Model Building through Regression

Sanjay Singh*†

*Department of Information and Communication Technology
Manipal Institute of Technology, MAHE
Manipal-576104, INDIA
sanjay.singh@manipal.edu

†Centre for Artificial and Machine Intelligence (CAMI)
MAHE, Manipal-576104, INDIA

February 11, 2019

Introduction

- Linear regression, a special form of function approximation, to model a given set of random variables
- In regression, we typically find the following scenarios:
  - One of the rv is considered to be of particular interest, that random variable is referred to as a dependent variable or response
  - The remaining rvs are called independent variabes, or regressors
  - The dependence of response on the regressors includes an additive error term to account for uncertainities in the manner in which the dependence is formulated
  - The error term is called the expectational error or explanational error
- Classes of regression model: linear and nonlinear

Linear Regression Model

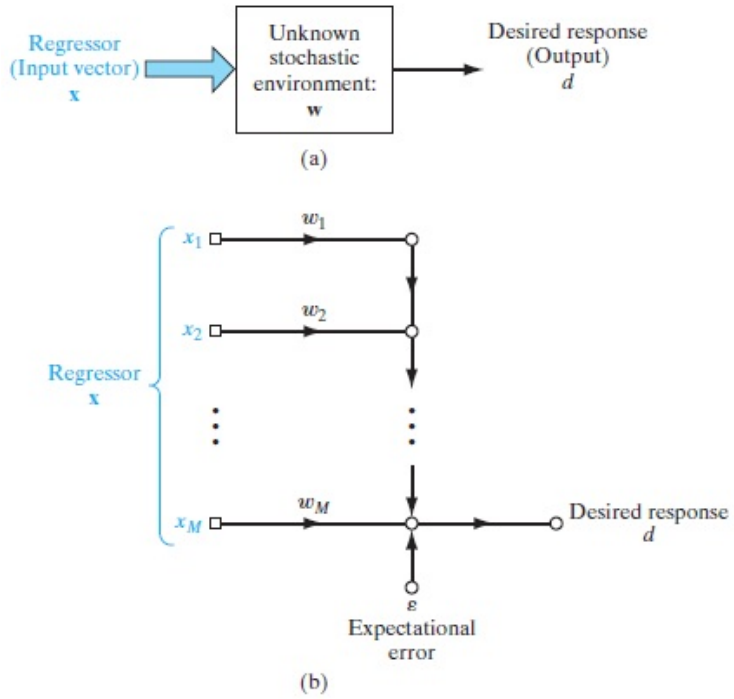

Figure 1: (a) Unknown stationary stochastic environment. (b) Linear regression model of the environment

- The environment is probed by applying a set of inputs, constituting the regressor

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$$

- The desired response, d is considered to be scalar for the convenience
- We don't know the functional dependence of d on the regressor $\mathbf{x}$, we propose a linear regression model, parameterized as

$$d = \sum_{j=1}^{m} w_j x_j + \epsilon$$

where $w_1, w_2, \ldots, w_m$ denote a set of fixed, but unknown parameters, meaning that the environment is stationary

- The term ϵ, represents the expactational error of the model, accounts for our ignorance about the environment
- Using matrix notation, we may write $d = \mathbf{w}^T\mathbf{x} + \epsilon$

- The common dimension, m is called the model order
- With environment being stochastic, it follows that $\mathbf{x}$, $\mathbf{w}$, and the expactational error ϵ are sample values (i.e., single-shot realizations) of the random vector $\mathbf{X}, D$, and E
- With such a stochastic setting as the background, the problem of interest may be stated as follows

  Given the joint statistics of the regressor $\mathbf{X}$, and the corresponding response D, estimate the unknown parameter vector $\mathbf{W}$
- By joint statistics, we mean the following
  - the correlation matrix of the regressor $\mathbf{X}$;
  - the variance of the desired response D;
  - the cross-correlation vector of the regressor $\mathbf{X}$ and the desired response D

