Machine Learning
Homework 1

Yao Song
301266041

September 25, 2014

1 Question 1

1.1 ML estimate of μ

$$\begin{aligned}
&\arg\max_{\mu} P(x_1, x_2, ..., x_N|\mu) && (1)\\
&= \arg\max_{\mu} P(x_1|\mu)P(x_2|\mu)...P(x_N|\mu) && (2)\\
&= \arg\max_{\mu} \exp\left\{-\frac{1}{2\sigma^2}[(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_N-\mu)^2]\right\} && (3)\\
&= \arg\min_{\mu}[(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_N-\mu)^2]. && (4)
\end{aligned}$$

Define $f(\mu) = (x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_N-\mu)^2$.
Let $\nabla_\mu f(\mu) = -2(\sum_i x_i - N\mu) = 0$, we get $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

1.2 MAP estimate of μ

$$\begin{aligned}
&\arg\max_{\mu} P(\mu|D) = \arg\max_{\mu} P(\mu)P(D|\mu) && (5)\\
&= \arg\max_{\mu} \exp\left\{-\frac{1}{2\sigma^2}[(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_N-\mu)^2 + (a-\mu)^2]\right\} && (6)\\
&= \arg\min_{\mu}[(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_N-\mu)^2 + (a-\mu)^2]]. && (7)
\end{aligned}$$

Like the ML estimate case, by taking derivative of the polynomial of μ, we get $\mu_{MAP} = \frac{1}{N+1}(a + \sum_{i=1}^{N} x_i)$

1.3 what the prior mean a does in the MAP case

The MAP estimate is a linear interpolation of the prior mean and the sample mean. The prior mean a help to adjust the sample mean estimate.

1.4 effect of a much smaller σ_p^2

If the prior distribution has a much smaller variance, than the MAP estimate will be much closer to the prior mean value a. Basically that means we have a quite confident prior distribution, much more confident than the likelihood distribution.

2 Question 2

2.1 optimal weights $\boldsymbol{\omega}$

Take gradient of the weighted sum of squares error function, we get the following equation:

$$\nabla_{\boldsymbol{\omega}} E(\boldsymbol{\omega}) = -\sum_{n=1}^{N} \alpha_n (t_n - \boldsymbol{\omega}^T \boldsymbol{\phi}(x_n))\boldsymbol{\phi}(x_n)^T. \tag{8}$$

Set the gradient to be zero, we obtain the following equation:

$$\sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(x_n)^T = \boldsymbol{\omega}^T \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T. \tag{9}$$

Thus

$$\boldsymbol{\omega}^T = (\sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(x_n)^T)(\sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T)^{-1}, \tag{10}$$

Define diagonal matrix $\mathbf{A} = diag(\alpha_1, \alpha_2, ..., \alpha_N)$, $\mathbf{t}^T = [t_1, t_2, ..., t_N]$, $\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_M(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_M(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_N) & \phi_2(x_N) & \cdots & \phi_M(x_N) \end{pmatrix}$, then the Equation (9) can be written as:

$$\boldsymbol{\Phi}^T \mathbf{A} \boldsymbol{t} = \boldsymbol{\omega}^T \boldsymbol{\Phi}^T \mathbf{A} \boldsymbol{\Phi}, \tag{11}$$

Thus,

$$\boldsymbol{\omega} = (\boldsymbol{\Phi}^T \mathbf{A} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{A} \boldsymbol{t}, \tag{12}$$

2.2 interpretation of weights α_n

In the Gaussian noise model, $\mathbf{t} = \mathbf{y(x,w)} + \boldsymbol{\epsilon}$, α_n is the variance of the noise term ϵ_n. Thus, the likelihood function of the data is written as follows:

$$p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{\alpha}) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T \boldsymbol{\phi(x_n)}, \alpha_n) \tag{13}$$

By taking log function over the above equation, the weighted sum of squares error function will arise. α_n is the variance of the noise term.

