CIS 520, Machine Learning, Fall 2021
Homework 2
Due: Sunday, September 26th, 11:59pm
Submit to Gradescope

Yixuan Meng, Zhouyang Fang

Sept 26th

1 Regression Models and Squared Errors

Regression problems involves instance spaces $\mathcal{X}$ and labels, and the predictions, which are real-valued as $\mathcal{Y} = \hat{\mathcal{Y}} = \mathbb{R}$. One is given a training sample $S = ((x_1, y_1), ..., (x_m, y_m)) \in (\mathcal{X} \times \mathbb{R})^m$, and the goal is to learn a regression model $f_S : \mathcal{X} \to \mathbb{R}$. The metric used to measure the performance of this regression model can vary, and one such metric is the squared loss function. The questions below ask you to work with regression problems and squared error losses.

1. The squared error is given by $\mathbb{E}_{(x,y)\sim p(X,Y)}[(f(x) - y)^2]$, where the examples are drawn from a joint probability distribution $p(X, Y)$ on $\mathcal{X} \times \mathbb{R}$. Find the lower bound of the expression $\mathbb{E}_{(x,y)\sim p(X,Y)}[(f(x) - y)^2]$. From this lower bound, what is the optimal expression of $f(x)$, in terms of x and Y?

$$\begin{aligned}\mathbb{E}_{(x,y)\sim p(X,Y)}[(f(x) - y)^2] &= \mathbb{E}[(y - \hat{y})^2] + \mathbb{E}[(\hat{y} - f)]^2 \\ &\geq \mathbb{E}[(y - \hat{y})^2].\end{aligned}$$

$$f(x) = \mathbb{E}(Y|X = x).$$

2. With this result, complete the following two problems. Consider the regression task in which instances contain two features, each taking values in $[0, 1]$, so that the instance space is $\mathcal{X} = [0, 1]^2$, and with label and prediction spaces belonging to the real space. Suppose examples $(\mathbf{x}, y)$ are drawn from the joint probability distribution D, whose marginal density on $\mathcal{X}$ is given by

$$\mu(\mathbf{x}) = 2x_1, \quad \forall \mathbf{x} = (x_1, x_2) \in \mathcal{X}$$

and the conditional distribution of Y given $\mathbf{x}$ is given by

$$Y|X = \mathbf{x} \sim \mathcal{N}(x_1 - 2x_2 + 1, 2)$$

What is the optimal regression model $f^*(X)$ and the minimum achievable squared error for D?

$$f(x) = \mathbb{E}(\mathbf{Y}|\mathbf{X} = x) = x_1 - 2x_2 + 1$$

$$L_D[f^*] = \mathbb{E}[\mathbf{Var}(\mathbf{Y}|\mathbf{X} = x)] = \mathbf{Var}(\mathbf{Y}|\mathbf{X} = x) = 2$$

3. Suppose you give your friend a training sample $S = ((\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m))$ containing m examples drawn i.i.d from D, and your friend learns a regression model given by

$$f_S(\mathbf{x}) = x_1 - 2x_2, \quad \forall \mathbf{x} = (x_1, x_2) \in \mathcal{X}$$

Find the squared error of f_S with respect to D.

$$L_D[f_S] = \mathbb{E}[(y - f_S)^2] = \mathbb{E}[(y - \hat{y})^2] + \mathbb{E}[(\hat{y} - f_S)]^2 = 2 + 1 = 3$$

4. Consider a linear model of the form

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{P} w_i x_i$$

together with a sum of squares error function of the form

$$L_P(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (f(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n)^2$$

where P is the dimensionality of the vector $\mathbf{x}$, N is the number of training examples, and $\mathbf{t}$ is the ground truth target . Now suppose that the Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, show that minimizing L_P averaged over the noise distribution is equivalent to minimizing the sum of squares error for noise-free input variables L_P with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

