Midterm 2 Solution

November 27, 2019

Problem 1

1. The weight of incorrectly classified sample increases (shown as the blue "+" sign). And other weights are not changed.

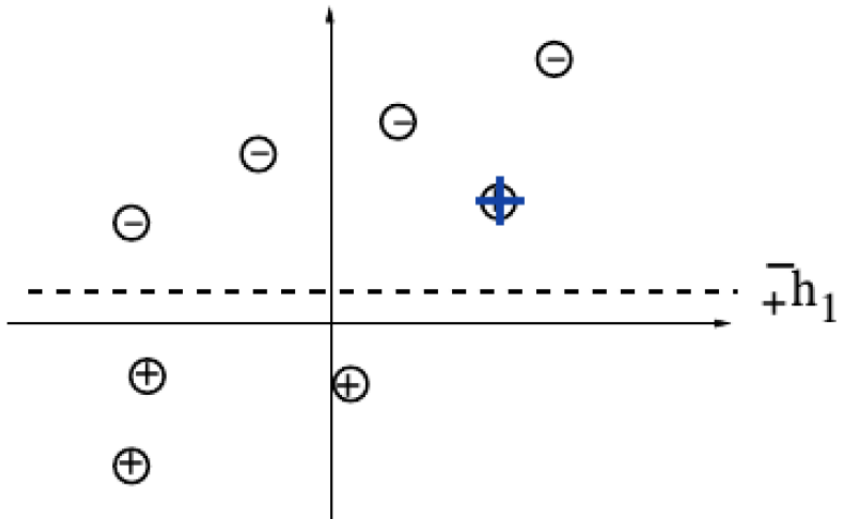


Figure 1: Weights $D_2(i)$ of each sample

2. Given $D_1(i) = \frac{1}{8} \quad \forall i$,

$$\epsilon_1 = \sum_{i=1}^{8} D_1(i) \mathbb{I}\{y_i \neq h_1(x_i)\} = \frac{1}{8}$$

Then $\alpha_1 = \frac{1}{2} ln(\frac{1-\epsilon_1}{\epsilon_1}) = \frac{1}{2} ln(7) = 0.9730$.

3. False.
   The votes α_i tend to increase as the algorithm proceeds. As the algorithm proceeds and we add classifiers with $\alpha_i > 0$, the weighted training error E_i decreases.

4. True.
   As defined in class, AdaBoost will choose classifiers with training error above $1/2$. Note that if the classifier does worse than $1/2$ we can always assign the sign of its predictions and therefore get a classifier that does slightly better than $1/2$.

Problem 2

1. **How many instances and how many features for each instance in the data set?**

   There are 4601 instance and 57 features in the data. The last column is the label for spam/regular instances.

   How many instances of spam versus regular emails are there in the data?

   There are 2788 emails labeled as 0 and 1813 labeled as 1

2. **Decision Tree Classifier** This is a decision tree with a depth of 3.