3 Question 3

For any matrix A, by the definition of eigenvalue, λ and eigenvector,v, we have the following equation:

$$Av = \lambda v, \tag{14}$$

which is equivalent to

$$Av - \lambda v = (A - \lambda I)v = 0, \tag{15}$$

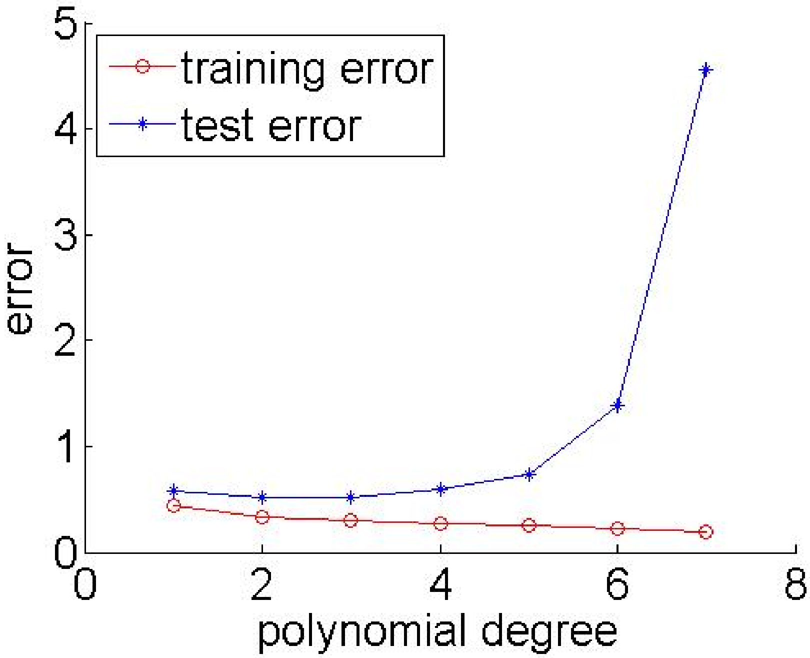
where I is the identity matrix with the same size as matrix A. By the fundamental theorem of linear algebra, if the above linear equation has a nonzero solution v, then $\det(A - \lambda I) = 0$.

$$\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)...(\lambda_n - \lambda) \tag{16}$$

If we set $\lambda = 0$, we get $\det A = \lambda_1 \lambda_2 ... \lambda_n$.