$$\begin{aligned}
\mathbb{E}[\tilde{L}] &= \mathbb{E}\left[\frac{1}{2}\sum_{n=1}^{N}(\tilde{f}_n - t_n)^2\right] \\
&= \mathbb{E}\left[\frac{1}{2}\sum_{n=1}^{N}\left(\left(w_0 + \sum_{i=1}^{P} w_i\left(x_{ni} + \epsilon_{ni}\right)\right) - \mathbf{t}_n\right)^2\right] \\
&= \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\Bigg[(f(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n) + \sum_{i=1}^{P} w_i\epsilon_{ni}\Bigg)^2\Bigg] \\
&= \frac{1}{2}\sum_{n=1}^{N}\Bigg(\mathbb{E}\left[(f(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n)^2\right] + 2\mathbb{E}\Bigg[\Bigg(f(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n * \left(\sum_{i=1}^{P} w_i\epsilon_{ni}\right)\Bigg] + \mathbb{E}\left[\left(\sum_{i=1}^{P} w_i\epsilon_{ni}\right)^2\right]\Bigg) \\
&= \frac{1}{2}\sum_{n=1}^{N}\Bigg(\mathbb{E}\left[(f(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n)^2\right] + \mathbb{E}\left[\left(\sum_{i=1}^{P} w_i\epsilon_{ni}\right)^2\right] + 2\Bigg(\Bigg[(f(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n) * \left(\sum_{i=1}^{P} w_i\mathbb{E}\left[\epsilon_{ni}\right]\right)\Bigg)\Bigg) \\
&= \frac{1}{2}\sum_{n=1}^{N}\left[(f(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n)^2 + \sigma^2\sum_{i=1}^{P} w_i^2\right]
\end{aligned}$$

2 MLE and MAP

1. Suppose that there are vaccinated (V) and unvaccinated (U) people in a population and the vaccinated proportion is θ. That is, a random draw from the population will result in drawing a vaccinated person with a probability θ and drawing a unvaccinated person with a probability $1 - \theta$. If we randomly draw two people with replacement, we can draw either one of these three combinations (VV, VU, UU). What are the probabilities of each of the outcomes?

$$\begin{aligned}P(VV) &= \theta^2\\ P(VU) &= 2\theta(1-\theta)\\ P(UU) &= (1-\theta)^2\end{aligned}$$

2. Suppose that in the population p_1 people draw VV, p_2 people draw VU and p_3 people draw UU. What is the log likelihood function $LL(P(D/\theta)$? Find the Maximum A Posteriori estimate of θ with an assumption that your prior belief about vaccination can be described with a Beta distribution of $\alpha = 2, \beta = 2$.

$$\begin{aligned}LL(P(D/\theta)) &= \log((\theta^2)^{p_1}(2\theta(1-\theta)^{p_2})((1-\theta)^2)^{p_3})\\ &= p_1\log(\theta^2) + p_2\log(2\theta(1-\theta)) + p_3\log((1-\theta)^2)\\ \hat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D)\\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\\ \text{Let f } &= \log P(D|\theta) + \log P(\theta)\\ \frac{df}{d\theta} &= p_1\frac{2\theta}{\theta^2} + p_2\frac{1-2\theta}{\theta(1-\theta)} + p_3\frac{2\theta-2}{(1-\theta)^2} + \frac{1-2\theta}{\theta(1-\theta)} = 0\\ 2p_1 + p_2\frac{1-2\theta}{1-\theta} + p_3\frac{1-2\theta}{1-\theta} + \frac{1-2\theta}{1-\theta} &= 0\\ 2p_1(1-\theta) + p_2(1-2\theta) + p_3(-2\theta) + 1 - 2\theta &= 0\\ (2p_1 + 2p_2 + 2p_3 + 2)\theta &= 2p_1 + p_2 + 1\\ \hat{\theta}_{MAP} &= \frac{2p_1 + p_2 + 1}{2p_1 + 2p_2 + 2p_3 + 2}\end{aligned}$$

3. Suppose that out of 100 people, 30 draw VV combination, 30 draw VU combination and 40 draw UU combination. What is the MAP estimate of θ (fraction of vaccination) given the same prior belief as in (b)? (Round up your answer to 2 decimal points.)

