Machine Learning, Spring 2018
Homework 4

Due on 23:59 May 1, 2018
Send to *cs*282_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.

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

□

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

Proof.

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

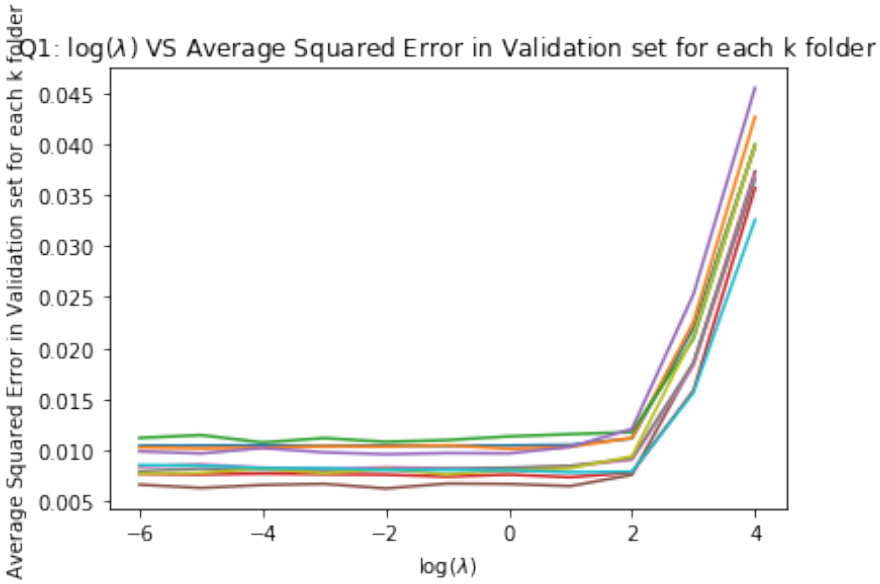

Figure 1: Q1: log(λ) VS Average Squared Error in valdation set for each k folder

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

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 L infinity 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 10^5, step is $\times 10$.
Consider the numbers of data in different folders are different, so I used the average of square loss in validation set.

log(λ) VS Average Squared Error in valdation set for average of K-folder.
Please look at fig.2

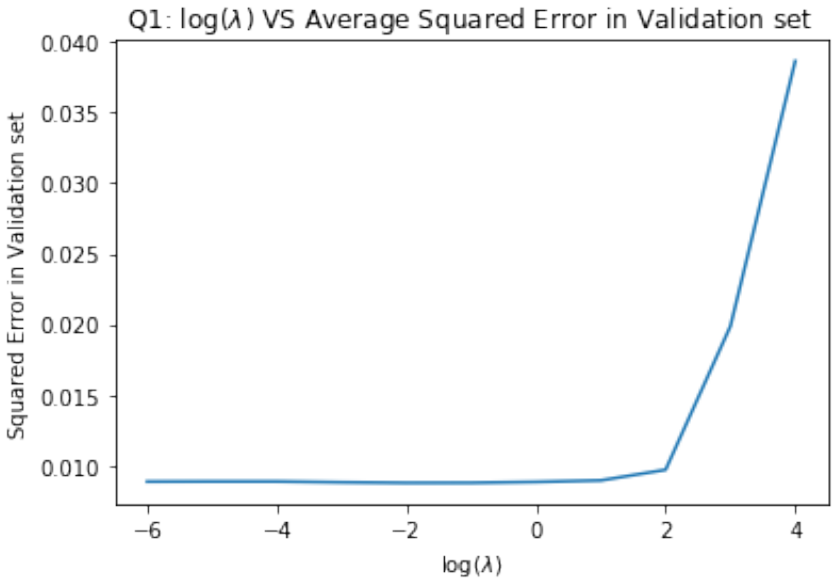

Figure 2: Q2: log(λ) VS Average Squared Error in valdation set for average of K-folder

(2) When $\lambda = 0.01$,
The average validation loss is minimum.
The best test set proformance: test loss = 4.18

(3) Please look at fig.3
The threshold is equal to 10^{-4}.
When $\lambda = 0.01$, the number of small coefficients is not large, so it can keep more features. Therefore $\lambda = 0.01$ is reasonable

(4) When $\lambda = 0.01$,
The best test set proformance: test loss = 4.18
The largest coefficient:
name: **PctIlleg**, value: -0.019
The smallest coefficient:
name: **PctKids2Par**, value: -0.01

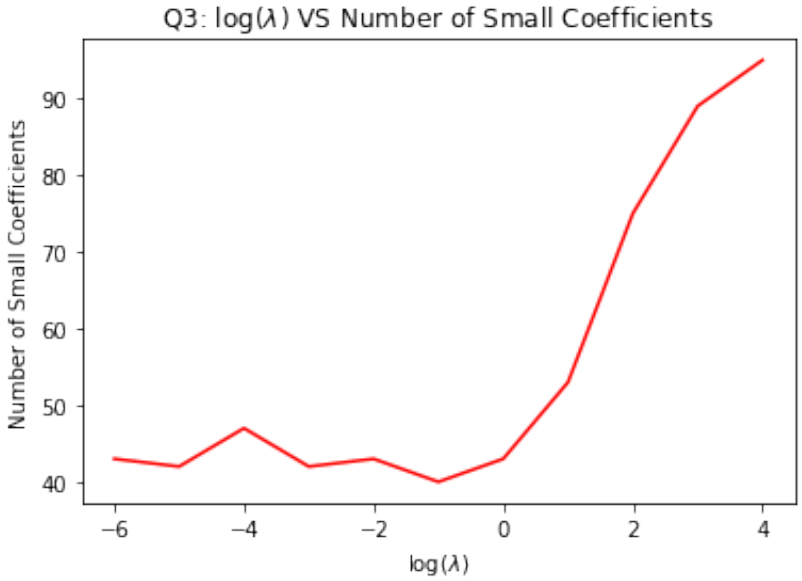

Figure 3: Q3: log(λ) VS Number of Small Coefficients