Introduction

- Regression is special form of function approximation that is used to formulate mathematical relationship between a dependent variable and a set of independent variables that also includes error term to account for uncertainties.
- There are two classes of regression models: *linear and nonlinear*. In *linear regression models, the dependence of the dependent variable (response variable) on the independent variable (regressors) is defined by a linear function.* Statistical analysis of linear regression is mathematically tractable. On the other hand, in *nonlinear regression models, this dependence is defined by a nonlinear function, hence* the mathematical difficulty in their analysis.

Linear Regression Model

- Consider an unknown stochastic environment that is probed by applying a set of inputs, constituting the regressor.

$$x = [x_1, x_2, ..., x_m]$$

- The resulting output of the environment, denoted by d, constitutes the corresponding response. We do not know the functional dependence of the response d on the regressor x, so we consider a linear regression model, parameterized as below:

$$d = \sum_{i=1}^{m} w_i x_i + \in$$

Linear Regression Model

- In the above equation w_i denote a set unknown, parameters. The additive term ε, representing the expectational error of the model, accounts for our ignorance about the environment.
- Using matrix notation, we may rewrite above equation in the compact form as below

$$d = w^T x + \in$$

Bayes Rule Revisted

$$\underbrace{P(A|B)}_{Posterior} = \frac{\overbrace{P(B|A)}^{Likelihood}\overbrace{P(A)}^{Prior}}{P(B)}$$

MLE Parameter Estimation

- Maximum likelihood estimation (MLE) is a method that determines values for the parameters of a model.
- The parameter values are found such that they maximize the likelihood that results in the best fit for the joint probability of the given data sample.
- Let D denotes the set of observed values and w denotes the parameter to be estimated. Now, MLE can be stated as:

$$w_{MLE} = \arg\max\ p(D \mid w)$$

$$\Rightarrow w_{MLE} = \arg\max \prod_i p(d_i \mid w)$$

MLE Parameter Estimation

- Maximizing a function is equal to maximizing the log of that function. So we can take log while maximizing likelihood. Hence,

$$w_{MLE} = \arg\max\left(\log\left(\prod_i p(d_i \mid w)\right)\right)$$

$$\Rightarrow w_{MLE} = \arg\max \sum_i \log(p(d_i \mid w))$$

MAP Parameter Estimation

- An alternative estimator is the Maximum a Priori (MAP) estimator, which finds the parameter w that maximizes the posterior.
- Let D denotes the set of observed values and w denotes the parameter to be estimated. Now, MAP can be stated as:

$$w_{MAP} = \arg\max\ p(w \mid D)$$

$$= \arg\max\ p(D \mid w).P(w)$$

$$\Rightarrow w_{MAP} = \arg\max \prod_i p(d_i \mid w).P(w)$$

MAP Parameter Estimation

- Like the case of MLE, we can take log while maximizing posterior. Hence,

$$w_{MAP} = \arg\max\left(\log\left(\prod_i p(d_i \mid w).P(w)\right)\right)$$

$$\Rightarrow w_{MAP} = \arg\max\left(\log\left(\prod_i p(d_i \mid w)\right) + \log(P(w))\right)$$

$$\Rightarrow w_{MAP} = \arg\max\left(\sum_i \log(p(d_i \mid w)) + \log(P(w))\right)$$

MLE vs MAP

- Consider a sequence of N coin tosses (call head = 0, tail = 1) Each outcome x_i is a binary random variable $\in \{0, 1\}$ Assume w to be probability of a head (parameter we wish to estimate).
- Here, likelihood has Bernoulli Distribution. Hence,

$$p(x_i \mid w) = w^{x_i}(1-w)^{1-x_i}$$

- Thus,

$$\sum_i \log(p(x_i \mid w)) = \sum_i x_i \log(w) + (1-x_i)\log(1-w)$$

MLE vs MAP

- Taking derivative of above relation w.r.t. w and equating with zero, we get.

$$w_{MLE} = \frac{\sum_i x_i}{N}$$

- Thus, when a coin is tossed 10 times and there are 7 heads and 3 tails.

$$w = \frac{7}{10} = 0.7$$

- MLE Can be problematic especially when the number of observations is very small (e.g., suppose we only observed heads in a small number of coin-tosses).

