CO496 Coursework 2

Jinsung Ha

5th November 2018

1 Linear Regression

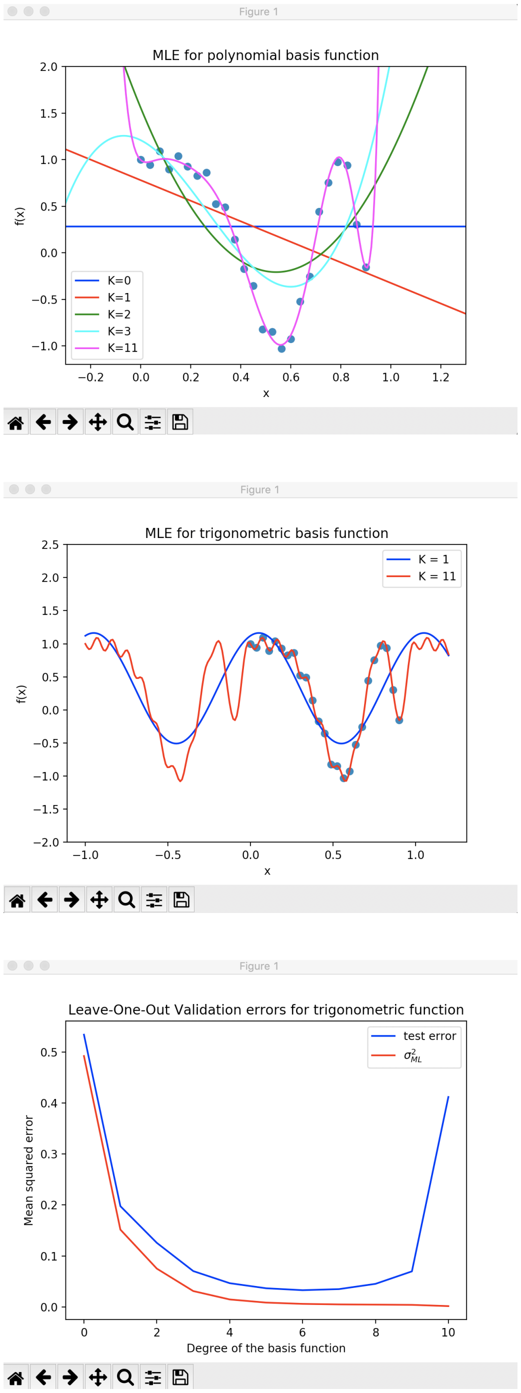

Figure 1: Answers to the question 1 (a), (b), (c)

d) A high level description of over-fitting is that the predictor fits too closely to the training data and does not generalize well to new data (Mitchell, 1997). Let's refer to the leave-one-out validation errors for trigonometric function of the figure 1. The shape of the σ_{ML}^2 graph suggests that the model (or predictor) learns well on the given data-set, thus decreasing in mean-squared-error as the degree increases. The shape of the test error graph shows that the model seems to generalize well until the degree of 6. However, as the degree increases, it fits to the given data-set too closely, so it does not generalize on the unseen data. Furthermore, the rest graphs of the figure 1 (i.e. MLE graphs) shows that the problem of over-fitting. Generally, it tends to over-fits to the given data-set as the degree increases. For example, the graph of MLE for Polynomial basis function shows that when K (i.e. degree) is 11, the graph almost perfectly go through all the existing data but does not generalize well. However, for K is 1, the opposite happens. The same properties of over-fitting and under-fitting occurs for trigonometric basis function as well.

2 Ridge Regression

a) Expand the linear regression with regularized least squares loss function

$$\begin{aligned} L(\boldsymbol{w}) &= \sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{\phi}(x_i))^2 + \lambda\sum_{j=1}^{M} w_j^2 \\ &= (\boldsymbol{y} - \boldsymbol{\Phi w})^T(\boldsymbol{y} - \boldsymbol{\Phi w}) + \lambda \boldsymbol{w}^T\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{\Phi w} - (\boldsymbol{\Phi w})^T\boldsymbol{y} + (\boldsymbol{\Phi w})^T(\boldsymbol{\Phi w}) + \lambda\boldsymbol{w}^T\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{\Phi w} - \boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi w} + \lambda\boldsymbol{w}^T\boldsymbol{w} \end{aligned} \tag{1}$$

