1.Probabilistic Modeling

In lecture we went over an example of modeling coin tossing – estimating a parameter μ, the probability the coin comes up heads. Consider instead the problem of modeling the outcome of the BC Provincial election. To simplify matters, assume one party will win a majority (i.e. either the NDP, Liberals, or Green Party wins). We wish to characterize uncertainty over our belief about the chances of each party winning.

What are the parameters μ of this distribution? (See PRML Appendix B.)

The problem can be modeled by a multidimensional distribution with 3 outcomes. We can set the following assumptions on this problem:

- Each trial is independent and mutually exclusive. The outcome of one trial doesn't affect the outcome on other trials.
- Only one member will run from each party and each trial has a discrete number of possible outcomes.
- There would be n repeated trials. So, each riding can be considered a single trial.
- The probability that an outcome will occur is constant.

Using these assumptions, there would be 3 parameters required for this distribution:

- $\mu N = Probability\ of\ the\ NDP\ winning\ a\ riding$
- $\mu L = Probability\ of\ the\ liberals\ winning\ a\ riding$
- $\mu G = Probability\ of\ the\ Green\ party\ winning\ a\ riding$

What would be the value of the parameters μ for an election where the outcome is an equal chance of each party winning?

If the outcome is an equal chance of each party winning, then all the parameters μ will be equal and take the same value. According to our assumption, all the events are mutually exclusive and make up all the possible outcomes.

$$\mu N = \mu L = \mu G$$

$$\mu N + \mu L + \mu G = 1$$

$$\mu N = \mu L = \mu G = 0.33$$

What would be the value of the parameters μ for an election that is completely "rigged"? E.g. the party currently in power is going to win.

All the parameters would be zero except for the parameter for the party that rigged the election. Party that rigged the election, their parameter μ would take on a value of 1.

$$\mu_N = \mu_L = 0$$

$$\mu_G = 1$$

Specify a prior P(μ) that encodes a belief that one party has rigged the election, but there is an equal chance that it is any of the 3 parties.

When there is an equal chance that it is any of the 3 parties who has rigged the election, the **P(μ)** will encode 3 events.

P(μ) = {(1, Φ, Φ**), (**Φ,1, Φ**), (**Φ, Φ,1**)}**

The prior considers all the three possibilities that any one of them can rig the election.

Suppose my prior is that the Green Party has completely rigged the election. Assume I see a set of polls where the NDP has the largest share of the vote in each poll. What would be my posterior probability on μ?

Firstly, if Green Party has completely rigged the election, then the prior would be $\mu_G = 1$. And if polls where the NDP had the largest share of the vote the posterior would return 0 since the probability that $P(\mu_g = 1|D = majority\ NDP)$ would be = 0. Using the data given to maximize the likelihood estimation to get the best estimate of the parameters μ.

Suppose if party i is elected, they will set university tuition to be ti dollars. Write down an equation for the expected amount tuition will be, given a prior P(μ)

If party I is elected, the posterior probability will be 1 for that party and all other becomes zero. The expected amount of tuition for that party will be given a prior **P(μ)** would be the multiple of posterior and the tuition i.e. **μi * ti = ti**

2 Precision Per Data point

The relation between the log likelihood function and the sum of square function.

$$P(t|X,w,\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | w^T \phi(x_n), \beta_n^{-1})$$

for each data point, precision would be β_n. i.e $\left[\beta_n = \frac{1}{\sigma_n^2}\right]$

$X = \{x_1, \ldots x_n\}$ with corresponding target values $t_1, \ldots, t_n$. We assume these data points are drawn independently and follows a normal or gaussian distribution.