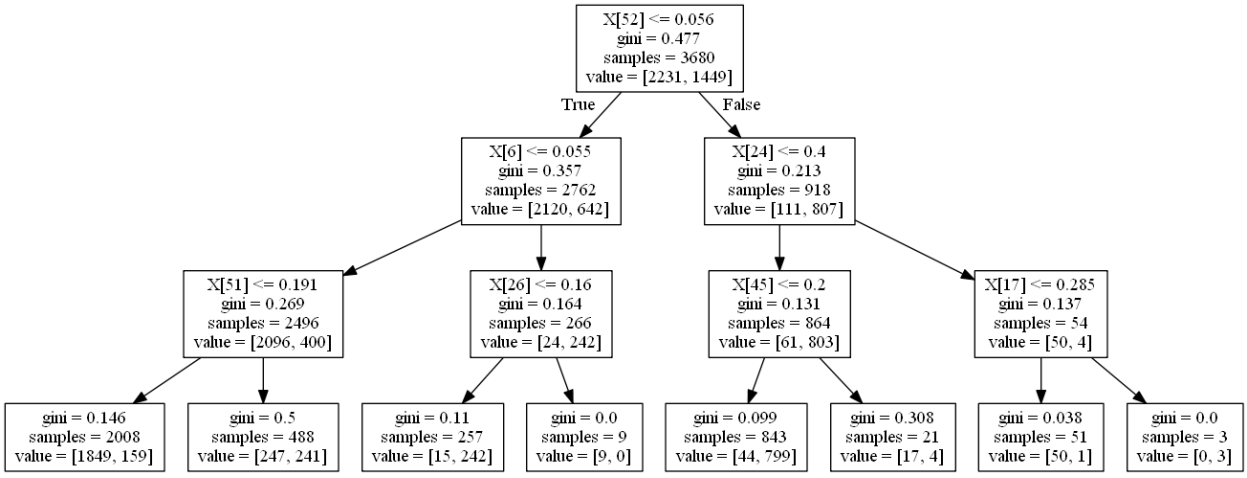
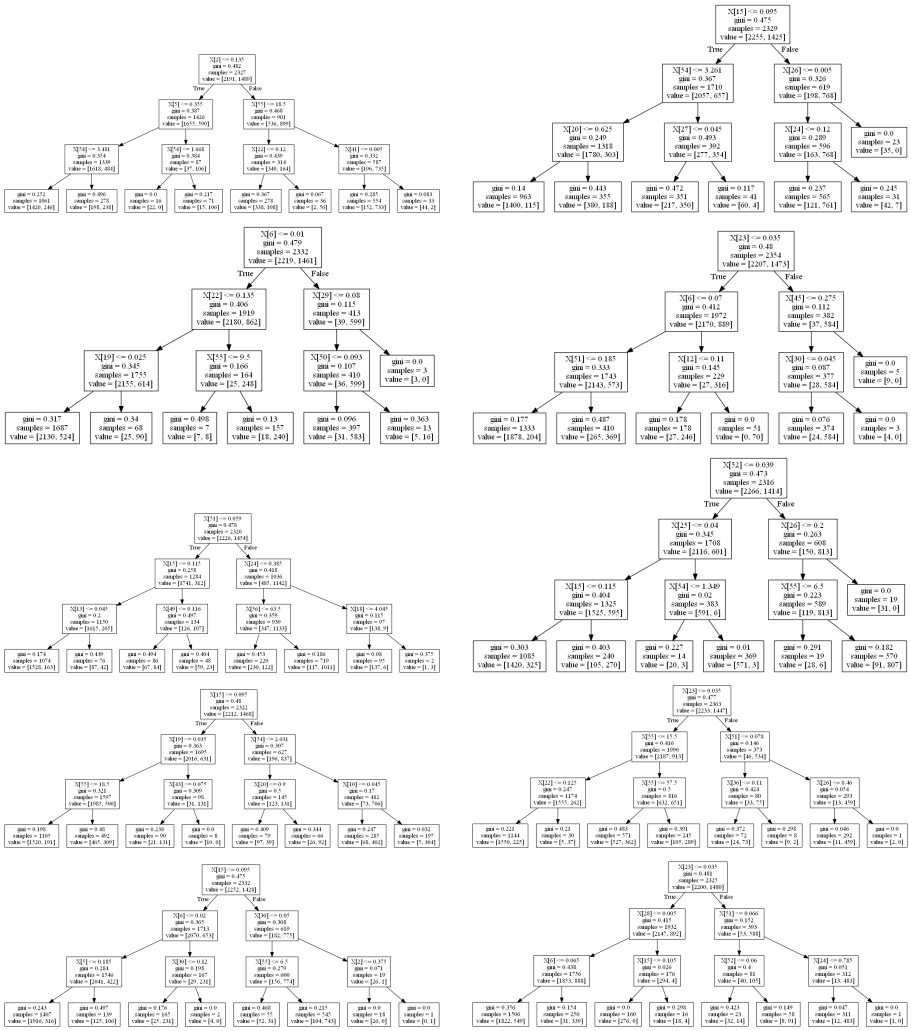


Figure 2: Decision Tree

The **prediction accuracy** of a decision tree is**88.2%**

It will be hard to visualize a large tree here. Please check students' code here to see if the decision is reasonable.

3. **Random Forest Classifier** This is a random forest with a max depth 3 for every estimator. As python default setting, there are 10 decsion trees in total in the forest.