Differentiating the equation (1) would result the following equation.

$$\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2(\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} - \boldsymbol{y}^T\boldsymbol{\Phi}) + 2\lambda\boldsymbol{w}^T \tag{2}$$

To minimize the loss function, equate the equation (2) to 0. This will yield the optimal $\boldsymbol{w}$ which gives us the minimum value of the loss function when it is used.

$$\boldsymbol{w}_{opt} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \tag{3}$$

We now consider the given regression problem where

$$\begin{aligned} p(\boldsymbol{w}) &= N(0, b^2\boldsymbol{I}) \\ b^2 &= \frac{1}{2\lambda} \\ -\log p(\boldsymbol{w}) &= \lambda\boldsymbol{w}^T\boldsymbol{w} \end{aligned} \tag{4}$$

To find MAP estimate, start from the equation derived from Bayes' rule.

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{X})p(\boldsymbol{w})}{\boldsymbol{p(X|y)}} \\
\log p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) &= \log p(\boldsymbol{y}|\boldsymbol{X}) + \log p(\boldsymbol{w}) - \log p(\boldsymbol{X}|\boldsymbol{y}) \\
-\log p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) &= -\log p(\boldsymbol{y}|\boldsymbol{X}) - \log p(\boldsymbol{w}) + \log p(\boldsymbol{X}|\boldsymbol{y}) \qquad (5)\\
&= \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\Phi w})^T(\boldsymbol{y}-\boldsymbol{\Phi w}) + \frac{1}{2b^2}\boldsymbol{w}^T\boldsymbol{w} + const \\
-\frac{\partial \log p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y})}{\partial \boldsymbol{w}} &= \frac{1}{\sigma^2}(\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} - \boldsymbol{y}^T\boldsymbol{\Phi}) + \frac{1}{b^2}\boldsymbol{w}^T
\end{aligned}$$

Find the MAP estimate $\boldsymbol{w}_{map}$ by setting this gradient to 0. From the lecture note, we know that the equation (5) can derive the equation below when its gradient is set to 0.

$$\boldsymbol{w}_{map} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \qquad (6)$$

Therefore, from the equation (3) and (6), we can conclude that $\boldsymbol{w}_{opt} = \boldsymbol{w}_{map}$ when $\lambda = \frac{\sigma^2}{b^2}$.

The regularized loss function contains mean-squared-errors and the regularizer where the errors is a distance from our prediction and the actual data. The regularizing term in the loss function reduces the effect of large $\boldsymbol{w}$. Therefore, it prevents over-fitting of the model thus benefiting our predictor.

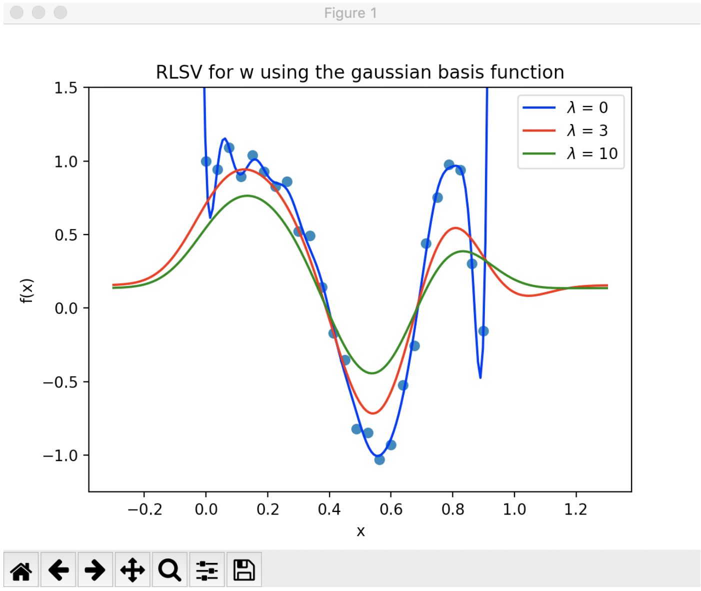

Figure 2: Answer to the question 2 (b)