Machine Learning Basics

Sunghee Yun

December 30, 2019

Contents

1 Optimal Predictor

Consider a regression problem where we predict $Y \in \mathbf{R}^m$ given $X \in \mathbf{R}^n$. We want to design a predictor $g : \mathbf{R}^n \to \mathbf{R}^m$ so that $g(X) \sim Y$ in some statistical sense. We first show that $g(X) = \mathbf{E}(Y|X)$ is the optimal predictor (or estimator) in least-mean-square sense.

We define $g^* : \mathbf{R}^n \to \mathbf{R}^m$ where $g(x) = \mathbf{E}(Y|X = x)$. Then

$$\begin{aligned}
&\mathbf{E}\, \|g(X) - Y\|_2^2 = \mathbf{E}\, \|g(X) - g^*(X) + g^*(X) - Y\|_2^2 \\
&= \quad \mathbf{E}\, \|g(X) - g^*(X)\|_2^2 + \mathbf{E}\, \|g^*(X) - Y\|_2^2 + 2\, \mathbf{E}(g(X) - g^*(X))^T (g^*(X) - Y) \\
&= \quad \mathbf{E}\, \|g(X) - g^*(X)\|_2^2 + \mathbf{E}\, \|g^*(X) - Y\|_2^2 + 2 \mathop{\mathbf{E}}_X \mathop{\mathbf{E}}_Y \left((g(X) - g^*(X))^T (g^*(X) - Y) | X \right) \\
&= \quad \mathbf{E}\, \|g(X) - g^*(X)\|_2^2 + \mathbf{E}\, \|g^*(X) - Y\|_2^2 + 2 \mathop{\mathbf{E}}_X (g(X) - g^*(X))^T \mathop{\mathbf{E}}_Y (g^*(X) - Y|X) \\
&= \quad \mathbf{E}\, \|g(X) - g^*(X)\|_2^2 + \mathbf{E}\, \|g^*(X) - Y\|_2^2 + 2 \mathop{\mathbf{E}}_X (g(X) - g^*(X))^T \left(g^*(X) - \mathbf{E}(Y|X) \right) \\
&= \quad \mathbf{E}\, \|g(X) - g^*(X)\|_2^2 + \mathbf{E}\, \|g^*(X) - Y\|_2^2 \geq \mathbf{E}\, \|g^*(X) - Y\|_2^2.
\end{aligned}$$

Therefore $g^*(X)$ is the optimal predictor for Y in the least-mean-square sense.

2 Bias and Variance

In §1, we proved that $g^*(X) = \mathbf{E}(Y|X)$ is the optimal predictor (or estimator) in the least-mean-square sense. However, unless we have the full knowledge of the joint probability distribution of X and Y, *i.e.*, $p(X, Y)$, or know $\mathbf{E}(Y|X = x)$ as a function of x, it is not possible to obtain g^*.

Here we assume that we obtain the predictor for Y given X from a dataset D where

$$D = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subseteq \mathbf{R}^n \times \mathbf{R}^m.^{1} \tag{1}$$

Now suppose that we have a predictor $g(\cdot; D) : \mathbf{R}^n \to \mathbf{R}^m$, which depends on D. Now let $\mathcal{D}$ denote the random variable for this data set, *i.e.*,

$$\mathcal{D} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\} \subseteq \mathbf{R}^n \times \mathbf{R}^m. \tag{2}$$

[1] Note that strictly speaking, $\mathcal{D}$ is *not* a set since the order of $(x_i, y_i) \in \mathbf{R}^n \times \mathbf{R}^m$ matters, *i.e.*, if the order is changed, we generally have different predictor, and we are allowed to have identical data point. Thus, we should say $\mathcal{D}$ is a (ordered) list of points, $(x_i, y_i) \in \mathbf{R}^n \times \mathbf{R}^m$.

