ISYE6740-HW6

pkubsad

March 2021

1 **House price dataset.**

(a) (10 points) Fit the Ridge regression model to predict Price from all variable. You can use one-hot keying to expand the categorical variable Status. Use 5-fold cross validation to select the regularizer optimal parameter, and show the CV curve. Report the fitted model (i.e., the parameters), and the sum-of-squares residuals. You can use any package.

Answer: I have used sklearn's linear model to perform all the analysis for Q1. After using RidgeCV method withrange from range of alphas: [**0.005 - 5000000000.0**], the optimal alpha value determined by the library $\alpha = 0.10772173450159389$.

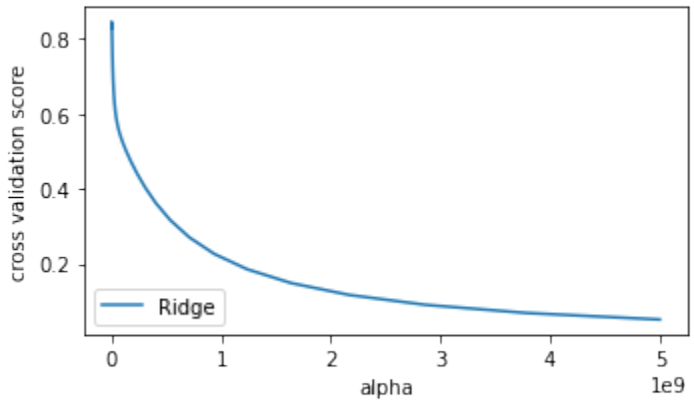

Figure 1: CV Curve

	0	1
0	50168.637286	Bathrooms
1	3632.595792	Bedrooms
2	-12274.491788	Foreclosure
3	1832.887507	Price/SQ.Ft
4	57635.404283	Regular
5	-19686.818918	Short Sale
6	207.593149	Size

Figure 2: Fitted Model

```
range of alphas: 0.005 5000000000.0
optimal alpha:  0.10772173450159389
mse for ridge regression:  21846054179.982845
r2 for ridge regression:  0.8204506806336742

array([[ 50168.63728641,   3632.59579198, -12274.49178788,
         1832.88750652,  57635.40428256, -19686.81891839,
          207.59314895]])
```

Figure 3: Residuals - Ridge

(b) (10 points) Use lasso to select variables. Use 5-fold cross validation to select the regularizer optimal parameter, and show the CV curve. Report the fitted model (i.e., the parameters selected and their coefficient). Show the Lasso solution path. You can use any package for this.

Answer: In the same lines as above, used sklearn's linear model library to perform this analysis. The output from LassoCV gives us the optimal regularizer parameter, $\alpha = 100.8230376891931$. Below are the illustrations of CV Curve, fitted model and solution path:

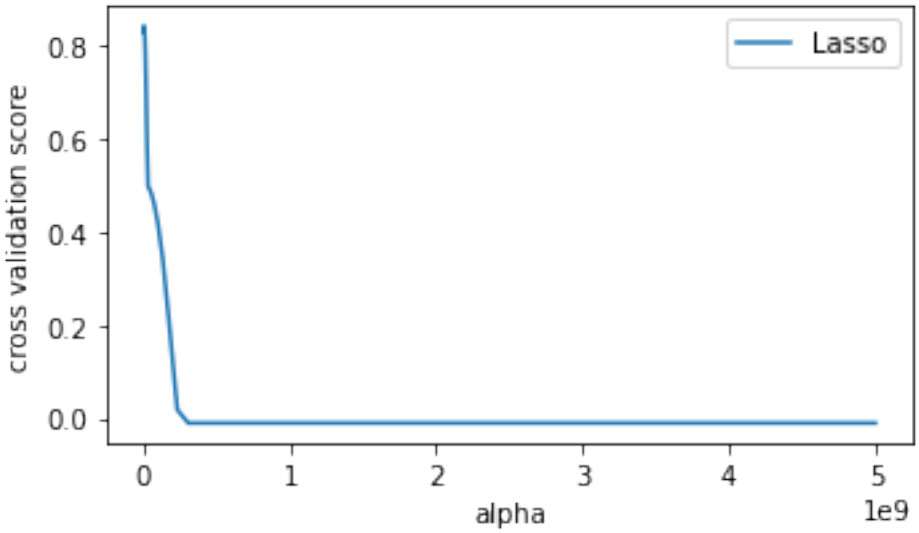

Figure 4: CV Curve - LASSO

	0	1
0	14743.630869	Bathrooms
1	-0.000000	Bedrooms
2	0.000000	Foreclosure
3	2018.141742	Price/SQ.Ft
4	42604.790481	Regular
5	-1962.037890	Short Sale
6	252.467580	Size