MLE vs MAP

- Taking derivative of above relation w.r.t. w and equating with zero, we get.

$$w_{MLE} = \frac{\sum_i x_i}{N}$$

- Thus, when a coin is tossed 10 times and there are 7 heads and 3 tails.

$$w = \frac{7}{10} = 0.7$$

- MLE Can be problematic especially when the number of observations is very small (e.g., suppose we only observed heads in a small number of coin-tosses).

MAP vs MLE

- In case of MAP,

$$w_{MAP} = \arg\max\left(\sum_i \log\big(p(x_i \mid w)\big) + \log\big(P(w)\big) \right)$$

- Where,

$$p(x_i \mid w) = w^{x_i}(1-w)^{1-x_i}$$

- Since $w \in (0,1)$, we assume that prior has Beta distribution. Thus,

$$p(w) = \frac{\lceil(\alpha+\beta)}{\lceil\alpha + \lceil\beta} \; w^{\alpha-1}(1-w)^{\beta-1}$$

MAP vs MLE

- In the above equation α and β are Hyperparameter that can be though as number of expected observations. In the coin toss example, $\alpha-1$ and $\beta-1$ are expected number of heads and tails respectively.
- Ignoring the constant term, log of posterior probability becomes

$$\sum_i \{x_i \log(w) + (1-x_i)\log(1-w)\} + (\alpha-1)\log w + (\beta-1)\log(1-w)$$

- Taking derivative of above equation w.r.t. w and equating it with zero, we get

MLE vs MAP

$$w_{MAP} = \frac{\sum_i x_i + \alpha - 1}{N + \alpha + \beta - 2}$$

- Thus, when a coin is tossed 10 times and there are 7 heads and 3 tails. If expected number of heads and tails are each 50 out of 100 toss.

$$w = \frac{7+50}{10+100} = 0.518$$

Relationship Between MLE and MAP

Comparing the equation of MAP with MLE, we can see that the only difference is that MAP includes prior in the formula, which means that the likelihood is weighted by the prior in MAP.

In the special case when prior follows a uniform distribution. MAP can be written as:

$$w_{MAP} = \arg\max\left(\sum_i \log\left(p(d_i \mid w)\right) + \log\left(P(w)\right)\right)$$

$$= \arg\max\left(\sum_i \log\left(p(d_i \mid w)\right) + cons\tan t\right)$$

$$= \arg\max\left(\sum_i \log\left(p(d_i \mid w)\right)\right) = w_{MLE}$$

Relationship Between MLE and MAP

- Based on the formula above, we can conclude that MLE is a special case of MAP, when prior follows a uniform distribution.

Regularized Least Mean Square Estimation

- Least squares fitting (also called least squares estimation) is a way to find the best fit curve or line for a set of points. In this technique, the sum of squares errors are used to estimate the best fit curve or line.
- Thus, loss function for least square estimation is given by

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}(d_i - y_i)^2$$

Regularized Least Mean Square Estimation

- One problem associated with ordinary LMS estimation is presence of outliers.
- Outliers can have a disproportionate effect if you use the least squares fitting method of finding an equation for a line or curve.
- Outliers naturally have larger errors and will affect the line more than points closer to the line.
- **Regularized least squares** is a way of solving least square regression problems with an extra constraint on the loss function. The constraint is called *regularization.*

Regularized Least Mean Square Estimation

- Above loss function relies solely on the training sample. Minimizing this loss function with respect to w yields a formula for the ordinary least-squares estimator that is identical to the maximum-likelihood estimation, and hence there is possibility of obtaining a solution that lacks uniqueness and stability.
- To overcome this problem, the practice is to expand the loss function by adding a new term as follows:

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}(d_i - y_i)^2 + \frac{\lambda}{2}\|w\|^2$$

Regularized Least Mean Square Estimation

- In the above equation, the scalar λ is referred to as the *regularization parameter.*
- When λ =0, the implication is that we have complete confidence in the observation model exemplified by the training sample. *At the other extreme, when* λ=∞, the implication is that we have no confidence in the observation model.
- In practice, the regularization parameter is chosen somewhere between these two limiting cases.
- We can clearly see that Regularized LMS estimation is identical to the MAP estimation.