$$p_1 = p_2 = 30 \tag{1}$$

$$p_3 = 40 \tag{2}$$

$$\hat{\theta} = 0.4505$$

4. We have a dataset with N records in which the i^{th} record has one real-valued input attribute x_i and one real-valued output attribute y_i. The model has one unknown parameter w to be learned from data, and the distribution of y_i is given by

$$y_i \sim \mathcal{N}(\log(wx_i), 1)$$

Suppose you decide to do a maximum likelihood estimation of w. What equation does w need to satisfy to be a maximum likelihood estimate?

$$
\begin{aligned}
LL &= \log(\Pi_i \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\log(wx_i))^2}{2\sigma^2}}) \\
&= \sum_i (\log(\frac{1}{\sqrt{2\pi}\sigma}) + \frac{-(y-\log(wx_i))^2}{2\sigma^2}) \\
\text{Let} \quad 0 &= \frac{\partial LL}{\partial w} \\
\sum_i (y_i - \log(wx_i))(-\frac{1}{wx_i})x_i &= 0 \\
\sum_i (y_i - \log(wx_i)) &= 0 \\
\sum_i (\log(wx_i)) &= \sum_i y_i \\
\log(\Pi_i wx_i) &= \sum_i y_i \\
\Pi_i wx_i &= e^{\sum_i y_i} \\
w &= \sqrt[N]{\frac{e^{\sum_i y_i}}{\Pi_i x_i}}
\end{aligned}
$$

3 Programming: Least Squares Regression

3.1 Implementing Linear Regression

Functions implemented on colab notebook.

3.2 Data Set 1 (synthetic 1-dimensional data)

This data set contains 100 training examples and 1000 test examples, all generated from a fixed distribution with random noise, i.e. $y = f(x) + \epsilon$. For this data set, you will run unregularized least squares regression.

