of ridge regression, for ridge regress

$$\underset{w \in \mathbf{R}^d}{min} J(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - w^T\Phi(\mathbf{x}_i))^2 + \frac{\lambda}{2}||w||^2 \tag{8}$$

Taking gradient of $J(w)$ w.r.t w and set it to zero, we get

$$w = -\frac{1}{\lambda}\sum_{i=1}^{n}(w^T\Phi(\mathbf{x}_i) - y_i)\Phi(\mathbf{x}_i) = \Phi^T\alpha \tag{9}$$

Here, let α is the vector with i^{th} entry $-\frac{1}{\lambda}(w^T\Phi(\mathbf{x}_i) - y_i)$ and Φ defined as same as previous. Plugging $w = \Phi^T\alpha$ into Eq. (**??**),

$$\begin{aligned}
J(w) &= \frac{1}{2}\sum_{i=1}^{n}(w^T\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^Tw - 2w^T\Phi(\mathbf{x}_i)y_i + y_i^2) + \frac{\lambda}{2}||w||^2 \\
&= \frac{1}{2}w^T\Phi^T\Phi w - w^T\Phi^Ty + \frac{1}{2}y^Ty + \frac{\lambda}{2}w^Tw \\
J(\alpha) &= \frac{1}{2}\alpha^T\Phi\Phi^T\Phi\Phi^T\alpha - \alpha^T\Phi\Phi^Ty + \frac{1}{2}y^Ty + \frac{\lambda}{2}\alpha^T\Phi\Phi^T\alpha \\
&= \frac{1}{2}\alpha^TKK\alpha - \alpha^TKy + \frac{1}{2}y^Ty + \frac{\lambda}{2}\alpha^TK\alpha \qquad (K = \Phi\Phi^T)
\end{aligned}$$

where K is the Kernel Matrix. Taking the gradient of $J(\alpha)$ and set it to zero, it gives

$$\alpha = (K + \lambda I)^{-1}y \tag{10}$$

1.3 Bayesian Learning

In Bayesian Learning, there are two problems, which is to get the posterior distribution of parameter w and to predict y usign the posterior distribution of w.

1.3.1 Postorier Distribution of w

we are more interested the posterior distribution of parameter w which contains more information than a single parameter w. The posterior distribution of w is

$$\begin{aligned}
p(w|D) &\propto p(D|w)p(w) \\
&= \prod_{i=1}^{n} p(y_i|\mathbf{x}_i; w)p(w) \\
&= \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\{-\frac{\beta}{2}(y_i - w^T\Phi(x_i))^2\} \frac{\alpha^{d/2}}{(2\pi)^{d/2}} \exp(-\frac{\alpha}{2}||w||^2) \\
&= (\frac{\beta}{2\pi})^{n/2} \exp\{-\frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T\Phi(x_i))^2\}(\frac{\alpha}{2\pi})^{d/2} \exp(-\frac{\alpha}{2}||w||^2) \\
&= (\frac{\beta}{2\pi})^{n/2}(\frac{\alpha}{2\pi})^{d/2} \exp\{-\frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T\Phi(x_i))^2 - \frac{\alpha}{2}||w||^2\} \\
&= (\frac{\beta}{2\pi})^{n/2}(\frac{\alpha}{2\pi})^{d/2} \exp\{-\frac{\beta}{2}\sum_{i=1}^{n}(y_i^2 - 2w^T\Phi(\mathbf{x}_i)y_i + w^T\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T w) - \frac{\alpha}{2}||w||^2\}
\end{aligned}$$

Ignoring some constant terms and using matrix form, it yields

$$\begin{aligned}
&(\frac{\beta}{2\pi})^{n/2}(\frac{\alpha}{2\pi})^{d/2} \exp\{-\frac{\beta}{2}\sum_{i=1}^{n}(y_i^2 - 2w^T\Phi(\mathbf{x}_i)y_i + w^T\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T w) - \frac{\alpha}{2}||w||^2\} \\
&(\frac{\beta}{2\pi})^{n/2}(\frac{\alpha}{2\pi})^{d/2} \exp\{-\frac{\beta}{2}w^T\Phi^T\Phi w - 2w^T\Phi^T y - \frac{\alpha}{2}||w||^2\} \\
&(\frac{\beta}{2\pi})^{n/2}(\frac{\alpha}{2\pi})^{d/2} \exp\{-\frac{1}{2}w^T(\beta\Phi^T\Phi + \alpha I)w - 2w^T\Phi^T y\}
\end{aligned}$$

The above can writen:

$$\exp\{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\} \tag{11}$$

where

$$\begin{aligned}
\Sigma^{-1} &= \beta\Phi^T\Phi + \alpha I \\
\mu &= \beta\Sigma\Phi y
\end{aligned}$$

So the posterior distribution of w is a Normal distribution $\mathcal{N}(w, |\mu, \Sigma^{-1})$, and μ, Σ are given by the above.

1.3.2 Predictive Distribution for New Data

For predicting the new data using the posterior distribution of w, we integrate

$$p(y|x) = \int p(y|x; w)p(w|D)dw \tag{12}$$

we know

$$\begin{aligned}
&p(w|D) \propto \mathcal{N}(w|\mu, \Sigma) \\
&p(y|x; w) \sim \mathcal{N}(w^T\Phi(x), \beta^{-1})
\end{aligned}$$

Let's compute mean and variance of $p(y|x)$ in the following way, we have $y = w^T\Phi(x) + \epsilon$, So the mean of y

$$\begin{aligned}
\hat{u} &= E[w^T\Phi(x) + \epsilon] \\
&= E[w^T\Phi(x)] + E[\epsilon] \\
&= E[w]^T\Phi(x) \\
&= \mu^T\Phi(x)
\end{aligned}$$

the variance of y,

$$\begin{aligned}
\hat{\Sigma} &= var(w^T\Phi(x) + \epsilon) \\
&= var(w^T\Phi(x)) + var(\epsilon) \\
&= E[(w^T\Phi(x) - \mu^T\Phi(x))^2] + \beta^{-1} \\
&= E[(w^T\Phi(x) - \mu^T\Phi(x))^2] + \beta^{-1} \\
&= E[\Phi(x)^T(w - \mu)(w - \mu)^T\Phi(x)] + \beta^{-1} \\
&= \Phi(x)^T E[(w - \mu)(w - \mu)^T]\Phi(x) + \beta^{-1} \\
&= \Phi(x)^T\Sigma\Phi(x) + \beta^{-1}
\end{aligned}$$

So the predictive distribution of y is also a Normal distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$, where $\hat{\mu}, \hat{\Sigma}$ are given in the above.