Then the mean square error of this predictor can be decomposed as following.

$$
\begin{aligned}
\mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - Y\|_2^2 &= \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - g^*(X) + g^*(X) - Y\|_2^2 \\
&= \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} (g(X;\mathcal{D}) - g^*(X))^T (g^*(X) - Y) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,\mathcal{D}} \mathop{\mathbf{E}}_{Y} \left((g(X;\mathcal{D}) - g^*(X))^T (g^*(X) - Y) | X, \mathcal{D}\right) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,\mathcal{D}} (g(X;\mathcal{D}) - g^*(X))^T \mathop{\mathbf{E}}_{Y} (g^*(X) - Y | X, \mathcal{D}) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,\mathcal{D}} (g(X;\mathcal{D}) - g^*(X))^T \mathop{\mathbf{E}}_{Y} (g^*(X) - Y | X) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,\mathcal{D}} (g(X;\mathcal{D}) - g^*(X))^T (g^*(X) - \mathop{\mathbf{E}}_{Y} (Y | X)) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) + \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X,\mathcal{D}} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,\mathcal{D}} (g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}))^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X} \mathop{\mathbf{E}}_{\mathcal{D}} \left((g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}))^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)) | X\right) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X} \mathop{\mathbf{E}}_{\mathcal{D}} (g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) | X)^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X} (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) | X)^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2.
\end{aligned}
$$

Note that we use the fact that $\mathbf{E}_{\mathcal{D}}\, g(X;\mathcal{D})$ is a function of X only (hence does not depend on X).

In the last equation, the first term is called the *variance* since it is the expected value (with respect to X) of variance of the predictor, $g(X;\mathcal{D})$, with respect to the dataset, $\mathcal{D}$. It represents the extent to which the prediction varies around its expected value. The second term is the expected value of the square of the bias where the bias is defined to be the difference between the expected value of prediction with respect to dataset and the optimal prediction. The second term itself is sometimes called *bias*. The third term is called *noise* since it is caused by the intrinsic noise residing in Y which cannot be reduced even with the optimal predictor (in least-mean-square sense).

The following equation summarizes these three quantities.

$$\begin{aligned}&\mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - Y\|_2^2 \\ &= \underbrace{\mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2}_{\text{variance}} + \underbrace{\mathop{\mathbf{E}}_{X} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - g^*(X)\|_2^2}_{\text{bias}} + \underbrace{\mathop{\mathbf{E}}_{X,Y} \|g^*(X) - Y\|_2^2}_{\text{noise}}.\end{aligned} \tag{3}$$

In general, we do not know the optimal predictor; if we knew, we would not need to train our in the first place. Thus we can only estimate $g(X;\mathcal{D}) - \mathbf{E}_{\mathcal{D}}\, g(X;\mathcal{D})$ and $\mathbf{E}_{\mathcal{D}}\, g(X;\mathcal{D}) - Y$. The mean square error can also be expressed in these two quantities as follows.

$$\begin{aligned}
\mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - Y\|_2^2 &= \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) + \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y\|_2^2 \\
&= \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,Y,\mathcal{D}} (g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}))^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,Y} \mathop{\mathbf{E}}_{\mathcal{D}} \left((g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}))^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y) | X, Y \right) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,Y} \mathop{\mathbf{E}}_{\mathcal{D}} (g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) | X, Y)^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y\|_2^2 \\
&\quad +2 \mathop{\mathbf{E}}_{X,Y} (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}))^T (\mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y) \\
&= \mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2 + \mathop{\mathbf{E}}_{X,Y} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y\|_2^2.
\end{aligned}$$

Equating the last equation with (3) yields

$$\mathop{\mathbf{E}}_{X,Y,\mathcal{D}} \|g(X;\mathcal{D}) - Y\|_2^2 = \underbrace{\mathop{\mathbf{E}}_{X,\mathcal{D}} \|g(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D})\|_2^2}_{\text{variance}} + \underbrace{\mathop{\mathbf{E}}_{X,Y} \| \mathop{\mathbf{E}}_{\mathcal{D}} g(X;\mathcal{D}) - Y\|_2^2}_{\text{bias + noise}} \tag{4}$$

Therefore in reality, we can only obtain the sum of the bias and noise (not separately) unless we know the quantity of the noise.