Figure 5: Lasso Model

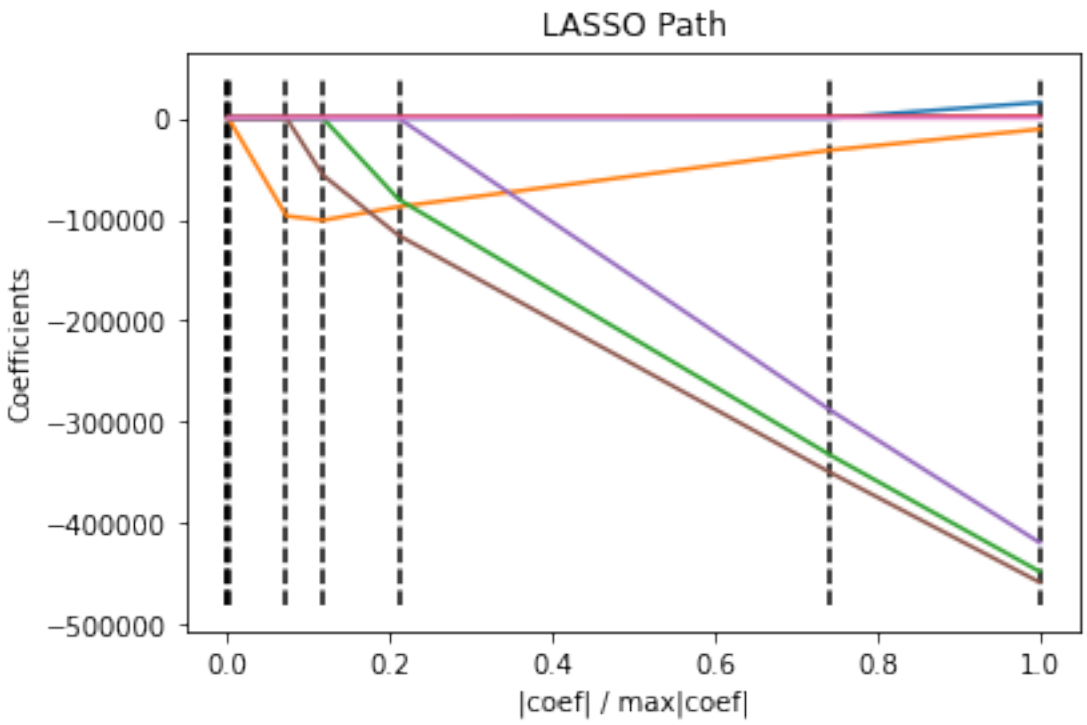

Figure 6: Lasso Path

(c) (5 points) Use elastic net to select variables. Report the fitted model (i.e., the parameters selected and their coefficient). Use 5-fold cross validation to select the regularizer optimal parameter. You can use any package for this.

Answer. Same lines as above, regularizer optimal parameter , $\alpha = 466.30167344161$

	0	1
0	15.764964	Bathrooms
1	-3.574742	Bedrooms
2	-1.649758	Foreclosure
3	2045.073923	Price/SQ.Ft
4	20.706935	Regular
5	-18.056825	Short Sale
6	267.275949	Size

Figure 7: Elastic Net Model

2 **AdaBoost.** (25 points)

(a) (15 points) For each iteration $t = 1, 2, 3$, compute ϵ_t, α_t, Z_t, D_t by hand (i.e., show the calculation steps) and draw the decision stumps on the figure (you can draw this by hand).