The **prediction accuracy** for random forest classifier is **88.7%**

It is hard to visualize all trees in assignment report. Student can just plot one or more sample here. Please check the code that they are using correct algorithm.

4. **AUC curve:**

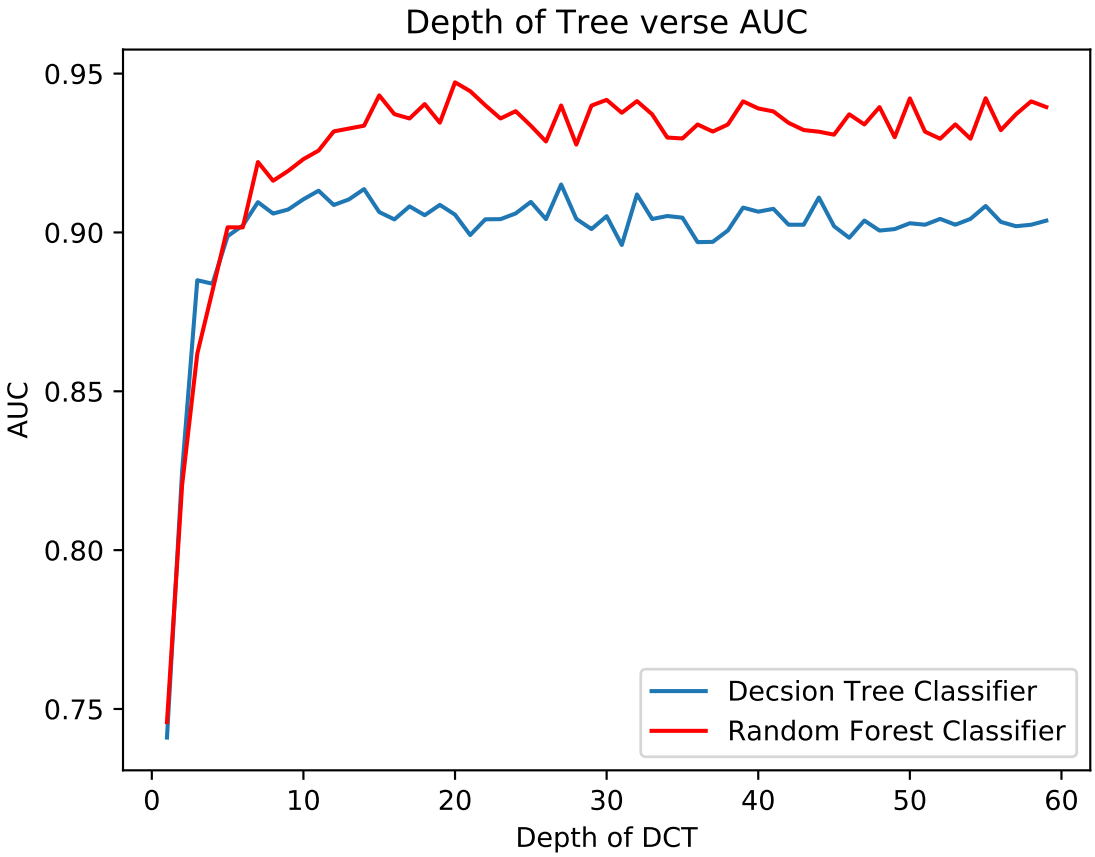
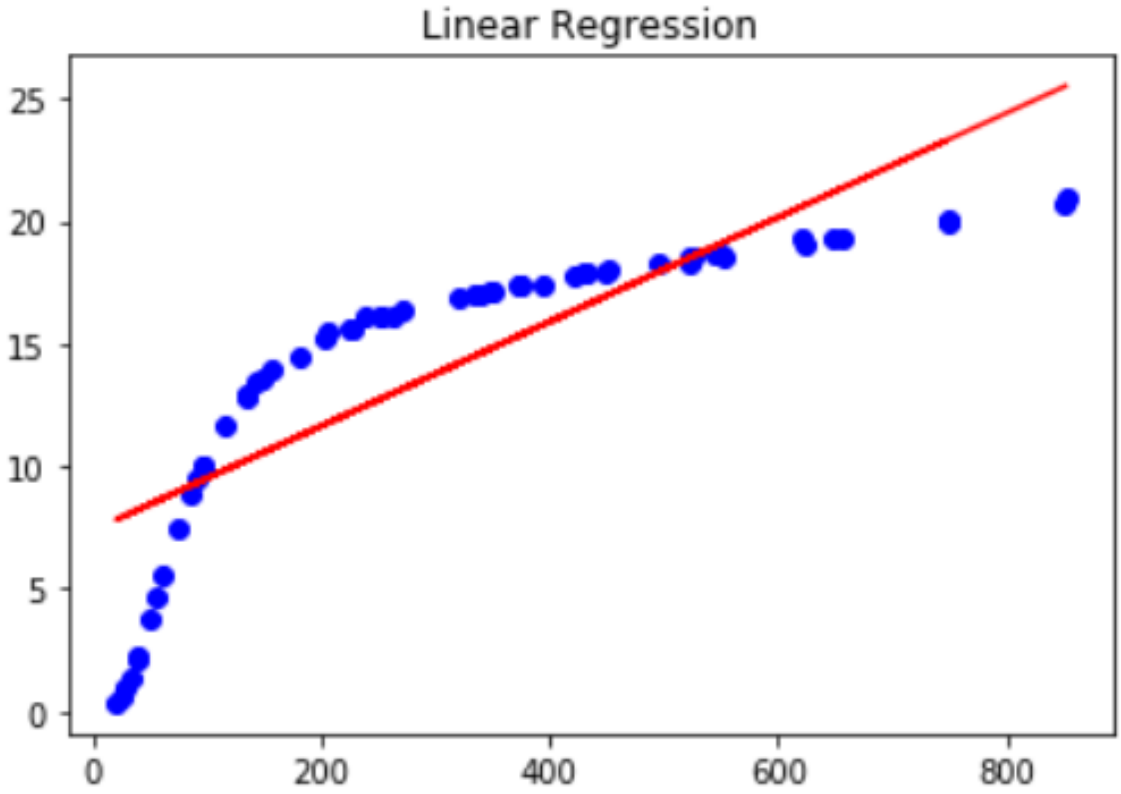