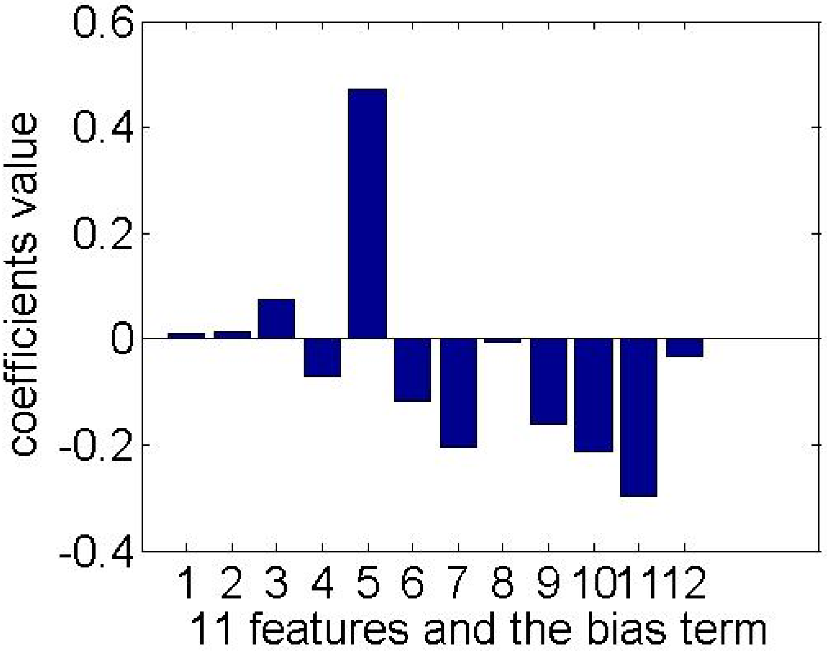
4 Question 4

4.1 polynomial regression from degree 1 to 7



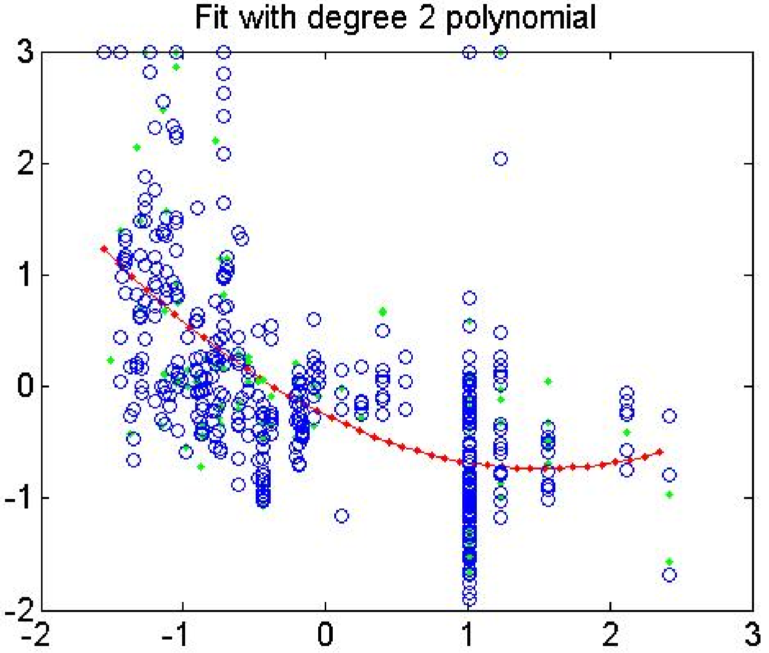
Comment: The training error goes to zero as the polynomial degree increases. The test error goes to the minimum and then increases as the polynomial degree increases. This is due to over fitting in higher degrees.

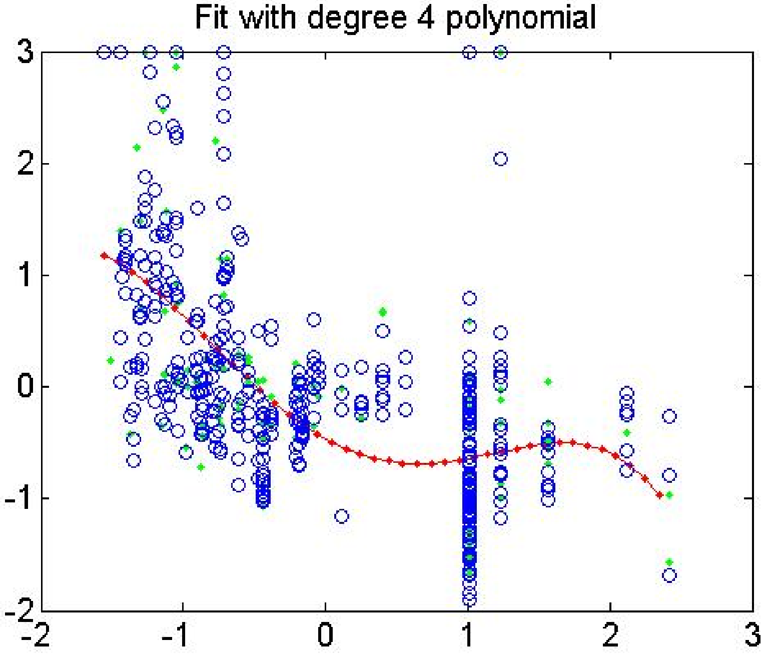
4.2 polynomial regression degree 1 weights