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$ where μ = mean & σ^2 = variance

Then, for each data point:

$$P(t|X,w,\beta_n^{-1}) = \mathcal{N}(t|y(x,w), \beta_n^{-1})$$

$$= \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n,w), \beta_n^{-1})$$

$$= \prod_{n=1}^{N} \sqrt{\frac{\beta_n}{2\pi}}\, e^{-\frac{\beta_n}{2}(t_n - y(x_n,w))^2}$$

Taking log on both sides.

$$\ln\left(P(t|X,w,\beta_n^{-1})\right) = \sum_{n=1}^{N} \ln\left(\sqrt{\frac{\beta_n}{2\pi}}\, e^{-\frac{\beta_n}{2}(t_n - y(x_n,w))^2}\right)$$

$$= \sum_{n=1}^{N} \frac{1}{2}\ln\beta_n - \ln(2\pi) - \frac{\beta_n}{2}(t_n - y(x_n,w))^2$$

$$= \frac{1}{2}\sum_{n=1}^{N} \ln\beta_n - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_{n=1}^{N} \beta_n (t_n - y(x_n,w))^2$$

We know, $E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2$ → sum of squared errors.

$$\underbrace{\ln\left(P(t|x,w,\beta_n^{-1})\right)}_{\text{log likelihood function}} = \frac{N}{2}\ln\beta_n - \frac{N}{2}\ln(2\pi) - \beta_n E_D(w)$$

$\beta_n E_D(w)$: sum of square error function.

3 Training vs. Test Error

For the questions below, assume that error means RMS (root mean squared error).

1. Suppose we perform unregularized regression on a dataset. Is the validation error always higher than the training error? Explain.

In general cases, the validation error tends to be higher than the training error, since the model is generated to minimize the error on the training data. However, there are cases where the validation error can be close or lower if the data is not randomized properly during initialization or the data set is too small. It is possible that the validation points happen to lie closer to the predicted curve than the training points.

2. Suppose we perform unregularized regression on a dataset. Is the training error with a degree 10 polynomial always lower than or equal to that using a degree 9 polynomial?
Explain.

It depends whether the data is normalized or not. For higher degree polynomial the training error will increase for un normalized data where as if it is normalized it will decrease as it would over fit.

3. Suppose we perform both regularized and unregularized regression on a dataset. Is the testing error with a degree 20 polynomial always lower using regularized regression compared to unregularized regression? Explain.

Regularization discourages the learning model to avoid the risk of over fitting, thus minimizing both the testing and training error. However, it is not definite that the testing error will always be lower or decrease when the model is regularized since the features that are used in the training set may not explain the model properly. The correlation of the data may vary which in turn may cause a high testing error despite the presence of a penalty term.

4 Basis Function Dependent Regularization

For L_1 / Lasso, $q=1$

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y_n(x_n,w) - t_n\}^2 + \lambda/2\, w^T w$$

For L_2 / Ridge, $q=2$

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y_n(x_n,w) - t_n\}^2 + \frac{\lambda}{2}|w|^2$$

Now for the gradient of the loss function with a choice of L_1 or L_2;

we assume.

$J_1 = \{w_0, w_2, w_4 .. w_n\}$ $\quad J_2 = \{w_1, w_3, w_5 .. w_n\}$

λ_L = penalty term for lasso $\quad \lambda_R$ = Penalty term for ridge

$$E(w) = \frac{1}{2}\sum_{n=1}^{n}\{y(x_n,w) - t_n\}^2 + \frac{1}{2}\sum_{J_1=1}^{L}\lambda_{J_1}^{(L)}|w_{J_1}| + \frac{1}{2}\sum_{J=2}^{K}\lambda_{J_2}^{(R)}(w_{J_2})^2$$

$$\nabla E(w) = 0^T = \left[\frac{\partial}{\partial w_0}(E(w)), \frac{\partial}{\partial w_1}(w), \frac{\partial}{\partial w_2}(E(w)), \dots\right]$$

For Lasso

$$\frac{\partial}{\partial w_L} = \sum_{n=1}^{n}\{y(x_n,w) - t_n)\,\phi(x_n) + \frac{\partial}{\partial w_L}\left[\frac{1}{2}\sum_{J_1=1}^{L}\lambda_{J_1}|w_{J_1}|\right]$$

For Ridge.

$$\frac{\partial}{\partial w_R} = \sum_{n=1}^{n}(y(x_n,w) - t_n)\,\phi(x_n) + \lambda_R w_R$$

5 Regression

5.1 Getting Started

Which country had the highest child mortality rate in 1990? What was the rate?
Niger had the highest under 5 mortality rates in 1990 with a mortality rate of 313.7 deaths per thousand live births.