Figure 3: AUC curve

The x-axis can be max depth,max number of leaf nodes, average tree size(total number of nodes), or other similar features of decision tree which is related to tree size.

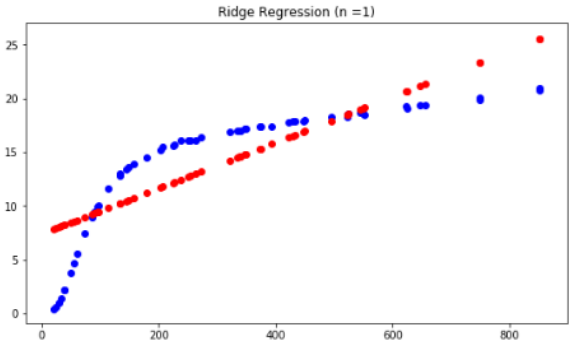
Problem 3

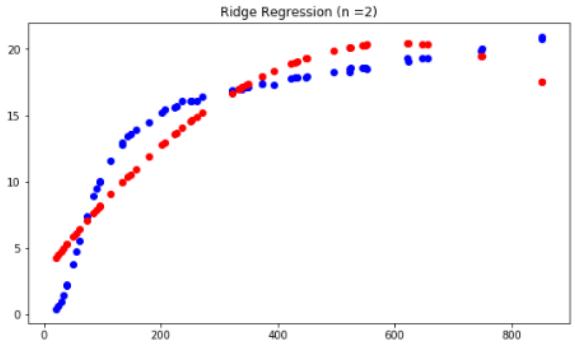
1. **Perform linear regression on the data. Report the fitted model and the fitting error.**