1. **Learning Curve.**

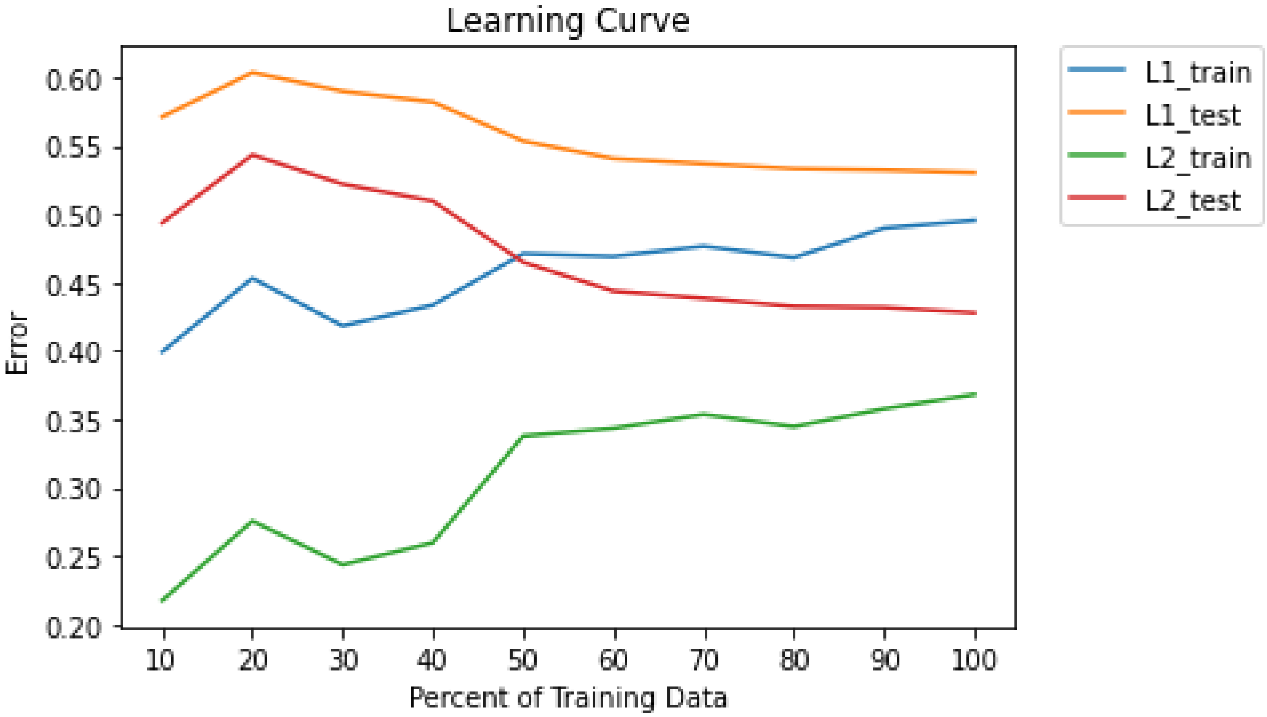

Figure 1: Learninig Curve

...

2. **Analysis of model learned from full training data.**

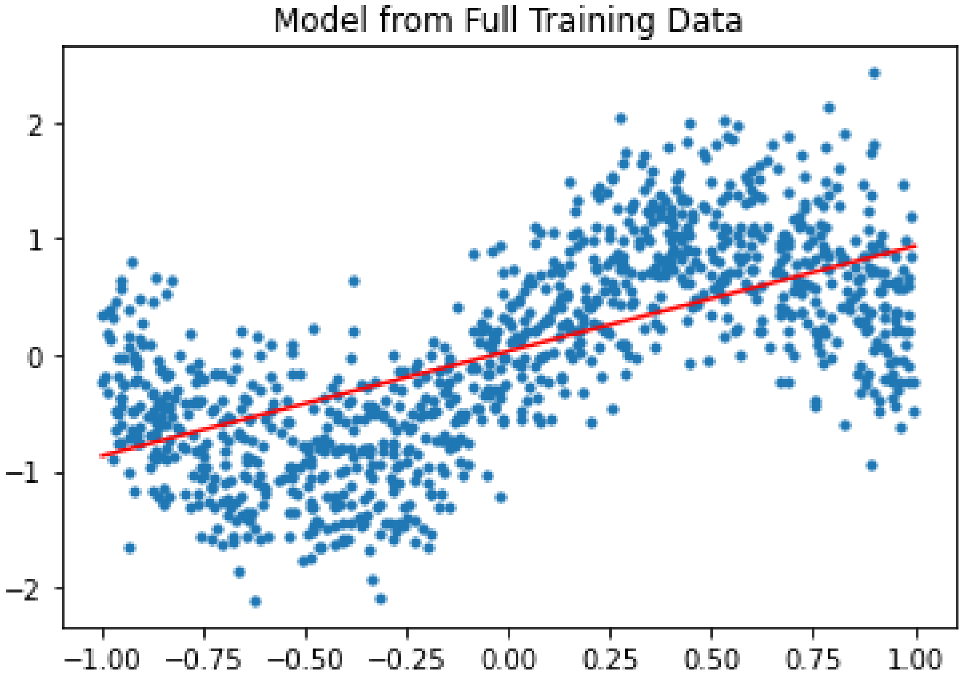

Figure 2: Model from Full Training Data

$\hat{w} = 0.90222293$
$\hat{b} = 0.03269313991926599$
L_2 training error = 0.3679586735242827
L_2 test error = 0.42779238775102935

3.3 Data Set 2 (real 12-dimensional data)

This is a real data set that involves predicting house prices from 12 features. The data set is a subset of a larger dataset from Kaggle (`https://www.kaggle.com/greenwing1985/housepricing`). This subset has 800 training examples and 200 test examples. For this data set, you will run L2-regularized least squares regression (ridge regression).