  It is assumed that the mean of both $\mathbf{X}$ and D are zero

Maximum a posteriori estimation of parameter vector

The Bayesian paradigms provides a powerful approach for addressing and quantifying the uncertainty that surrounds the choice of parameter vector $\mathbf{w}$ in the linear regression model

- The regressor $\mathbf{X}$ acts as the ”excitation,” bearing no relation to the parameter vector $\mathbf{w}$
- Information about the unknown parameter vector $\mathbf{W}$ is contained in the desired response D that acts as the observable of the environment
- Accordingly, we focus on the joint probability density function of $\mathbf{W}$, and D conditional on $\mathbf{X}$, denoting as $p_{w,D|x}(w, d|x)$

- From probability theory, $p_{w,x|D}(w,d|x) = p_{w|D,x}(w|d,x)p_D(d)$
- Equivalently, $p_{w,x|D}(w,d|x) = p_{D|w,x}(d|w,x)p_w(w)$
- We may write,

$$p_{w|D,x}(w|d,x) = \frac{p_{D|w,x}(d|w,x)p_w(w)}{p_D(d)} \quad p_D(d) \neq 0$$

- The last equation is a special form of Bayes theorem, it embodies four density functions, characterized as
  - **Observation density**: $p_{D|\mathbf{w},\mathbf{x}}(d|\mathbf{w},\mathbf{x})$-observation of the environmental response d due to the regressor $\mathbf{x}$, given the parameter vector $\mathbf{w}$
  - **Prior**: $p_w(w)$ referring to information about $\mathbf{w}$, prior to any observation made on the environment (hence forth, prior will be denoted as $\pi(w)$)
  - **Posterior density**: $p_{\mathbf{w}|D,\mathbf{x}}(\mathbf{w}|d,\mathbf{x})$, referring to the parameter vector $\mathbf{w}$ after the observation of the environment has been completed (hereafter, posterior will be denoted as $\pi(\mathbf{w}|d,\mathbf{x})$
  - **Evidence**: $p_D(d)$, referring to the information contained in the response d for statistical analysis

- The observation density $p_{D|\mathbf{w},\mathbf{x}}(d|\mathbf{w},\mathbf{x})$ is commonly reformulated as the likelihood function defined by

$$l(\mathbf{w}|d,\mathbf{x}) = p_{D|\mathbf{w},\mathbf{x}}(d|\mathbf{w},\mathbf{x})$$

- As far as the estimation of $\mathbf{w}$ is concerned, the evidence $p_D(d)$ in the denominator in Bayes relation plays the role of a normalizing constant
- So we may express the Bayes relation here as
  The posterior density of the vector $\mathbf{w}$ parameterizing the regression model is proportional to the product of likelihood function and the prior

$$\pi(\mathbf{w}|d,\mathbf{x}) \propto l(\mathbf{w}|d,\mathbf{x})\pi(\mathbf{w})$$

- The likelihood function $l(\mathbf{w}|d,\mathbf{x})$ provides the basis for the maximum-likelihood (ML) estimate of the parameter vector $\mathbf{w}$, as shown by

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} l(\mathbf{w}|d,\mathbf{x})$$

- For more profound estimate of the parameter vector $\mathbf{w}$, we look to the posterior density $\pi(\mathbf{w}|d, \mathbf{x})$
- We define the **maximum a posterior (MAP) estimate** of $\mathbf{w}$ by the formula

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} \pi(\mathbf{w}|d, \mathbf{x})$$

- MAP estimator is more profound than the ML estimator due to following reasons:
  1. The Bayesian paradigm for parameter estimation rooted in the Bayes' theorem and exemplified by the MAP estimator exploits all the conceivable information about the parameter $\mathbf{w}$. In contrast, the ML estimator lies on the fringe of the Bayesian paradigm, ignoring the prior
  2. ML estimator relies soley on the observation model $(d, \mathbf{x})$ and may lead to a nonunique solution. To enforce uniqueness and stability on the solution, the prior $\pi(\mathbf{w})$ has to be incorporated into the formulation of the estimator, which is done in the MAP estimator

