Machine Learning, Spring 2018
Homework 4

Due on 23:59 May 1, 2018
Send to $cs282\_01@163.com$
with subject "Chinese name+student number+HW4"

1 Hoeffding Inequality

(1) $\mathrm{P}(v \leq 0.1) = \binom{10}{0}\mu^0(1-\mu)^{10} + \binom{10}{1}\mu^1(1-\mu)^9 = 9.1 \times 10^{-9}$

(2) $\because \mathrm{P}[|v-\mu| > \varepsilon] \leq 2\exp\left(-2\varepsilon^2 N\right)$
Set $\varepsilon = 0.8$, then we get the bound is
$2\exp\left(-2\varepsilon^2 N\right) = 2\exp\left(-2 \times 0.8^2 \times 10\right) = 5.5 \times 10^{-6}$

2 Bias-variance decomposition

(1) **Lemma:** $Var(z) = \mathbb{E}[(z-\bar{z})^2] = \mathbb{E}[z^2] - \bar{z}^2$

Proof.

$$Var(z) = \mathbb{E}[(z-\bar{z})^2] = \mathbb{E}[z^2 + \bar{z}^2 - 2z\bar{z}] = \mathbb{E}[z^2] + \bar{z}^2 - 2\bar{z}^2 = \mathbb{E}[z^2] - \bar{z}^2$$

$\square$

Then, show that $\textbf{variance} + \textbf{bias}^2 + \sigma^2 = \mathbb{E}_{\mathcal{D},\epsilon}[(y^* - h(x^*))^2]$

Proof.

$$\overline{y^*} = \overline{f(x^*) + \epsilon} = \overline{f(x^*)} + 0 = f(x^*) \Rightarrow \overline{y^*} = f(x^*)$$

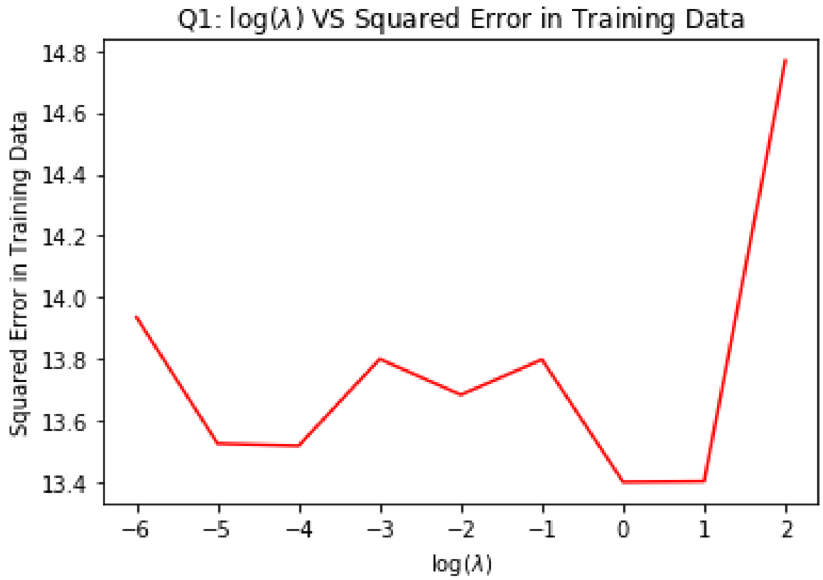

Figure 1: Q1: $\log(\lambda)$ VS Squared Error in Training Data

$$\begin{aligned}
\textbf{variance} + \textbf{bias}^2 + \sigma^2 &= \mathbb{E}_{\mathcal{D}}[(h(x^*) - \overline{h(x^*)})^2] + [\overline{h(x^*)} - f(x^*)]^2 + \mathbb{E}_{\epsilon}[(y^* - f(x^*))^2] \\
&= Var_{\mathcal{D}}[h(x^*)] + f^2(x^*) + \overline{h(x^*)}^2 - 2f(x^*) \cdot \overline{h(x^*)} + \mathbb{E}_{\epsilon}[(y^* - \overline{y^*})^2] \\
&= Var_{\mathcal{D}}[h(x^*)] + \overline{y^*}^2 + \overline{h(x^*)}^2 - 2\overline{y^*} \cdot \overline{h(x^*)} + Var_{\epsilon}[y^*] \\
&= (Var_{\mathcal{D}}[h(x^*)] + \overline{h(x^*)}^2) + (\overline{y^*}^2 + Var_{\epsilon}[y^*]) - 2\overline{y^*} \cdot \overline{h(x^*)} \\
&= \mathbb{E}_{\mathcal{D}}[h^2(x^*)] + \mathbb{E}_{\epsilon}[(y^*)^2] - 2\mathbb{E}_{\epsilon}[y^*] \cdot \mathbb{E}_{\mathcal{D}}[h(x^*)] \\
&= \mathbb{E}_{\mathcal{D},\epsilon}[(y^* - h(x^*))^2]
\end{aligned} \tag{1}$$