3 Maximum Likelihood Estimation

Suppose that $X \in \mathbf{R}^n$ and $Y \in \mathbf{R}^m$ are random variables representing inputs (or independent variables or predictors or features) and outputs (or dependent variables or responses). We want to find a parameterized model to predict Y from X.

We consider the parameter $\theta \in \Theta$ where $\Theta \subset \mathbf{R}^l$ is the set of all possible parameter values, and the model, $g : \mathbf{R}^n \times \mathbf{R}^l \to \mathbf{R}^m$, such that $g(X;\theta)$ is close to Y in some statistical sense.

We further assume that the conditional probability of Y given $g(X;\theta)$ can be characterized by $\beta \in \mathcal{B}$, *i.e.*, $p(Y|g(X;\theta))$ is a function of β where $\mathcal{B}$ is the set of all possible values for β. To express that $p(Y|g(X;\theta))$ is a function of β, we will use a notation, $p(Y|X;\theta,\beta)$, for $p(Y|g(X;\theta))$.

Now suppose that we have observed N independent data sample, $\{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathbf{R}^n$ and $y_i \in \mathbf{R}^m$. We want to find $\theta \in \Theta$ which maximizes the probability of this event, *i.e.*,

$$p\left((X_1, Y_1) = (x_1, y_1), \ldots (X_N, Y_N) = (x_N, y_N)\right) \tag{5}$$

For notational convenience, we define two random variables, $\tilde{X} \in \mathbf{R}^{n\times N}$ and $\tilde{Y} \in \mathbf{R}^{m\times N}$, such that

$$\tilde{X} = \left[\begin{array}{ccc} X_1 & \ldots & X_N \end{array}\right] \tag{6}$$

and

$$\tilde{Y} = \left[\begin{array}{ccc} Y_1 & \ldots & Y_N \end{array}\right]. \tag{7}$$

We also defined $\tilde{x} \in \mathbf{R}^{n\times N}$ and $\tilde{y} \in \mathbf{R}^{m\times N}$, such that

$$\tilde{x} = \left[\begin{array}{ccc} x_1 & \ldots & x_N \end{array}\right] \tag{8}$$

and

$$\tilde{y} = \left[\begin{array}{ccc} y_1 & \ldots & y_N \end{array}\right]. \tag{9}$$

Then (5) becomes

$$p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}). \tag{10}$$

Bayes's theorem and the independence among (X_i, Y_i) imply

$$\begin{aligned} p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) &= \prod_{i=1}^{N} p(X_i = x_i, Y_i = y_i) \\ &= \prod_{i=1}^{N} p(X_i = x_i | Y_i = y_i; \theta, \beta) p(X_i = x_i) \\ &\propto \prod_{i=1}^{N} p(X_i = x_i | Y_i = y_i; \theta, \beta). \end{aligned}$$

Here $p(X_i = x_i | Y_i = y_i; \theta, \beta)$ is called *likelihood function.*

The maximum likelihood estimation (MLE) (or learning) is to find θ (and β) which maximizes this probability. Since the log function is a strictly increasing function, taking log on this probability does give us the same results when maximizing it.

$$\log p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) = \sum_{i=1}^{N} \log p(X_i = x_i | Y_i = y_i; \theta, \beta) + \text{constant} \tag{11}$$

In many cases, this log form is preferable due to its numerical property.