Parameter estmation in Gaussian environment

- Consider the training set $\mathcal{T} = \{\mathbf{x}_i, d_i\}_{i=1}^{N}$
- To proceed with the task of parameter estimation, we make the following assumptions
  1. **Statistical independence and identical distribution**-the N examples are statistically independent and identically distributed (iid)
  2. **Gaussianity**-the environment, responsible for generation of training sample $\mathcal{T}$ is Gaussian $(e_i \sim \mathcal{N}(0, \sigma^2))$
  3. **Stationarity** -the environment is stationary, which means that the parameter vector $\mathbf{w}$ is fixed, but unknown, throughout the N trials of the experiment

- For the ith trial of the experiment performed on the environment, we have

$$d_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i, \quad i = 1, 2, \ldots, N$$

where $d_i, \mathbf{x}_i$, and ϵ_i are sample values of the random variables $D, \mathbf{X}$, and E

- $\mathbb{E}$ denotes the statistical expectation operator
- Due to Gaussianity assumption, $\mathbb{E}[E_i] = 0, \forall i$, and

$$var[E_i] = \mathbb{E}[E_i^2] = \sigma^2, \ \forall i$$

- For a given regressor $\mathbf{x}_i$

$$\begin{aligned} \mathbb{E}[D_i] &= \mathbb{E}[\mathbf{w}^T \mathbf{x}_i + \epsilon_i] \\ &= \mathbb{E}[\mathbf{w}^T \mathbf{x}_i] + \mathbb{E}[\epsilon_i] \\ &= \mathbf{w}^T \mathbf{x}_i, \ i = 1, 2, \ldots, N \\ var[D_i] &= \mathbb{E}[(D_i - \mathbb{E}[D_i])^2] \\ &= \mathbb{E}[E_i^2] \\ &= \sigma^2 \end{aligned}$$

- The likelihood function for the ith trial is defined as

$$l(\mathbf{w}|d_i, \mathbf{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(d_i - \mathbf{w}^T \mathbf{x}_i)^2\right), \ i = 1, 2, \ldots, N$$

- Invoking the iid characterization of the N trials of the experiment on the environment, we express the overall likelihood function for the experiment as

$$\begin{aligned} l(\mathbf{w}|d, \mathbf{x}) &= \prod_{i=1}^{N} l(\mathbf{w}|d_i, \mathbf{x}_i) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(d_i - \mathbf{w}^T \mathbf{x}_i)^2\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^N} \prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}(d_i - \mathbf{w}^T \mathbf{x}_i)^2\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i - \mathbf{w}^T \mathbf{x}_i)^2\right) \end{aligned}$$

- $l(\mathbf{w}|d, \mathbf{x})$ accounts for the total empirical knowledge about the weight vector $\mathbf{w}$ contained in the training sample $\mathcal{T}$

- The other source of information that to be accounted for is contained in the prior $\pi(\mathbf{w})$
- Following the zero-mean assumption for $\mathbf{w}$, and following iid characterization of the m elements of $\mathbf{w}$, we write

$$\begin{aligned}\pi(\mathbf{w}) &= \prod_{i=1}^{m} \pi(w_i) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{w_i^2}{2\sigma_w^2}\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma_w)^m} \prod_{i=1}^{m} \exp\left(-\frac{w_i^2}{2\sigma_w^2}\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma_w)^m} \exp\left(-\frac{1}{2\sigma_w^2} \sum_{i=1}^{m} w_i^2\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma_w)^m} \exp\left(-\frac{1}{2\sigma_w^2} ||\mathbf{w}||^2\right)\end{aligned}$$