Answer. Dataset: $X, Y, Z = \{-1, 0, +1\}, \{-0.5, 0.5, +1\}, \{0, 1, -1\}, \{0.5, 1, -1\}, \{1, 0, +1\}, \{1, -1, +1\}, \{0, -1, -1$
Iteration 1: Classifier: If $X < -0.25$ then $+1$ and $X > -0.25$ then -1

Result of Classification: = [1,1,-1,-1,-1,-1,-1,-1]

$D_1 = 1 * 1/m = 1 * 1/8 = [0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125]$

$\epsilon_t = \frac{misclassifiedpoints}{totalpoints} = \frac{2}{8} = 0.25$

$\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t}) = \frac{1}{2} ln(\frac{1-0.25}{0.25}) = 0.54930614433$

$Z_t = \sum_i^m D_t(i) e^{-\alpha_t y^i h_t(x^i)} = 6 * 0.125 * e^{-\alpha_t} + 2 * 0.125 e^{\alpha_t} = 0.8660225$

$D_2 = \frac{D_1}{Z_1} e^{-\alpha_t y^i h_1(x^i)} = [0.08333333, 0.08333333, 0.08333333, 0.08333333, 0.25, 0.25, 0.08333333, 0.08333333]$

Calculating in the same lines:
Iteration 2: Classifier: If $X > +0.75$ then $+1$ and $X < 0.75$ then -1
Iteration 3: Classifier: If $Y > -0.25$ then $+1$ and $Y < -0.25$ then -1

	ϵ_t	α_t	Z_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(4)$	$D_t(5)$	$D_t(6)$	$D_t(7)$	$D_t(8)$
1	0.25	0.549306	0.866025	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
2	0.166667	0.804719	0.745356	0.083333	0.083333	0.083333	0.083333	0.25	0.25	0.083333	0.083333
3	0.1	1.098612	0.6	0.25	0.25	0.05	0.05	0.15	0.15	0.05	0.05

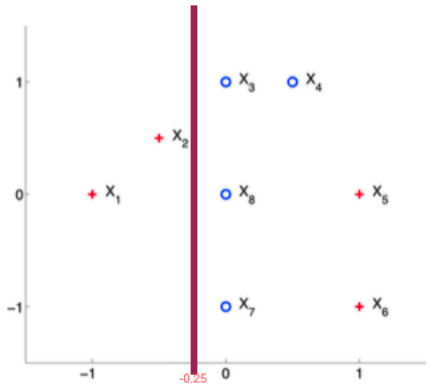

Figure 8: Iteration 1

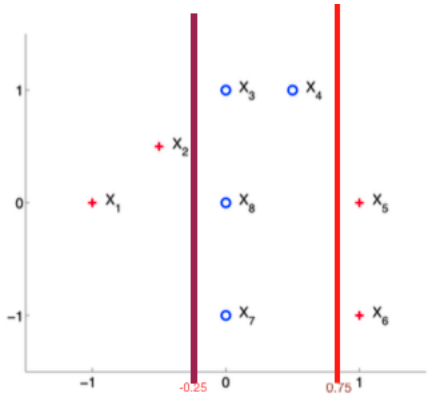

Figure 9: Iteration 2

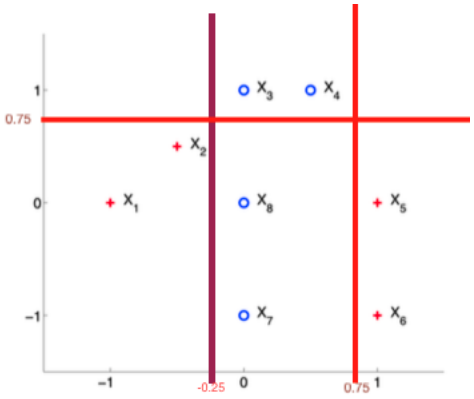

Figure 10: Iteration 3

(b) (10 points) What is the training error of this AdaBoost? Give a short explanation for why AdaBoost outperforms a single decision stump.

Answer.

The training error of AdaBoost is 0. All the points are classified correctly. Adaboost algorithm classifies at multiple iterations. At each iteration, the classification is focussed on particular part of the dataset. With wieghted focus in each iteration, this can create better classifier than other classification models.

3 **Random forest and one-class SVM for email spam classifier** (25 points)

(a) (5 points) Build a CART model and visualize the fitted classification tree.

Answer.