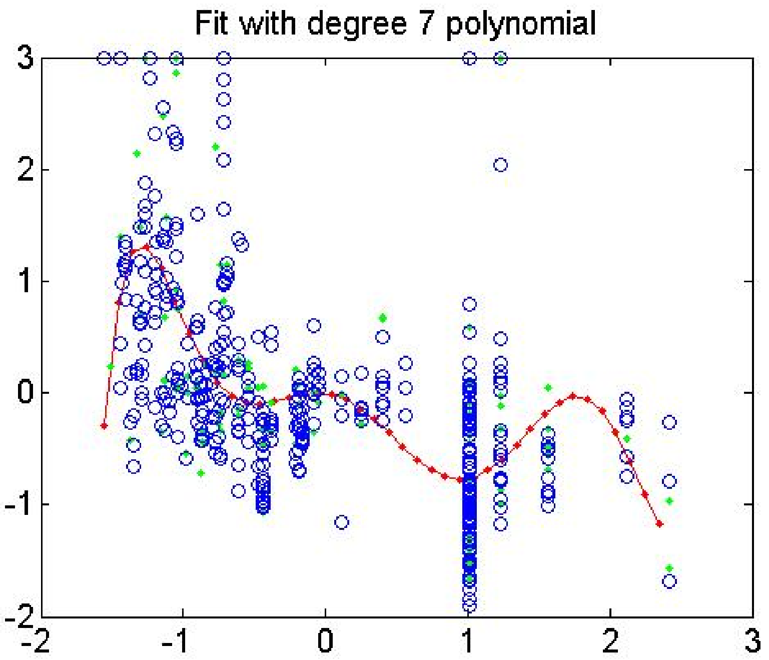


Comment: The 5th feature has the largest positive value of the weight. A higher value of average number of rooms per dwelling implies a higher house value. The 7th feature is negative. That means the feature has a negative effect on house value. That is reasonable because people prefer house near Boston employment center.

