

# Estimation for MPG using Linear regression and Gaussian Process regression

Yota Nonaka

## 1 Explanation of the data

### 1.1 Source

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.[1] The dataset was used in the 1983 American Statistical Association Exposition.

### 1.2 Data set information

Mpg (miles per gallon) is an indicator of fuel consumption used in the United States and the United Kingdom. It indicates how many miles (about 1.6 km) can be traveled with one gallon (about 4.55 liters) of fuel, and it can be said that the larger this value, the better the fuel efficiency.

The goal of this experiment is to predict the value of the fuel economy (mpg) from 3 discrete variables and 5 continuous variables.

### 1.3 Attribute Information

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

The purpose of this survey is to estimate mpg using data from (2) to (9). Note that data (9) is not used for regression. Also, some samples whose horsepower is unknown are removed.

## 2 Methods

In implementing this regression, two different approaches were used. One is Linear Regression (LR) and the other is Gaussian Process Regression (GPR). Also, three kinds of basis functions were used in LR.

### 2.1 Linear Regression

The  $k$ 'th component of  $n$ 'th data  $(\mathbf{x}_n)_k$  is converted to  $\phi_k((\mathbf{x}_n)_k; \alpha_k)$ , using a parameter  $\alpha_k$ . Here,  $\phi$  is a basis function. Note that  $\phi_{k=0} = 1$  is satisfied for all basis functions used here. The predicted value is represented by  $y_n = \sum_k w_k \phi_k((\mathbf{x}_n)_k; \alpha_k)$ . To get optimal weight  $\mathbf{w}$  and parameter  $\alpha$ , the mean square error should be minimized.

### 2.1.1 Identity function

When identity function is used as a basis function,  $\phi_k(\mathbf{x}) = x_k$  holds. Using a hyper parameter  $\lambda$ ,

$$\mathbf{w} = (X^T X + \lambda)^{-1} X^T \mathbf{y}$$

gives optimal weight parameters. The hyper parameter  $\lambda$  for regularization is also used when using other basis functions.

### 2.1.2 Sigmoid function

When sigmoid function is used as a basis function,  $\phi_k(\mathbf{x}) = \frac{1 - e^{-x_k + \alpha_k}}{1 + e^{-x_k + \alpha_k}}$  holds. Using the design matrix  $\Phi$ ,

$$\mathbf{w} = (\Phi^T \Phi + \lambda)^{-1} \Phi^T \mathbf{y}$$

gives optimal weight parameters.

The gradient of the loss function  $L$  for  $\alpha$  is

$$\frac{\partial L}{\partial \alpha_k} = 2w_k \sum_n \frac{\partial \phi_k}{\partial \alpha_k}(\mathbf{x}_n) \left[ -y_n + \sum_j \phi_j(\mathbf{x}_n) w_j \right]$$

Using this,  $\alpha$  is optimized if appropriate initial conditions are given. Note that  $\frac{\partial \phi_k}{\partial \alpha_k} = \frac{1}{2}(\phi_k^2 - 1)$  is useful when sigmoid is used as basis function.

### 2.1.3 Gaussian function

When a Gaussian function is chosen as the basis function, the parameters can be optimized in exactly the same way as the sigmoid function. Since basis function is  $\phi_k(\mathbf{x}) = \exp \left[ -\frac{(x_k - \mu_k)^2}{2\sigma_k^2} \right]$ ,

$$\begin{aligned} \frac{\partial \phi_k}{\partial \mu_k} &= \frac{x_k - \mu_k}{\sigma_k^2} \phi_k \\ \frac{\partial \phi_k}{\partial \sigma_k} &= \frac{(x_k - \mu_k)^2}{\sigma_k^3} \phi_k \end{aligned}$$

can be used to calculate gradient.

## 2.2 Gaussian process regression

In the method of setting the basis function of the LR described above, the correlation between each component of the data is not considered. However, incorporating correlations into basis functions can lead to problems such as too many combinations of data components. Therefore, we decided to use Gaussian Process Regression (GPR) which regress without explicitly specifying the basis functions. The kernel function is Gaussian kernel, and the hyper parameters contained in the kernel were tuned with the verification data.

In GPR, given training data  $\mathbf{x}_n$  and true output  $\mathbf{y}$ , the predicted value for the data  $\mathbf{x}^*$  is  $y^* = \mathbf{k}_*^T K^{-1} \mathbf{y}$ . Here,  $(\mathbf{k}_*)_n = k(\mathbf{x}^*, \mathbf{x}_n)$  and  $K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'})$  are used.

### 3 Experiment

Of all 392 data, 20% was be used as test data. For the remaining 80% data, 5 split cross validation was performed and the average relative error for each split was calculated. Parameters such as  $\alpha$  are optimized for training data, and hyper parameters such as  $\lambda$  for validation data. Finally, test data was used to confirm the relative error of the model considered to be the most accurate.

### 4 Results

Table1 compares the relative error from each model. From this table, it can be read that GPR is the most accurate model in this research. In addition, the standard deviation for cross validation is small, so stability is also guaranteed.

Finally, regression for test data was performed using GPR, and the relative error was 8.69%. Since the test

| Table 1: Comparison of accuracy for models |                          |                           |         |
|--|--------------------------|---------------------------|---------|
| Method                                     | basis function or kernel | average of relative error | (STDEV) |
| Linear Regression                          | id                       | 12.2%                     | 1.2%    |
|  | sigmoid                  | 18.0%                     | 6.5%    |
|  | gaussian                 | 17.2%                     | 7.2%    |
| Gaussian Process                           | RBF                      | 9.5%                      | 1.4%    |

data is also maintained with high accuracy, it is also guaranteed that over-fitting has not occurred.

### 5 Discussion

From Table1, when we compare basis functions of LR, it can be said that function forms like identity functions capture features of mpg regression rather than sigmoids and Gaussians.

We also examined the contribution of each component to mpg, taking into account the typical magnitude of each component of the data, using the weight parameter of LR with identity function. At this time, it was found that the year of manufacture had a largest positive effect on mpg. Next, as a major contribution, the weight of the vehicle had a large negative effect on mpg. These are intuitively quite natural. Other contributions were as large as 10% of the magnitude of former two components.

It can also be said that the idea that each component gives an independent basis function is not sufficient because GPR is the most accurate model. That is, the correlation of each component needs to be taken into consideration in the selection of basis functions. However, as mentioned above, it is necessary to consider a huge number of combinations for that, and since the number of data is not so large, it may be said that the GPR is a practical solution.

### References

- [1] <http://archive.ics.uci.edu/ml/datasets/Auto+MPG>