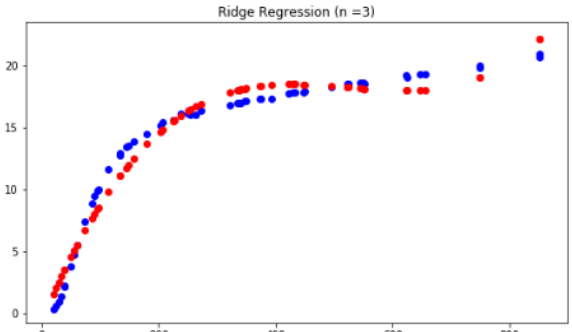


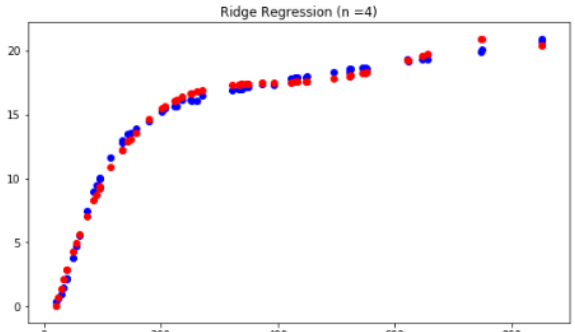
The function is $y = 6.895 + 0.0217x$ The error of linear regression model(MSE) is 11.006

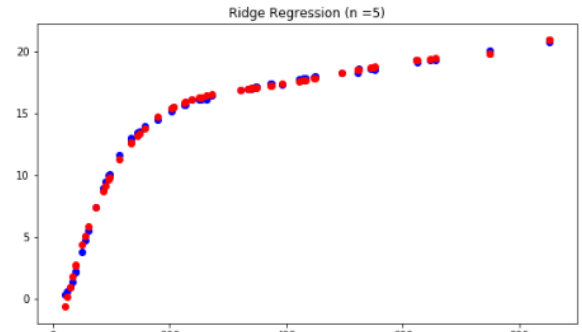
2. **Perform nonlinear regression with polynomial regression function up to degree n = 10 and use ridge regression (see Lecture Slides for "Bias-Variance Trade-off"). Write down your formulation and strategy for doing this, the form of the ridge regression.**

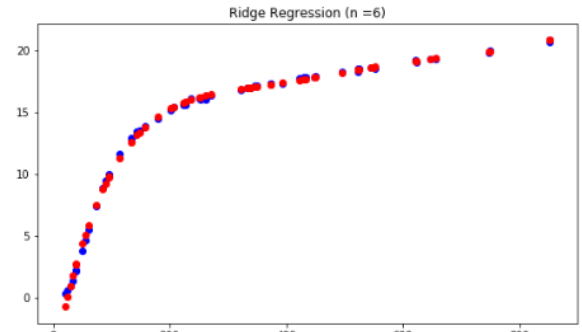

Ridge Regression (n =1)

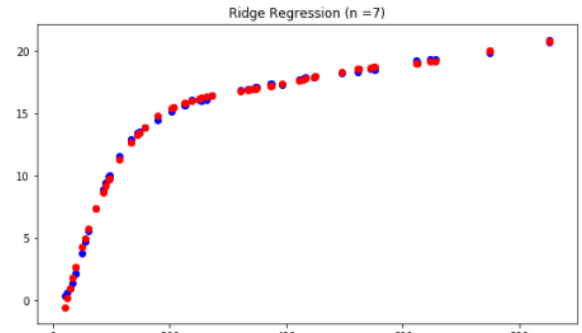

Ridge Regression (n =2)

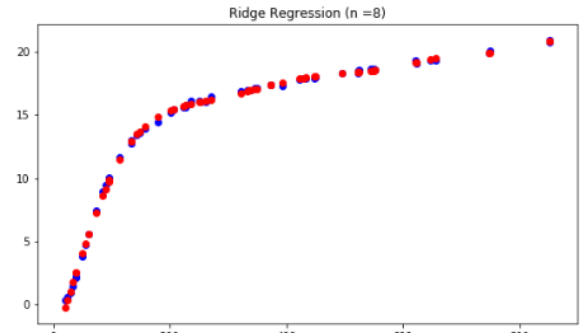

Ridge Regression (n =3)

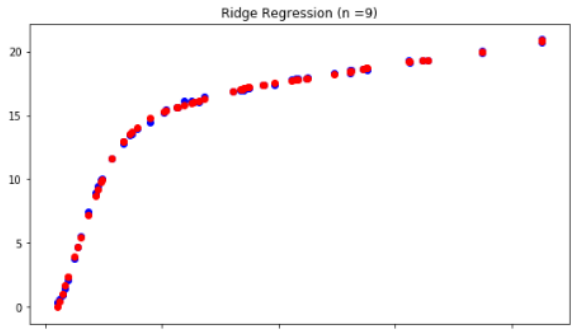

Ridge Regression (n =4)