The optimal value of θ for the maximum likelihood estimation (when β is fixed) can be written as

$$\theta_{\mathrm{ML}}(\beta) = \operatorname*{argmax}_{\theta\in\Theta} \sum_{i=1}^{N} \log p(X_i = x_i | Y_i = y_i; \theta, \beta). \tag{12}$$

Note that the solution is a function of β when it is fixed. We sometimes want to find the optimal value for β, too. In this case, we have

$$(\theta_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \operatorname*{argmax}_{(\theta,\beta)\in\Theta\times\mathcal{B}} \sum_{i=1}^{N} \log p(X_i = x_i | Y_i = y_i; \theta, \beta). \tag{13}$$

Note that both of these are point estimation, *i.e.*, we want to find one value for θ (and β) to maximize the log likelihood function.

Now the MLE model is given by $g(\cdot; \theta_{\mathrm{ML}}) : \mathbf{R}^n \to \mathbf{R}^m$, *i.e.*, given $x \in \mathbf{R}^n$, we can predict $y \in \mathbf{R}^m$ by $g(x; \theta_{\mathrm{ML}})$.

4 Maximum a Posteriori Estimation

In MLE, we regard θ as a deterministic variable, which is called Frequentist perspective.

However, we sometimes have prior knowledge of the distribution of θ (or belief about θ). In this situation, we want to find probability distribution of θ after observing some evidence, *e.g.*, the N data sample we have observed, $\{(x_i, y_i)\}_{i=1}^N$.

The natural way of finding the distribution of θ after observing this evidence is to evaluate the condition probability of θ given $\tilde{x}$ and $\tilde{y}$, *i.e.*,

$$p(\theta|\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}). \tag{14}$$

Note that θ is considered to be a *random* variable unlike in MLE case.

Here we introduce a new parameter $\alpha \in \mathcal{A}$ that characterizes the distribution of θ where $\mathcal{A}$ is the set of all the possible values of α.

Since the data samples are assumed to be independent, the Bayes' theorem implies that

$$\begin{aligned}
&p(\theta|\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}; \alpha, \beta) = p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}|\theta; \alpha, \beta)p(\theta; \alpha, \beta)/p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) \\
&\quad = \quad p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}|\theta; \beta)p(\theta; \alpha)/p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) \\
&\quad \propto \quad p(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}|\theta; \beta)p(\theta; \alpha) \\
&\quad = \quad p(\theta; \alpha) \prod_{i=1}^{N} p(X_i = x_i, Y_i = y_i|\theta; \beta) \\
&\quad = \quad p(\theta; \alpha) \prod_{i=1}^{N} p(Y_i = y_i|X_i = x_i, \theta; \beta)p(X_i = x_i|\theta; \beta) \\
&\quad = \quad p(\theta; \alpha) \prod_{i=1}^{N} p(Y_i = y_i|X_i = x_i, \theta; \beta)p(X_i = x_i) \\
&\quad \propto \quad p(\theta; \alpha) \prod_{i=1}^{N} p(Y_i = y_i|X_i = x_i, \theta; \beta)
\end{aligned}$$

The maximum a posteriori (MAP) estimation is to find θ (when β is fixed) which maximizes this posteriori probability. Thus, the MAP solution can be expressed as

$$\theta_{\mathrm{MAP}}(\alpha, \beta) = \operatorname*{argmax}_{\theta \in \Theta} \left(\log p(\theta; \alpha) + \sum_{i=1}^{N} \log p(Y_i = y_i|X_i = x_i, \theta; \beta) \right) \tag{15}$$

where $p(Y_i = y_i|X_i = x_i, \theta; \beta)$ is called *likelihood function*.

Note the difference between the likelihood function, $p(Y_i = y_i|X_i = x_i; \theta, \beta)$, in (12) and the likelihood function, $p(Y_i = y_i|X_i = x_i, \theta; \beta)$, in (15) where θ in (12) is an optimization variable and θ in (15) is a variable for a random variable.

Now the MAP model is given by $g(\cdot; \theta_{\mathrm{MAP}}) : \mathbf{R}^n \to \mathbf{R}^m$, *i.e.*, given $x \in \mathbf{R}^n$, we can predict $y \in \mathbf{R}^m$ by $g(x; \theta_{\mathrm{MAP}})$.