Which country had the highest child mortality rate in 2011? What was the rate?
Sierra Leone had the highest under 5 mortality rates in 2011 with a mortality rate of 185.3 deaths per thousand live births.

Some countries are missing some features (see original .xlsx/.csv spreadsheet). How is this handled in the function assignment1.load unicef data()?

The function reads data from excel and uses float to convert the data from strings into doubles. If the data to be converted is not a number, (including underscores) the value for that property is replaced with the median value for that property.

5.2 Polynomial Regression

Plots of the training error and test error are shown below in Figure 1 and Figure 2. Figure 1 shows the results before the data was normalized. From the graph we can see that the training error is increasing with polynomial degree. This is since the data was not normalized before running the regression model. Figure 2 shows the results after the input data was normalized. After normalizing the data, the training error now decreases with an increasing polynomial degree.

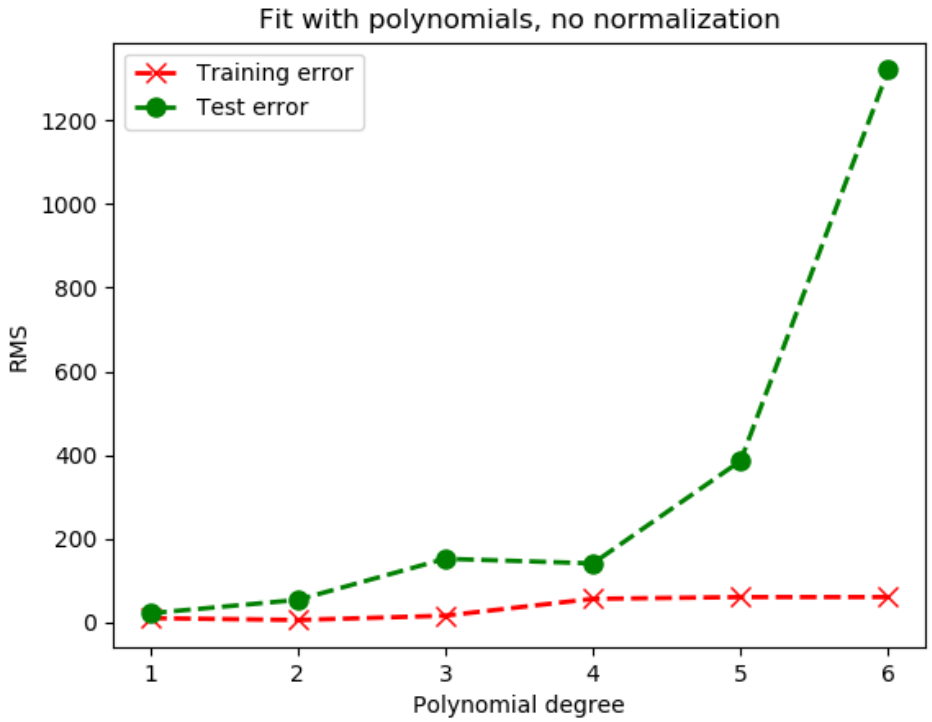


Figure 1: Testing error and training error vs polynomial degree before normalizing the input