4.3 1d vis

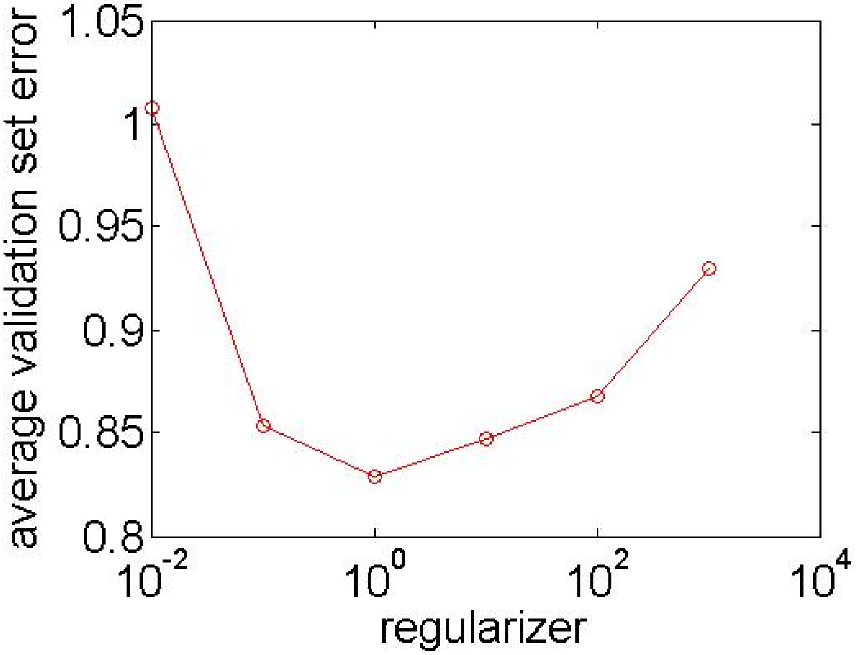

Fit with degree 2 polynomial
3
2
1
0
-1
-2
-2
-1
0
1
2
3


Fit with degree 4 polynomial
3
2
1
0
-1
-2
-2
-1
0
1
2
3



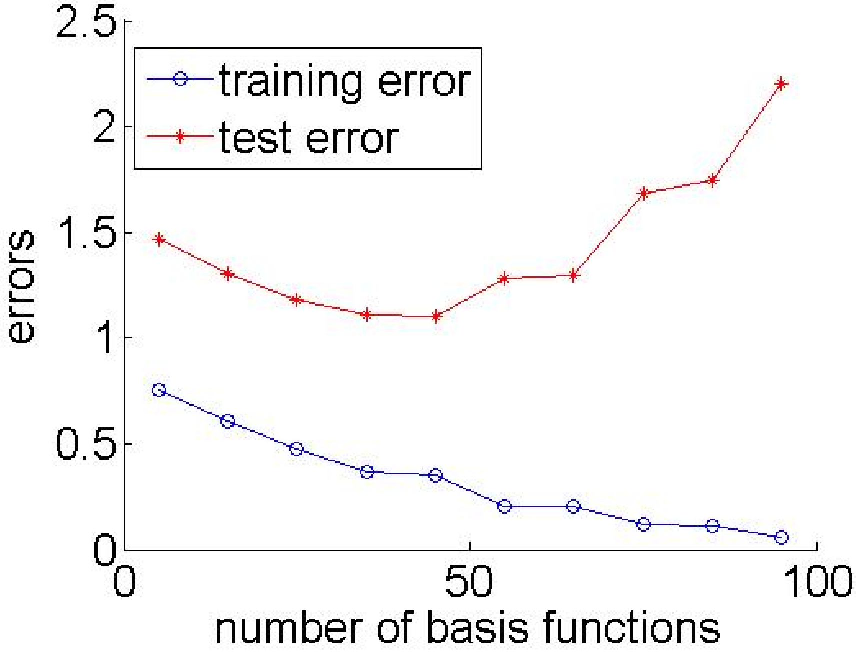
Comment: higher degree will give us a more complex curve. However, the points are so randomly distributed, the regression is still not well enough.

4.4 polynomial regression regularized



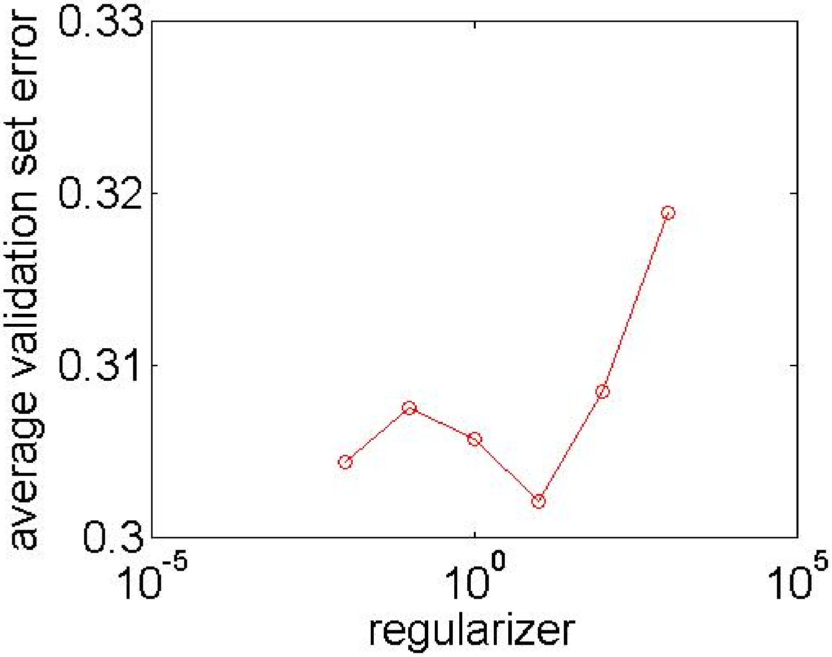
the case when $\lambda = 0$ is not in the figure. When $\lambda = 0$, $error = 1.0550$. The best λ is 1.