LMS Regression

- Given the n data points $(x_1,y_1),(x_2,y_2)\ldots\ldots (x_n, y_n)$, we wish to find the linear model

$$y = \beta_0 + \beta_1 x + e$$

- Here the basic matrices are

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathrm{x} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Thus,

$$y = x\beta + e$$

$$\Rightarrow e = y - x\beta$$

LMS Regression

- Thus, loss function for LMS estimation is given by

$$s = \frac{1}{2}\sum_i e_i^2 = \frac{1}{2}\sum_i e^T e = \frac{1}{2}\sum_i (y - x\beta)^T (y - x\beta)$$

- Simplifying above equation, taking its derivative w.r.t. β and equating it with zero, we get the following solution.

$$\beta = (x^T x)^{-1} x^T y$$

LMS Regression

- Given the following dataset. Fit a straight line through the data points using LMS estimation of parameters. And predict demand for the price 64.

Price(x)	Demand(y)
49	124
69	95
89	71
99	45
109	18

Regularized LMS Regression

- Given the n data points $(x_1,y_1),(x_2,y_2)\ldots\ldots (x_n, y_n)$, we wish to find the linear model

$$y = \beta_0 + \beta_1 x + e$$

- Here the basic matrices are

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathrm{x} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Thus,

$$y = x\beta + e$$

$$\Rightarrow e = y - x\beta$$

Regularized LMS Regression

- Thus, loss function for Regularized LMS estimation is given by

$$s = \frac{1}{2}\sum_i e_i^2 + \frac{\lambda}{2}\|\beta\|^2 = \frac{1}{2}\sum_i e^T e + \frac{\lambda}{2}\|\beta\|^2 = \frac{1}{2}\sum_i (y - x\beta)^T (y - x\beta) + \frac{\lambda}{2}\|\beta\|^2$$

- Simplifying above equation, taking its derivative w.r.t. β and equating it with zero, we get the following solution.

$$\beta = (x^T x + \lambda I)^{-1} x^T y$$

Regularized LMS Regression

- Given the following dataset. Fit a straight line through the data points using Regularized LMS estimation of parameters. And predict demand for the price 64. Assume λ=2.

Price(x)	Demand(y)
49	124
69	95
89	71
99	45
109	18

MLE and MAP Estimation in Regression

- Maximizing log likelihood is equivalent to minimizing mean squared error. Thus, regression with LMS estimation is equivalent to regression with MLE estimation.
- Again, maximizing log of posteriori is equivalent to minimizing regularized mean squared error. Thus, regression with regularized LMS estimation is equivalent to regression with MAP estimation.

Minimum Description Length Principle

- The minimum description length (MDL) criteria states that the best description of the data is given by the model which compresses it the best.
- Put another way, learning a model for the data or predicting it is about capturing the regularities in the data and any regularity in the data can be used to compress it.
- Thus, the more we can compress a data, the more we have learnt about it and the better we can predict it.

Minimum Description Length Principle

- MDL is also connected to Occam's Razor used in machine learning which states that "other things being equal, a simpler explanation is better than a more complex one."
- In MDL, the complexity of a model is interpreted as the length of the code obtained when that model is used to compress the data.
- MDL principle is used for selecting the best statistical model.

Minimum Description Length Principle

- Let h is the hypothesis that is used to describe the data sequence d. Suppose L(h) is the description length of h and L(d|h) is the description length of data given h. Then, MDL is about Minimizing the following sum.

$$L(h,d) = L(h) + L(d \mid h)$$

- MDL principle can be derived from MAP parameter estimation.

Minimum Description Length Principle

- We know that MAP estimation is given by

$$\begin{aligned} h_{MAP} &= \arg\max (p(d|h).p(h)) \\ &= \arg\max (\log (p(d|h)) + \log (p(h))) \\ &= \arg\min (-\log (p(d|h)) - \log (p(h))) \end{aligned}$$

- From information(Shanon) theory, length of h is given by –log(h).
- Thus, above relation can be written as

$$h_{MAP} = \arg\min (length(d|h) + length(h)) = L(p,d)$$