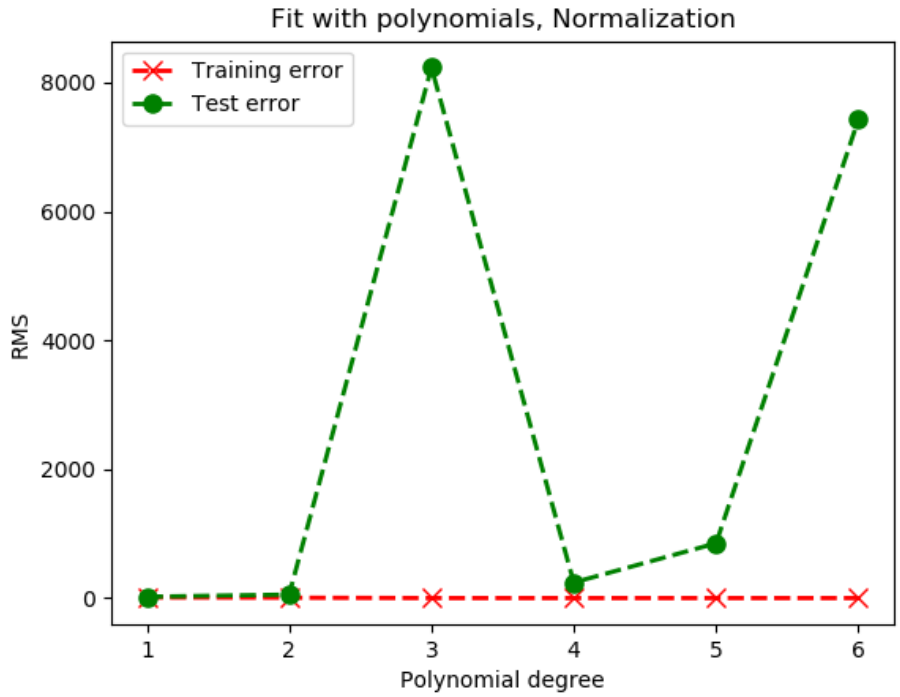


Figure 2: Testing error and training error vs polynomial degree after normalizing the input

For question 2, one dimensional polynomial regression was used. The results are shown below in a bar graph. Figure 3 shows the training error and test error for 8 input features. Figure 4,5,6 shows the individual features with 3rd degree polynomial regression.