Therefore, $\mathbb{E}_{\mathcal{D},\epsilon}[(y^* - h(x^*))^2] = \textbf{variance} + \textbf{bias}^2 + \sigma^2$. □

3 Cross Validation And L2 Regularization

In this program, I set the learning rate is equal to 0.00001 and termination condition is the L1 norm of the gradient of weight not larger than 7×10^{-5}.

(1) Please look at fig.1
The λ range from 10^{-6} to 10^3, step is $\times 10$, and I chose the smallest validation loss among 10 folders for each λ as the result.

(2) Please look at fig.2

(3) The threshold is equal to 10^{-8}.

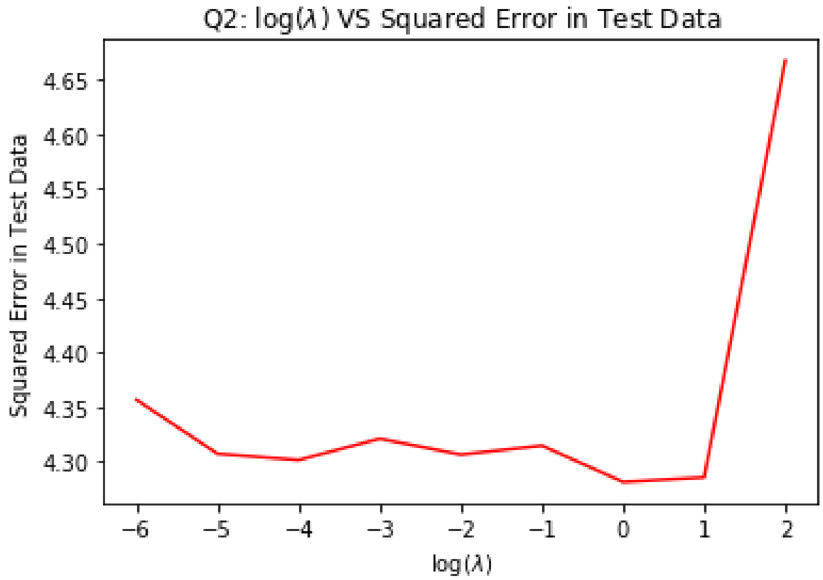

Figure 2: Q2: $\log(\lambda)$ VS Squared Error in Test Data

When $\lambda = 1$, both of training loss and test loss are minimum, so I chose $\lambda = 1$.

(4) When $\lambda = 1$,
The best test set proformance: test loss = 4.28
The largest coefficient: 0.23
The smallest coefficient: -0.045

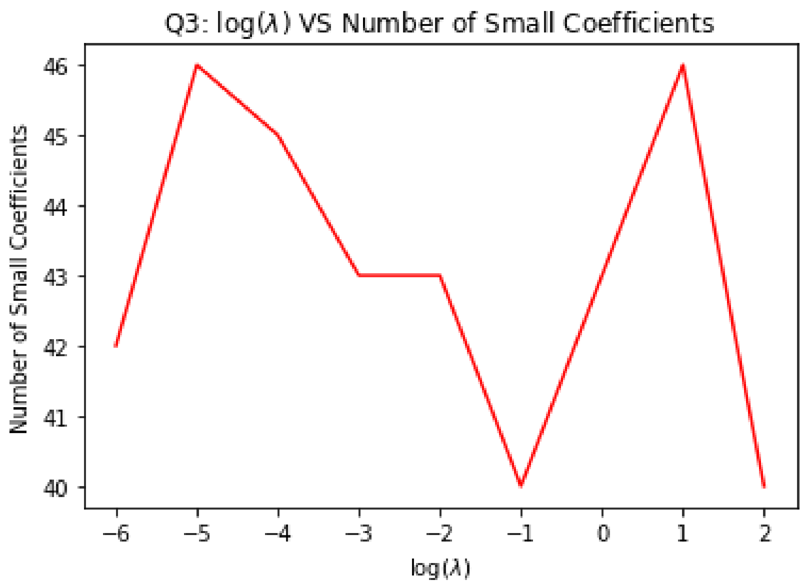

Figure 3: Q3: $\log(\lambda)$ VS Number of Small Coefficients