4.5 Gaussian regression



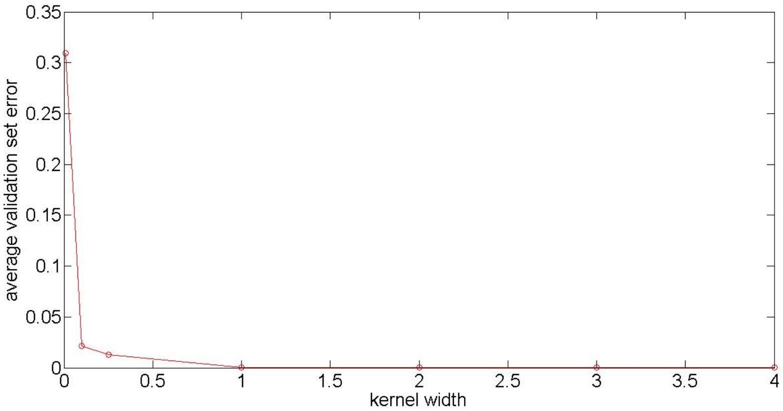
comment: the training error goes down as the number of basis functions increase. The test error decreases to minimum then increases with number of basis functions. This is due to over fitting.

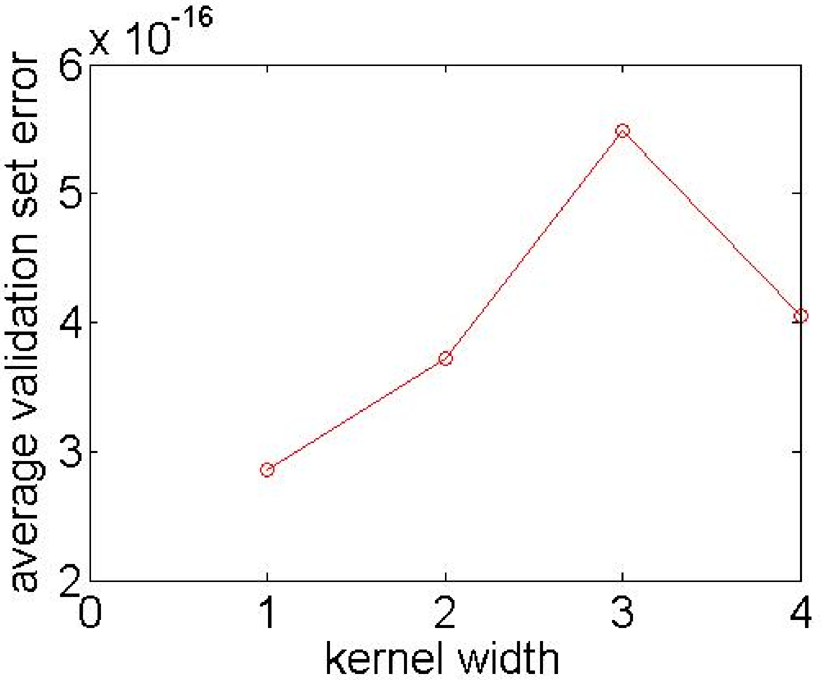
4.6 Gaussian regression regularized



the case when $\lambda = 0$ is not in the figure.when $\lambda = 0$, $error = 0.3683$,which is much bigger than the ones in the figure. The best λ is 10.

5 Question 5





The first figure shows all the values of h in the same figure. We can observe small values of h results in a higher estimation error. This is because smaller h makes the kernel function to be zero. In that case, I just estimate the expected t value to be an average of the training t values. The second figure zooms in to show the validation error when h =1,2,3,4. It is obvious that h = 1 is the best value.