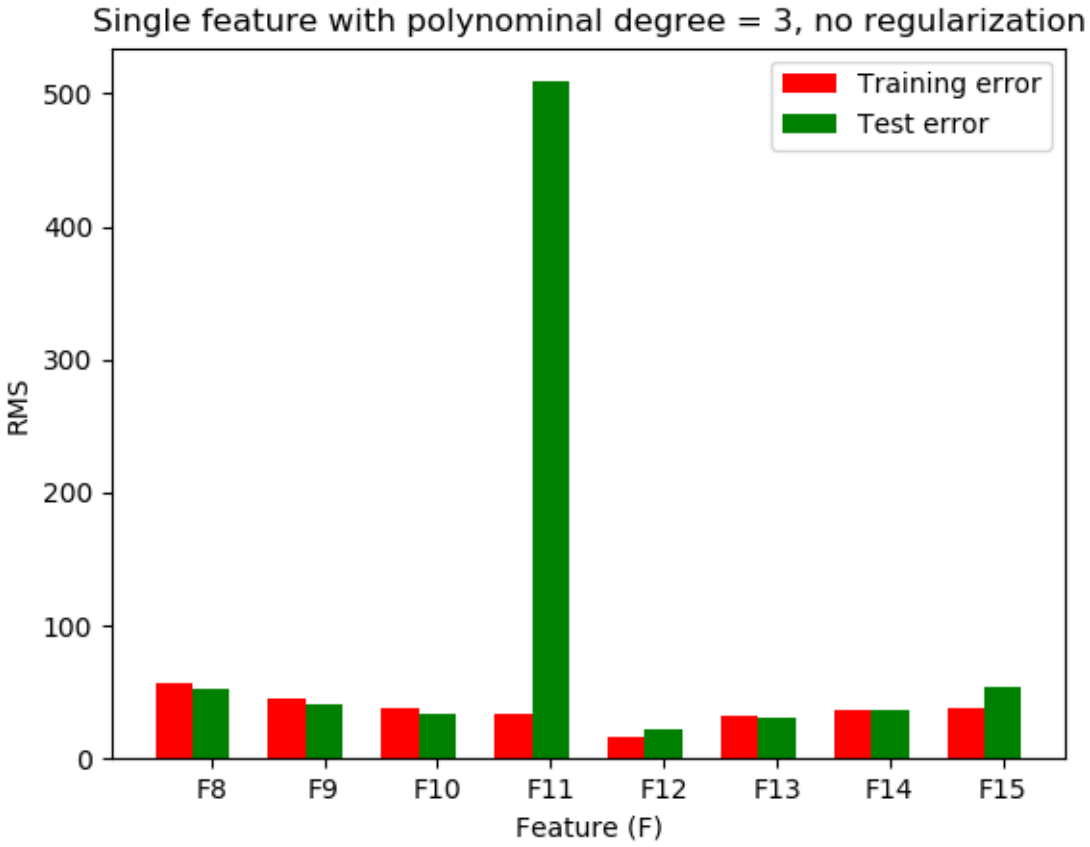


Figure 3: Training and Testing error for each feature

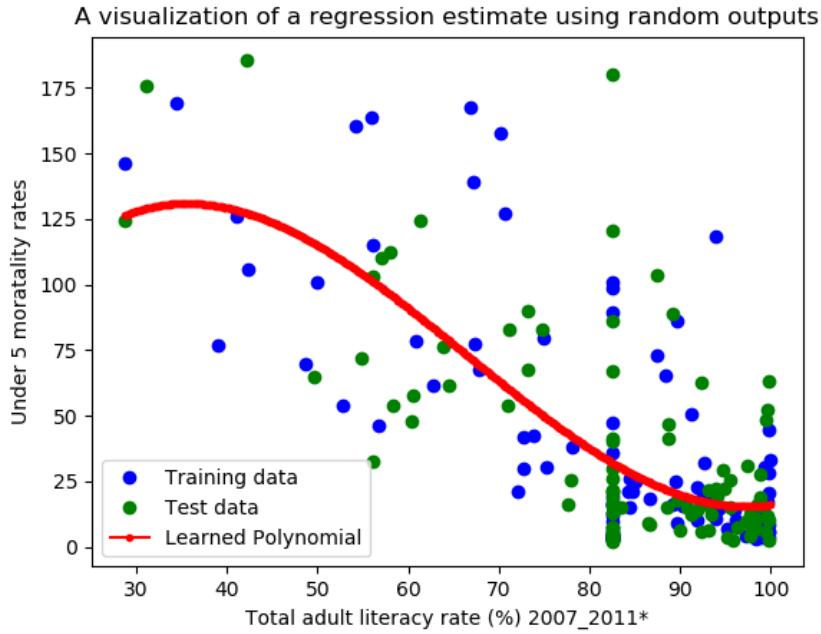
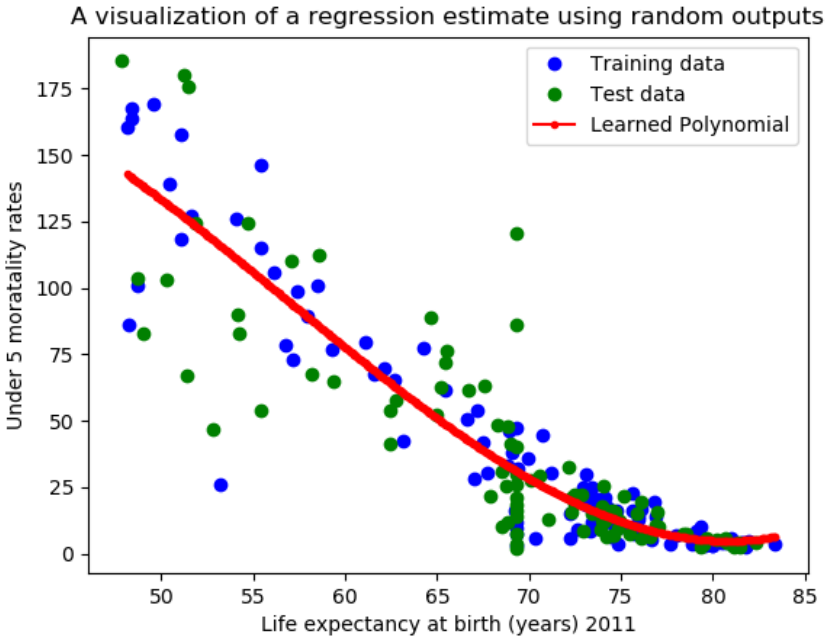