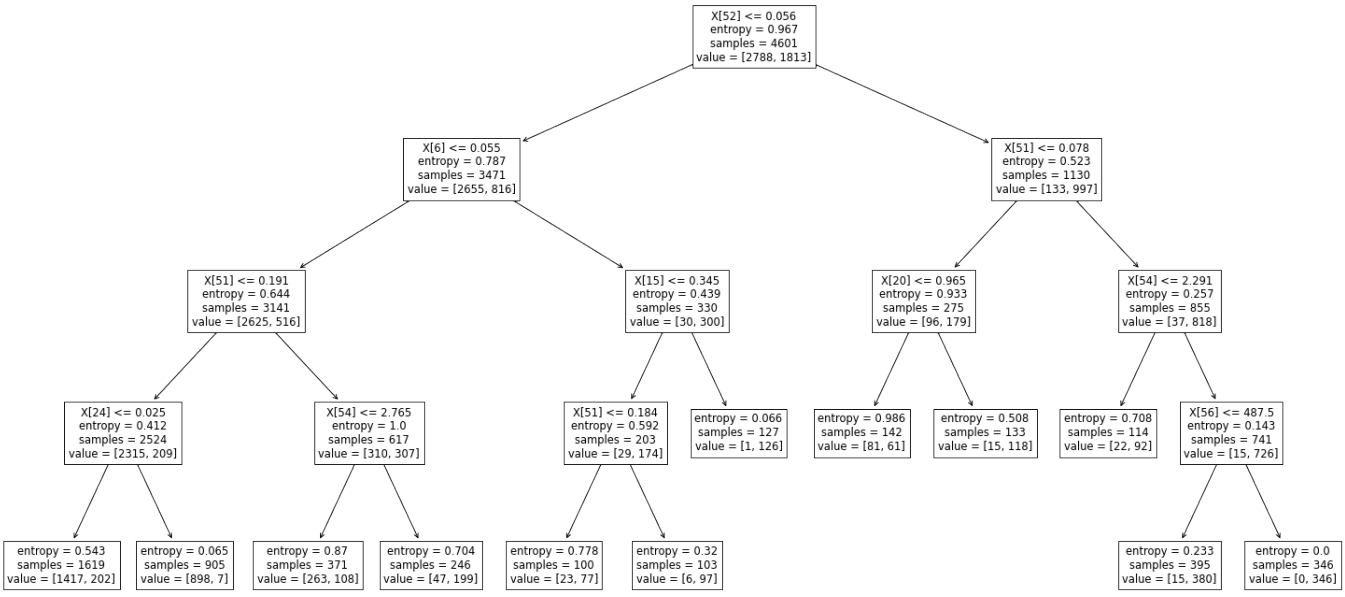

Figure 11: CART model tree classification

(b) (10 points) Now also build a random forest model. Partition the data to use the first 80% for training and the remaining 20% for testing. Compare and report the test error for your classification tree and random forest models on testing data. Plot the curve of test error (total misclassification error rate) versus the number of trees for the random forest, and plot the test error for the CART model (which should be a constant with respect to the number of trees).

Answer. I have used sklearn's Random Forest classifier and looping through nTrees =[1,1000]. For each iteration of nTree, I have calculated the misclassification rate as

$$miscalculation\_rate = \frac{FalsePositives + FalseNegatives}{TotalNumberofdatapoints}$$