1. **Regression on 5% of the training data.**

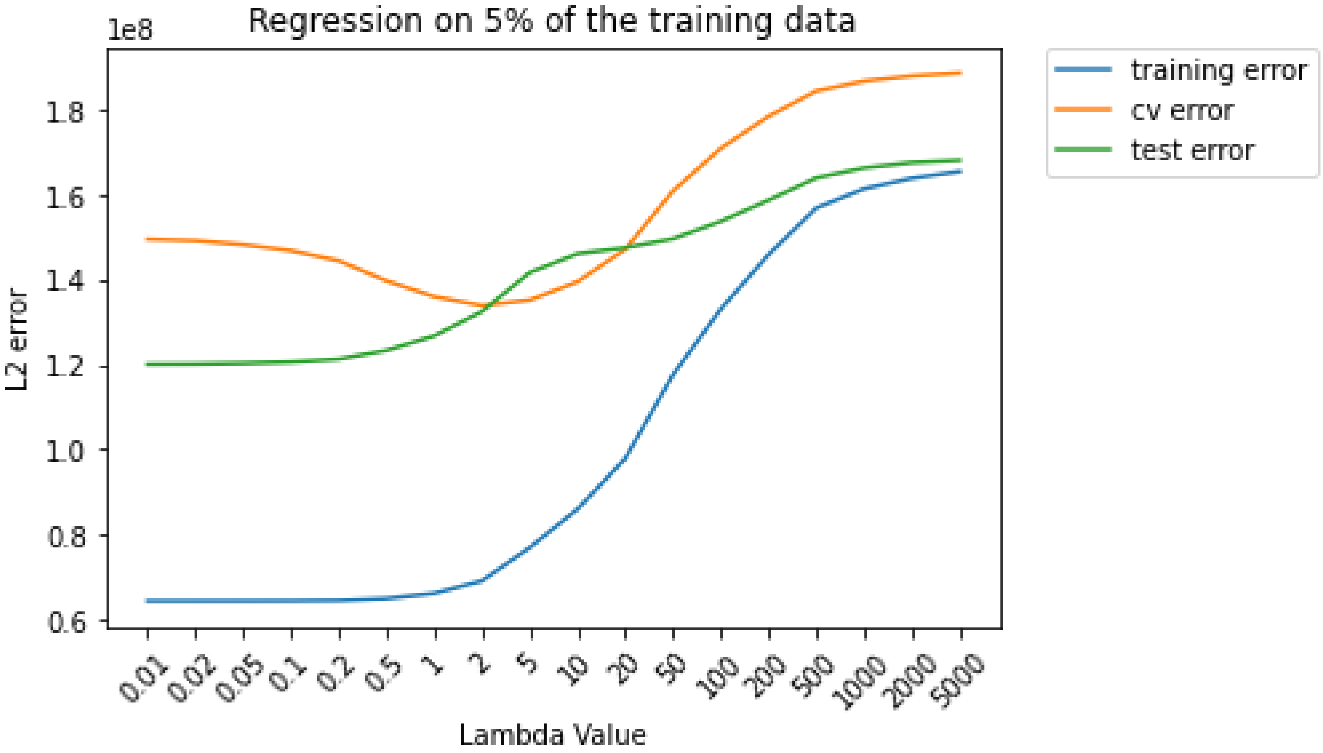

Figure 3: Regression on 5% of the training data.

The chosen λ is 2, and

$$
\begin{aligned}
w &= \begin{bmatrix} 61.39087067 \\ 2661.48434529 \\ -111.33278024 \\ 3323.06099759 \\ -5981.58769171 \\ -11202.53022041 \\ 4988.28083995 \\ 203.70887336 \\ -274.62007833 \\ 1774.20037102 \\ 3853.76840516 \\ 763.78363582 \end{bmatrix}, \\
b &= 9574.355516880052, \\
Error_{train} &= 69111801.41325754, \\
Error_{cv} &= 133870238.35320623, \\
Error_{test} &= 132465674.90224601.
\end{aligned}
$$

