CO-496 Coursework #2

Zhenhao Li

October 31, 2018

1 Linear Regression

a) By first finding the maximum likelihood solution for the parameters σ^2 and $\boldsymbol{w}$ in terms of $\boldsymbol{\Phi}$, plot the predictive mean at test points in the interval $[-0.3, 1.3]$ in the case of polynomial basis functions of order 0, 1, 2, 3 and also order 11. Plot all the curves on the same axes, showing also the data.

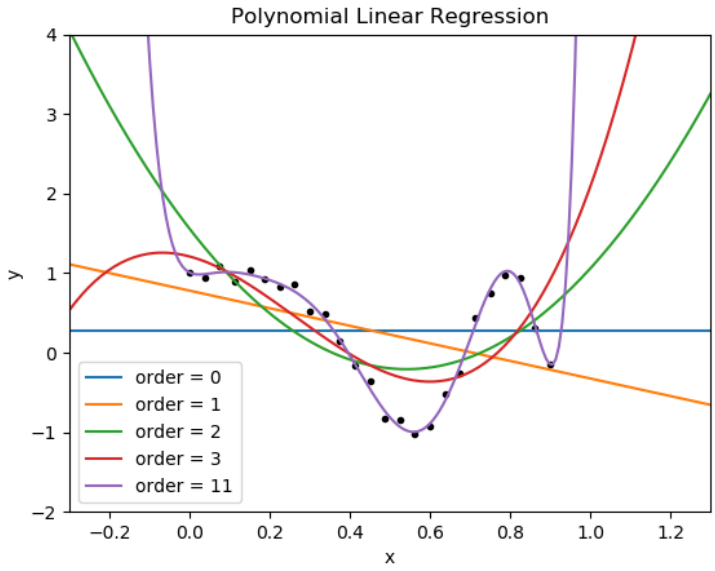

Figure 1: MLE with polynomial basis functions

b) Repeat the previous part but this time with trigonometric basis functions of orders 1 and 11. Use test points in $[-1, 1.2]$ to see the periodicity. Note that your basis functions should be of size $2J + 1$ for order J (i.e. dont forget the bias term)

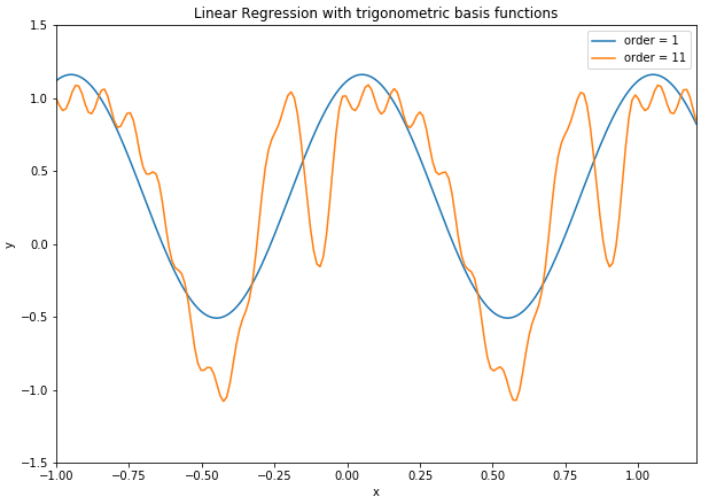

Figure 2: MLE with trigonometric basis functions

c) In this part you will investigate over-fitting with leave-one-out cross validation. You should use trigonometric basis functions of order 0 to 10 inclusive and for each choice use leave-one-out cross validation to estimate the average squared test error. Plot this average error on a graph against order of basis together. On the same graph plot also the maximum likelihood value for σ^2.

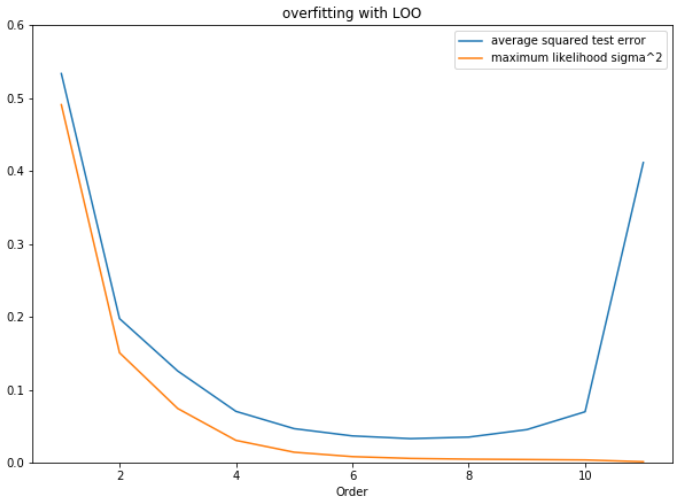

Figure 3: Cross-validation with LOO

d) Briefly describe the concept of over-fitting, using your graph in the previous

part as an illustrative example. You should also refer to your plots from the first two parts of this question.