- Substituting for $l(\mathbf{w}|d,\mathbf{x})$, and $\pi(\mathbf{w})$, in $\pi(\mathbf{w}|d,\mathbf{x}) \propto l(\mathbf{w}|d,\mathbf{x})\pi(\mathbf{w})$, yields

$$\begin{aligned}\pi(\mathbf{w}|d,\mathbf{x}) \propto & \frac{1}{(\sqrt{2\pi}\sigma)^N} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (d_i - \mathbf{w}^T\mathbf{x}_i)^2\right) \times \\ & \frac{1}{(\sqrt{2\pi}\sigma_w)^m} \exp\left(-\frac{1}{2\sigma_w^2} ||\mathbf{w}||^2\right) \\ & \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (d_i - \mathbf{w}^T\mathbf{x}_i)^2 - \frac{1}{2\sigma_w^2} ||\mathbf{w}||^2\right]\end{aligned}$$

- Now, we can apply MAP formula, $\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \log(\pi(\mathbf{w}|d,\mathbf{x}))$
- On substituting, we get

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (d_i - \mathbf{w}^T\mathbf{x}_i)^2 - \frac{1}{2\sigma_w^2} ||\mathbf{w}||^2\right]$$

- $\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \left[-\frac{1}{2} \sum_{i=1}^{N} (d_i - \mathbf{w}^T\mathbf{x}_i)^2 - \frac{\lambda}{2} ||\mathbf{w}||^2\right]$ where $\lambda = \frac{\sigma^2}{\sigma_w^2}$

- We define the quadratic function

$$\mathcal{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(d_i - \mathbf{w}^T\mathbf{x}_i)^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

- Maximization of the argument in $\hat{\mathbf{w}}_{MAP}$ wrt $\mathbf{w}$ is equivalent to minimization of $\mathcal{E}(\mathbf{w})$
- The optimum estimate $\hat{\mathbf{w}}_{MAP}$ is obtained by differentiating the function $\mathcal{E}(\mathbf{w})$ wrt $\mathbf{w}$ and setting the results to zero, we obtain the desired MAP estimate of the $m \times 1$ parameter vector as

$$\hat{\mathbf{w}}_{MAP}(N) = [\mathbf{R}_{xx}(N) + \lambda\mathbf{I}]^{-1}\mathbf{r}_{dx}(N)$$

Derivation detail

- $\mathcal{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(d_i - \mathbf{w}^T\mathbf{x}_i)^2 + \frac{\lambda}{2}||\mathbf{w}||^2$
- Lets consider one example case

$$\begin{aligned}\mathcal{E}(w) &= \frac{1}{2}(d_i - \mathbf{w}^T\mathbf{x}_i)^T(d_i - \mathbf{w}^T\mathbf{x}_i) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \\ &= \frac{1}{2}\left[(d_i^T - \mathbf{x}_i^T\mathbf{w})(d_i - \mathbf{w}^T\mathbf{x}_i) + \lambda\mathbf{w}^T\mathbf{w}\right] \\ &= \frac{1}{2}\left[d_i^T d_i - d_i^T\mathbf{w}^T\mathbf{x}_i - \mathbf{x}_i^T\mathbf{w}d_i + \mathbf{x}_i^T\mathbf{w}\mathbf{w}^T\mathbf{x}_i + \lambda\mathbf{w}^T\mathbf{w}\right] \\ &= \frac{1}{2}\left[d_i^T d_i - 2(\mathbf{x}_i d_i)^T\mathbf{w} + \mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} + \lambda\mathbf{w}^T\mathbf{w}\right]\end{aligned}$$

- Now lets compute $\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w})$

$$\begin{aligned}\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w}) &= \nabla_{\mathbf{w}}\left(\frac{1}{2}\left[d_i^T d_i - 2(\mathbf{x}_i d_i)^T\mathbf{w} + \mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} + \lambda\mathbf{w}^T\mathbf{w}\right]\right) \\ &= \frac{1}{2}\left(\nabla_{\mathbf{w}}(d_i^T d_i) - 2\nabla_{\mathbf{w}}(\mathbf{x}_i d_i)^T\mathbf{w} + \nabla_{\mathbf{w}}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} + \nabla_{\mathbf{w}}\lambda\mathbf{w}^T\mathbf{w}\right) \\ &= \frac{1}{2}\left(0 - 2\mathbf{x}_i d_i + 2\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} + \lambda\mathbf{w}\right)\end{aligned}$$