2. **Regression on 100% of the training data.**

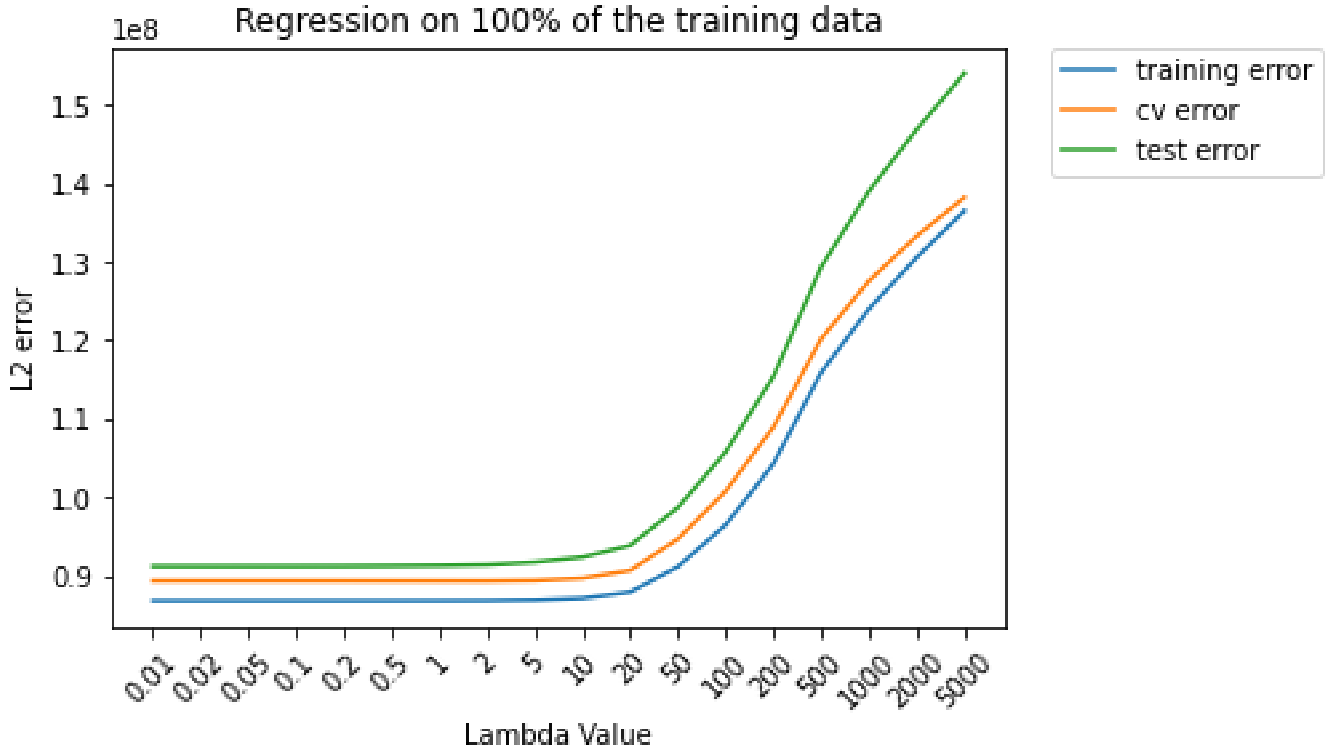

Figure 4: Regression on 100% of the training data.

The chosen λ is still 2, and

$$
\begin{aligned}
w &= \begin{bmatrix} 38.94417856 \\ 959.21868228 \\ 715.92173717 \\ 1296.93194987 \\ -8971.98954866 \\ -13971.43652607 \\ 4119.84285173 \\ 421.44893227 \\ 663.09527805 \\ 3951.27092709 \\ 897.82857383 \\ 1229.96403083 \end{bmatrix}, \\
b &= 25305.58267569583, \\
Error_{train} &= 86794902.9200716, \\
Error_{cv} &= 89333713.9426765, \\
Error_{test} &= 91385314.14078139.
\end{aligned}
$$

3. **Comparison of models learned by two methods.** For each of the two training sets considered above (5% and 100%), compare the training and test errors of the models learned using ridge regression. What can you conclude from this about the value of regularization for small and large training sets?

   The training error is smaller in small training set than large set. However, the test error is bigger in small training set than large set. The value of regularization for small and large training sets are the same, both $\lambda = 2$.

4. **Theoretical Value of λ.** For each of the two training sets considered above (5% and 100%), which λ should be larger by theory? Why? Do those values align with the conclusion you made in part 3.3.3?

   The smaller set should have a larger λ by theory, since the small one tends to be overfitting. When data is not enough, the parameters need more regularization and rely more on prior estimation. However, the actual result obtained from the model is two λs are the same.