5 Bayesian prior update

Note that both MLE and MAP estimation is a point estimation, *i.e.*, to find one solution that maximizes some probability.

However, the posterior probability $p(\theta|\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}; \alpha, \beta)$ can be used to update the prior probability.

In Bayesian probability theory, if the posterior distributions are in the same probability distribution family as the prior probability distribution, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood function.

In this case, we can update the prior by updating α. Suppose that we have initial prior, $\alpha^{(0)} \in \mathcal{A}$. After we observing first data samples, $(\tilde{x}^{(1)}, \tilde{y}^{(1)})$, we evaluate the posterior probability $p(\theta | \tilde{X} = \tilde{x}^{(1)}, \tilde{Y} = \tilde{y}^{(1)}; \alpha^{(0)}, \beta)$ which can be characterized by some θ^+ due to conjugate distribution assumption. We let $\alpha^{(1)}$ be this updated parameter. We can repeat this process every time we observe new set of data samples. This process can be expressed as

$$\alpha^{(0)} \xrightarrow{\tilde{x}^{(1)}, \tilde{y}^{(1)}} \alpha^{(1)} \xrightarrow{\tilde{x}^{(2)}, \tilde{y}^{(2)}} \alpha^{(2)} \xrightarrow{\tilde{x}^{(2)}, \tilde{y}^{(2)}} \alpha^{(3)} \cdots \tag{16}$$

We can see this process as the one similar to what happens in our brain. A simplified version of explaining human learning process is to update its prior knowledge whenever it observes new evidence. For example, if one has observed that when it rains, the temperature is high, her prior knowledge is that

$$\text{rain} \rightarrow \text{high temperature.} \tag{17}$$

However, if she experiences a rainy day with low temperature, her knowledge is updated as something like

$$\text{rain} \rightarrow \begin{cases} \text{high temperature} & \text{with probability } 0.9 \\ \text{low temperature} & \text{with probability } 0.1 \end{cases} \tag{18}$$

Now this becomes her new prior. If she observes more rainy cold days, her knowledge is updated as something like

$$\text{rain} \rightarrow \begin{cases} \text{high temperature} & \text{with probability } 0.7 \\ \text{low temperature} & \text{with probability } 0.3 \end{cases} \tag{19}$$

This analogy tells why the prior in Bayesian statistics is sometimes called Bayesian belief. This prior belief is something that can be constantly updated with new evidence.

6 Predictive Distribution

If θ is fixed, the probability of $y \in \mathbf{R}^m$ given $x \in \mathbf{R}^n$, $p(Y = y | X = x; \theta, \beta)$, is solely characterized by $\beta \in \mathcal{B}$. However, if we regard θ as a random variable with distribution characterized by α, the probability of $y \in \mathbf{R}^m$ given $x \in \mathbf{R}^n$ can be evaluated by

$$\begin{aligned} p(Y = y | X = x; \alpha, \beta) &= \int_{\theta \in \Theta} p(Y = y, \theta | X = x; \alpha, \beta) d\theta \\ &= \int_{\theta \in \Theta} p(Y = y | X = x, \theta; \alpha, \beta) p(\theta | X = x; \alpha, \beta) d\theta \\ &= \int_{\theta \in \Theta} p(Y = y | X = x, \theta; \beta) p(\theta; \alpha) d\theta, \end{aligned}$$

i.e.,

$$p(Y = y | X = x; \alpha, \beta) = \int_{\theta \in \Theta} p(Y = y | X = x, \theta; \beta) p(\theta; \alpha) d\theta. \tag{20}$$

This is called predictive distribution. This Bayesian statistical predictor, if (20) can be efficiently evaluated, not only gives the point estimation, *e.g.*, by mean or mode, but also the distribution of the output. One advantage of this approach is that we can evaluate the confidence interval