Figure 4: Plot of U5MR vs. GNI including the training set, test set, and the function obtained through regression



F

Figure 5: Plot of U5MR vs. Life Expectancy at birth including the training set, test set, and the function obtained through regression

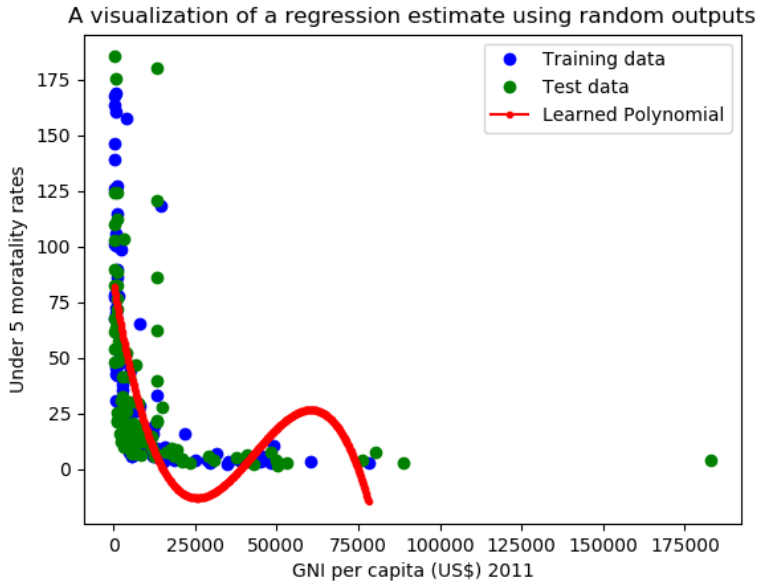


Figure 6: Plot of U5MR vs. adult literacy rate including the training set, test set, and the function obtained through regression

5.3 ReLU Basis Function

Train Error: 29.121771

Test Error: 34.223586

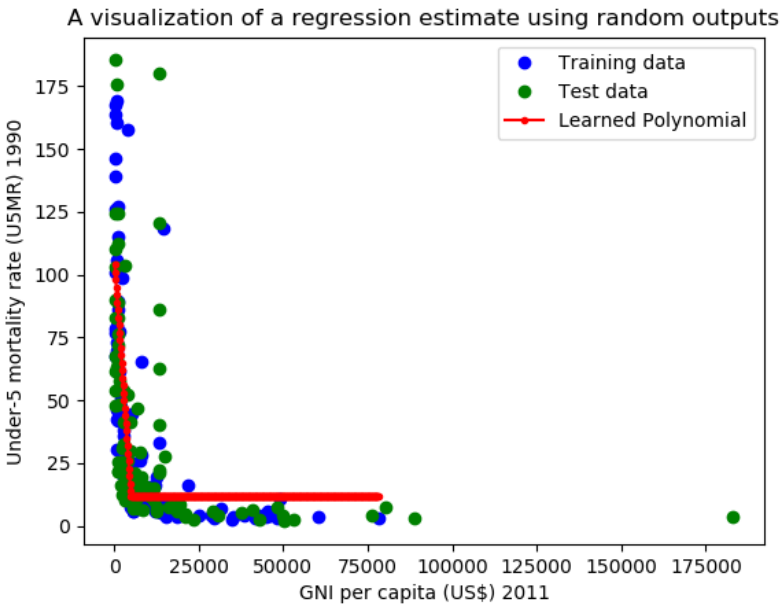


Figure 7: U5MR vs. GNI per capita fitted using Relu Regression

5.4 Regularized Polynomial Regression

Examining the graph, it appears that lambda = 100 or 1000 would give the best results. Moreover, it is to be noted that the validation error for lambda = 0 could not be plotted since the x –axis used a log scale. **The validation error for lambda = 0 was 96.23**.