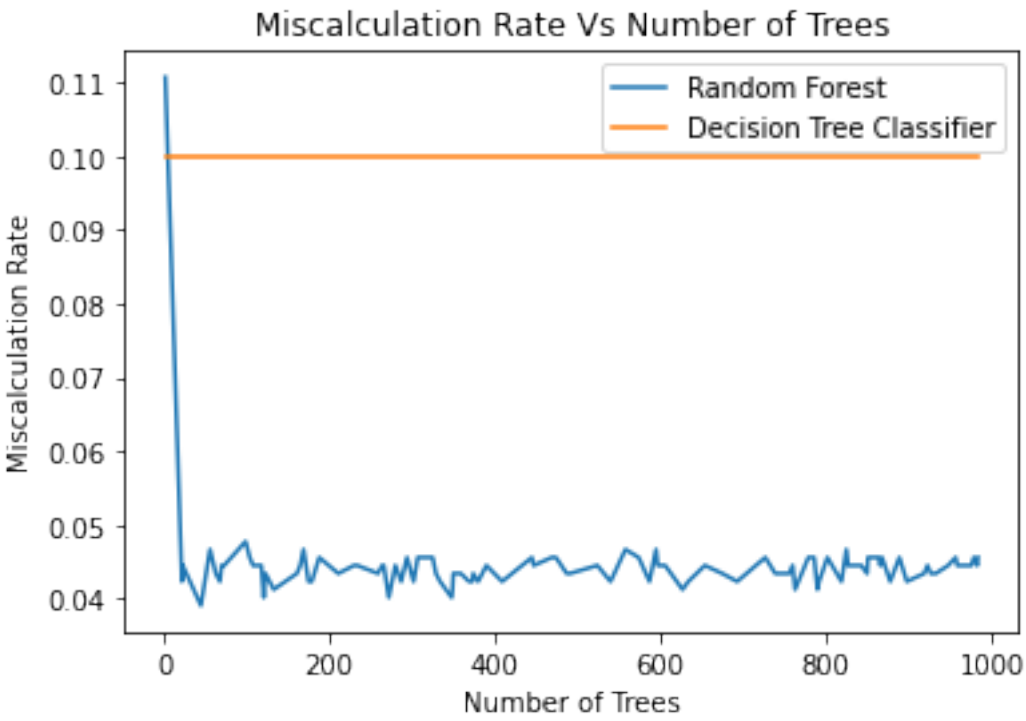

Figure 12: Miscalculation Rate vs Number of Trees

Used sklearn's GridSearchCV method to calculate the optimal Decision Tree Classifier and ran the same classification with Random Forest and Decision Tree. Decision tree is single tree classifier and hence the misclassfication rate stays the same irrespective of the number of trees used for classification. The results of the classification are as follows (Confusion matrix, miscalculation rate):

```
confusion matrix for random forest:
[[520  11]
 [ 31 359]]
miscalc rate for random forest:
0.04560260586319218
================
confusion matrix for decision tree:
[[513  18]
 [ 74 316]]
miscalc rate for decision tree:
0.0998914223669924
```

Figure 13: Miscalculation Rate of Random Forest Classifier vs Decision Tree Classifier

(c) (10 points) Now we will use a one-class SVM approach for spam filtering. Partition the data to use the first 80% for training and the remaining 20% for testing. Extract all *non-spam* emails from the training block (80% of data you have selected) to build the one-class kernel SVM using RBF kernel (you can turn the kernel bandwidth to achieve good performance). Then apply it on the 20% of data reserved for testing (thus this is a novelty detection situation), and report the total misclassification error rate on these testing data.

Answer. I have split the data using sklearn's train_test_split method to split the data 80,20. In the train set, filtered it for col 57 == 0 (Non-spam emails). Used sklearn's OneClassSVM on the filtered data set to build the model. Observations from the result:
Number of data point in training set = **3680**
Number of non-spam emails in training set=**2257**
Number of spam email in training set = **1423**
Number of data point in testing set = **921**
Number of non-spam emails in testing set=**531**
Number of spam email in testing set = **390**

Number of spam email reported by model's prediction in training set = **1581**
Number of spam email reported by model's prediction in testing set = **527**
Miscalculation rate = **0.2975**

Snapshot from programme's output:

```
(3680, 58)
(921, 58)
counts in train data:
0    2257
1    1423
Name: 57, dtype: int64
counts in test data:
0    531
1    390
Name: 57, dtype: int64
No of spams detected by model in training set:
1581
No of spams detected by model in test set:
527
No of non-spams detected by model in test set:
394
miscalculation rate:
0.2975027144408252
```

Figure 14: Programme output for one class svm

4.0 **Locally weighted linear regression and bias-variance tradeoff.** (25 points)

a (5 points) Show that the ridge regression which introduces a squared ℓ_2 norm penalty on the parameter in the maximum likelihood estimate of β can be written as follows

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ (X\beta - y)^T W (X\beta - y) + \lambda \|\beta\|_2^2 \right\}$$

for property defined diagonal matrix W, matrix X and vector y.

Answer.: Given linear regression model

$$y_i = \beta^{*^T} . x_i + \epsilon_i$$

We have to determine the likelihood function $L(\beta, (x_i, y_i))$ and maximum likelihood function will be deteremined by doing partial derivative wrt β. $\frac{\partial L(\beta,(x_i,y_i))}{\partial \beta} = 0$. Recall that $y_i \tilde{} N(x_i \beta^{*^T}, \sigma^2)$, we can establish the Ridge rigression's penalty term as :

$$\hat{\beta}(\lambda) = argmin_{\beta} \{ -\frac{1}{2} \sum (y_i - x_i\beta)^2 + \lambda ||\beta||_2^2 \}$$

Reduced to matrix form, we can write $\sum (Y - X\beta)^2 = ||Y - X\beta||_2^2 = (Y - X\beta)^T (Y - X\beta)$. Replacing in the equation

$$\hat{\beta}(\lambda) = argmin_{\beta} \{ -\frac{1}{2} (Y - X\beta)^T (Y - X\beta) + \lambda ||\beta||_2^2 \}$$