- As the order increases, the σ^2 decreases and approaches to zero. This is because when the order increases, the model tries to pass through exactly every training point, thus resulting in overfitting.
- As the order increases, the test error first decreases because the model fits the data well. When the order becomes large (over 10), the model tends to overfit and therefore has bad generalizing ability. As a result, it cannot predict precisely on the data it has not seen, so the test error increases tremendously with larger order.
- Overfitting is a phenomenon that the model tries to fit precisely on every training data and thus loses ability to generalize on data it has not seen. With overfitting, the training error is relatively small while the test error is extremely high.
- From Fig.1, it can be seen that, if the order is too small (see the blue and orange lines), the model tends to underfit. Under this circumstance, both the training error and the test error is relatively high. When it comes to moderate orders, the model can describe the general distribution of the data points. However, if the order is too high, the model passes through each training points and has high test error(because the function tends to infinity as x goes larger or smaller). In Fig.2, the model with order one can describe the data points well (because the original distribution is in trigonometric form), but the model with order 11 overfits.

2 Ridge Regression

a) A non-probabilistic approach to linear regression is to set $y_i = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\phi}(x_i)$ and then find the best $\boldsymbol{w}$ by minimizing some loss function. Show that linear regression with regularized least squares loss function

$$L(\boldsymbol{w}) = \sum_{i=1}^{N}(y_i - \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\phi}(x_i))^2 + \lambda \sum_{j=1}^{M} w_j^2$$

is equivalent to the MAP estimate for w with the factorized Gaussian likelihood $y_i \sim \mathcal{N}(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\phi}(x_i), \sigma^2)$ and a certain prior for $\boldsymbol{w}$. You must state the prior you place on $\boldsymbol{w}$ and explicitly connect the parameters of your prior and the parameters of the likelihood to λ. Explain also the intuition behind this loss function.

- The prior $p(\boldsymbol{\theta}) = \mathcal{N}(0, b^2\boldsymbol{I})$, where $b^2 = \frac{1}{2}\lambda$
- The negative log likelihood is $-log(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\Phi\theta})^T(\boldsymbol{y} - \boldsymbol{\Phi\theta})$
- The negative log prior is equal to $\lambda\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{\theta} + const$, This means that for a quadratic regularization, the regularization parameter λ corresponds to twice the precision (inverse variance) of the Gaussian (iso- tropic) prior

$p(\theta)$. Therefore, the log-prior reflects the impact of the regularizer that penalizes implausible values.

- $\boldsymbol{w}_{MAP} = (\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I})^{-1}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{y}$
- when introduce of λ increases the coefficient of Φ, thus to minimize the loss function, $\boldsymbol{w}$ with higher order tends to be smaller(with same λ, higher order phi_i increases the loss function more.) Therefore, it might prevent the model from overfitting.

b) Find the regularized least squares value for $\boldsymbol{w}$ using 20 Gaussian basis functions of scale 0.1 with means equally spaced in $[0, 1]$ and plot the regression function for test points between -0.3 and 1.3 for three values of λ of your choosing. Your chosen values of λ should illustrate under-fitting, over-fitting and somewhere more satisfactory in between. Make sure you label which is which, together with the values of λ that you used.

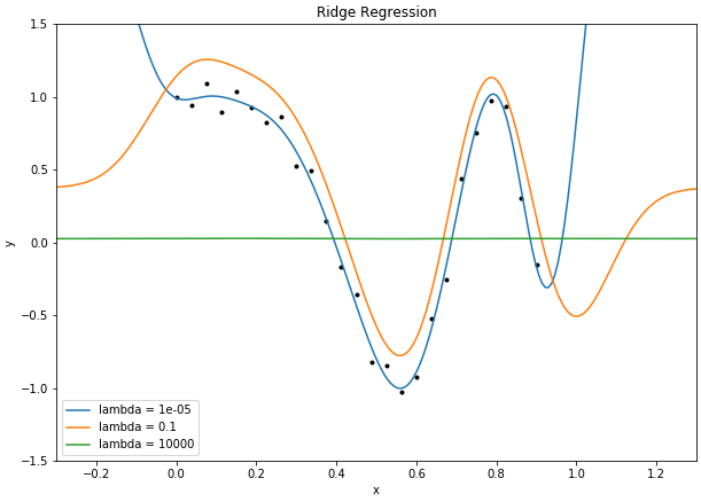

Figure 4: Ridge Regression with different λ