- Now $\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w}) = 0$, yields

$$-\mathbf{x}_i d_i + \mathbf{x}_i \mathbf{x}_i^T \mathbf{w} + \lambda \mathbf{w} = 0$$

$$\mathbf{w}[\mathbf{x}_i \mathbf{x}_i^T + \lambda \mathbf{I}] = \mathbf{x}_i d_i$$

$$\mathbf{w} = (\mathbf{x}_i \mathbf{x}_i^T + \lambda \mathbf{I})^{-1} \mathbf{x}_i d_i$$

$$= \left[\underbrace{\sum_{i=1}^{N} \sum_{j}^{N} \mathbf{x}_i \mathbf{x}_i^T}_{\mathbf{R}_{xx}(N)} + \lambda \mathbf{I} \right]^{-1} \times \underbrace{\sum_{i=1}^{N} \mathbf{x}_i d_i}_{\mathbf{r}_{dx}}$$

- $\hat{\mathbf{w}}_{MAP}(N) = [\mathbf{R}_{xx}(N) + \lambda \mathbf{I}]^{-1} \mathbf{r}_{dx}(N)$

  where $\mathbf{R}_{xx}(N) = -\sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{x}_i \mathbf{x}_j^T$ is the time-average $m \times m$ correlation matrix applied to the environment on ith and jth experimental trials
- The time-average $m \times 1$ cross-correlation vector of the regressor $\mathbf{x}$ and the desired response d is defined as

$$\mathbf{r}_{dx}(N) = -\sum_{i=1}^{N} \mathbf{x}_i d_i$$

- $\mathbf{R}_{xx}(N)$, and $\mathbf{r}_{dx}(N)$ are both averaged over all the N examples of the training sample $\mathcal{T}$-hence the use of the term *time-averaged*

- Suppose we assign a large value to σ_w^2, saying that prior distribution of each element of $\mathbf{w}$ is uniform over a wide range of possible values
- Under this condition, the parameter λ is zero, hence the formula for $\hat{\mathbf{w}}_{MAP}$ reduces to that of ML estimate,

$$\hat{\mathbf{w}}_{ML} = \hat{\mathbf{R}}_{xx}^{-1}(N)\hat{\mathbf{r}}_{dx}(N)$$

  which supports the point-ML estimator relies solely on the observation model exemplified by the training sample $\mathcal{T}$
- The ML estimator $\hat{\mathbf{w}}_{ML}$ is an unbiased estimator, that is

$$\lim_{N\to\infty} \hat{\mathbf{w}}_{ML}(N) = \mathbf{w}$$

- In contrast, the MAP estimator is biased estimator, In improving the stability of the maximum likelihood estimator through the use of regularization (i.e., incorporation of prior knowledge), the resulting maximum a posteriori estimator becomes biased
- We need to have tradeoff between stability and bias

Bias

- The bias of an estimator is defined as

$$\text{bias}(\hat{\mathbf{w}}_N) = \mathbb{E}(\hat{\mathbf{w}}_N) - \mathbf{w}$$

  where the expectation is over the data, and $\mathbf{w}$ is true underlying value of $\mathbf{w}$ used to define the data generating distribution
- An estimator $\hat{\mathbf{w}}_N$ is said to be unbiased if $\text{bias}(\hat{\mathbf{w}}_N) = 0$, which implies that $\mathbb{E}(\hat{\mathbf{w}}_N) = \mathbf{w}$
- An estimator is said to be asymptotically unbiased if $\lim_{N\to\infty} \mathbb{E}(\hat{\mathbf{w}}_N) = \mathbf{w}$