Adjusting for constant and sign,as it is a derivative equal to 0

$$\hat{\beta}(\lambda) = argmin_{\beta} \{ (X\beta - Y)^T (X\beta - Y) + \lambda ||\beta||_2^2 \}$$

Introducing the wieghtage matrix (W) representing the local weights, we can rewrite the equation as

$$\hat{\beta}(\lambda) = argmin_{\beta} \{ (X\beta - Y)^T . W (X\beta - Y) + \lambda ||\beta||_2^2 \}$$

W is a nXn dimensional diagonal matrix with $W_{ii} \varepsilon [0, 1]$ representing the weight of the i^{th} observation.

b (5 points) Find the close-form solution for $\widehat{\beta}(\lambda)$ and its distribution conditioning on $\{x_i\}$.

Answer. Source: https://arxiv.org/pdf/1509.09169.pdf, page 50
In order to find the closed form solution of the $\widehat{\beta}$, we take the derivative of $\hat{\beta}(\lambda)$ with respect to β equal to zero. Then we solve for $\hat{\beta}$:

$$\frac{\partial \hat{\beta}(\lambda)}{\beta} = 0$$

$$2X^T WY - 2X^T WX\beta - 2\Delta\beta + 2\Delta\beta_0 = 0. where \Delta = \lambda I$$

This is solved by:

$$\widehat{\beta(\lambda)} = (X^T WX + \Delta)^{-1} (X^T WY)$$

$$\widehat{\beta(\lambda)} = (X^T WX + \lambda I)^{-1} (X^T WY)$$

The Expectation function is given by

$$E(\hat{\beta}) = (X^T WX + \lambda I)^{-1} (X^T WX\beta)$$

The variance of β is given by

$$Var[\hat{\beta}] = \sigma^2(X^TWX + \lambda I)^{-1}X^TW^2X(X^TWX + \lambda I)^{-1}$$

The error term is gaussian, y is guassian, $\hat{\beta}$ is also gaussian $\hat{\beta} \sim \mathcal{N}(\mathbb{E}[\hat{\beta}], Var[\hat{\beta}])$

c (5 points) Derive the bias as a function of λ and some fixed test point x.

Answer. As mentioned in the hw walk through we have to find Bias term as,

$$Bias = \mathbb{E}[\hat{\beta}(\lambda)^T x] - \beta^* x$$

$$Bias = \mathbb{E}[((X^TWX + \lambda I)^{-1}(X^TWY))^T x] - \beta^* x$$

Based on conditional probability, we can use the $E(\gamma y) = \gamma E(y) = \gamma X\beta^*$, for any given random varaible y.

$$Bias = x\mathbb{E}[((X^TWX + \lambda I)^{-1}(X^TWY))^T] - \beta^* x$$

$$Bias = x\mathbb{E}[((X^TWX + \lambda I)^{-1}(X^TWY))^T] - \beta^* x$$

$$Bias = x((X^TWX + \lambda I)^{-1}(X^TWX\beta^*))^T - \beta^* x$$

d (5 points) Derive the variance term as a function of λ.
Answer. We can calculate the variance term as follows:

$$VAR(\hat{\beta}(\lambda)^T x) = x.VAR(\hat{\beta}(\lambda)^T$$

substituting the equation from q 4.ii for $VAR[\hat{\beta}(\lambda)]$

$$= x(\sigma^2(X^TWX + \lambda I)^{-1}X^TW^2X(X^TWX + \lambda I)^{-1})^T$$

e (5 points) Now assuming the data are one-dimensional, the training dataset consists of two samples $x_1 = 1.5$ and $x_2 = 1$, and the test sample $x = 0.5$. The true parameter $\beta_0^* = 1$, $\beta_1^* = 0.5$, the noise variance is given by $\sigma_1^2 = 2$, $\sigma_2^2 = 1$. Plot the MSE (Bias square plus variance) as a function of the regularization parameter λ.

Answer.

$$MSE = Bias^2 + Variance$$

$$MSE = [x((X^TWX+\lambda I)^{-1}(X^TWX\beta^*))^T - \beta^* x]^2 + [x(\sigma^2(X^TWX+\lambda I)^{-1}X^TW^2X(X^TWX+\lambda I)^{-1})^T]$$