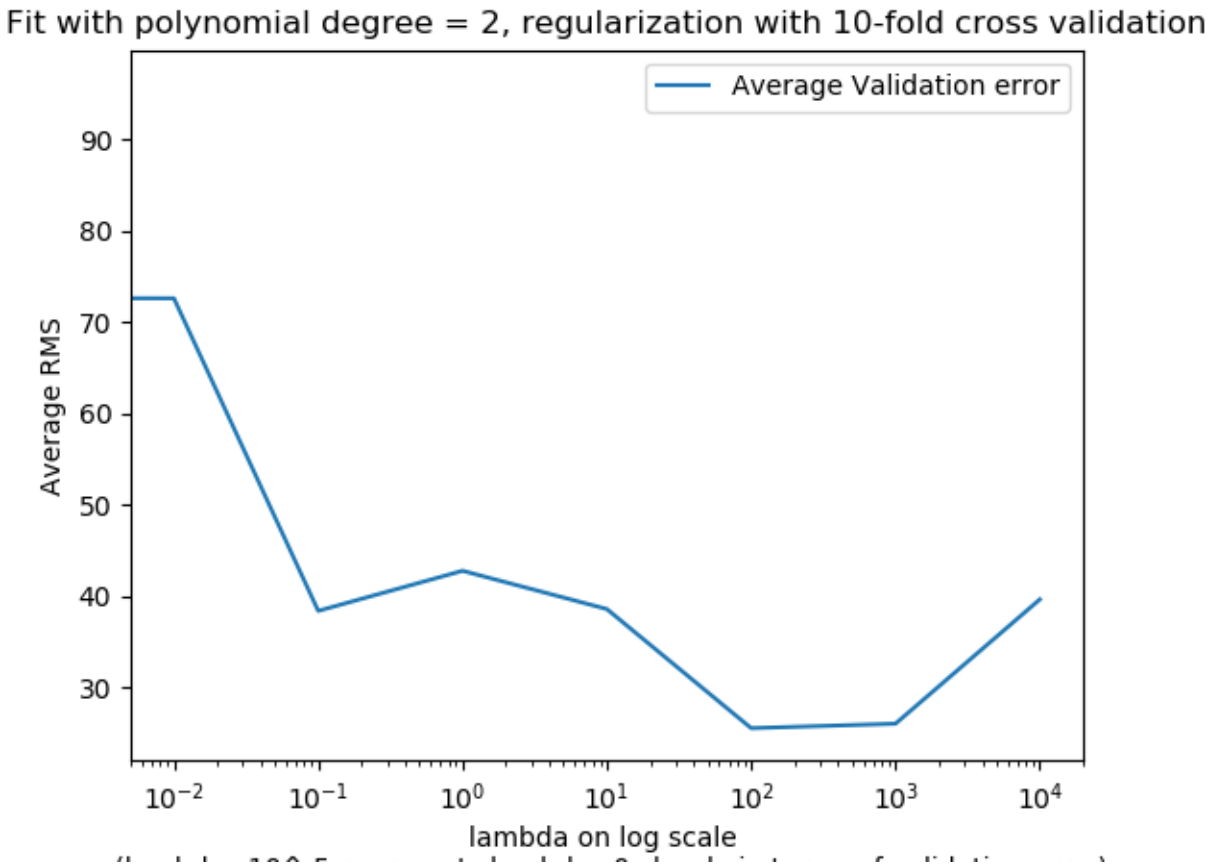


Figure 8: Average validation error for different regularization coefficients