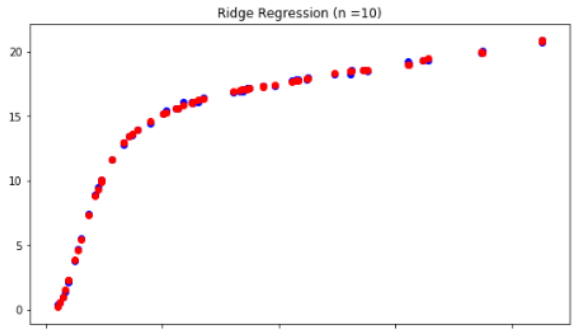

Ridge Regression (n =5)


Ridge Regression (n =6)

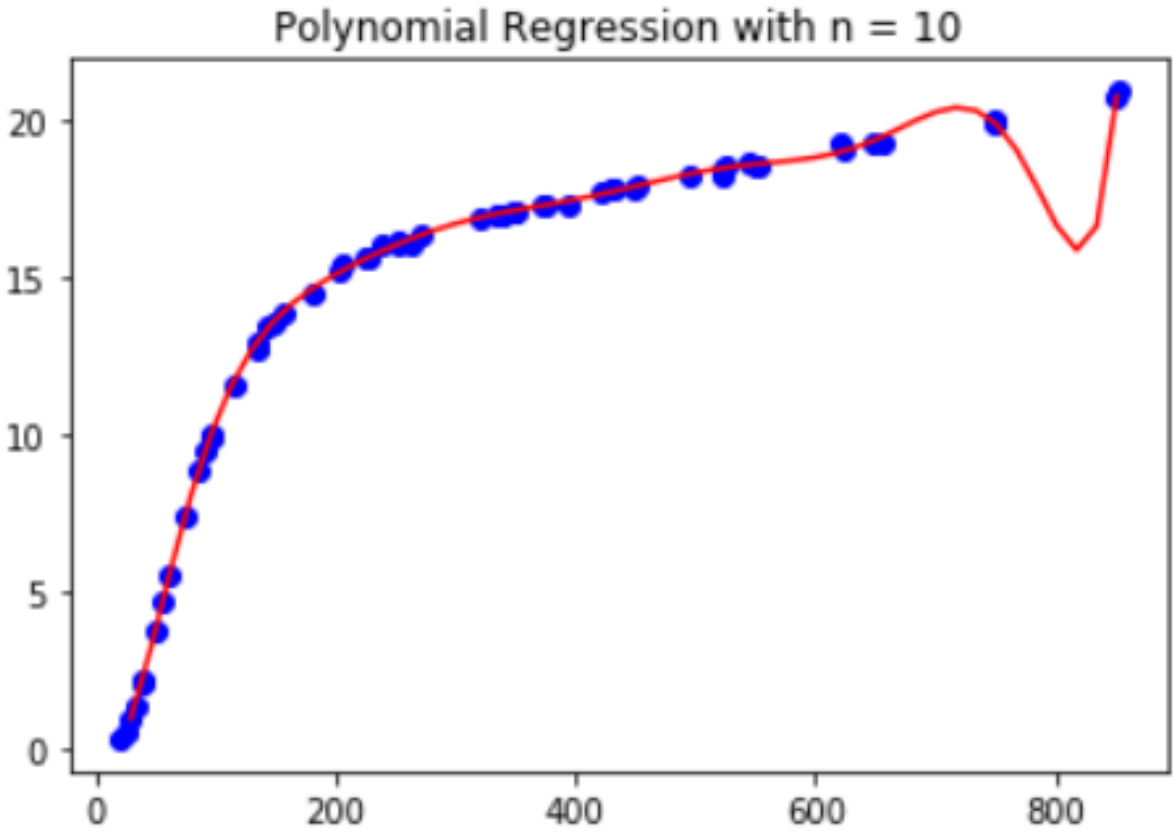

Ridge Regression (n =7)


Ridge Regression (n =8)


Ridge Regression (n =9)

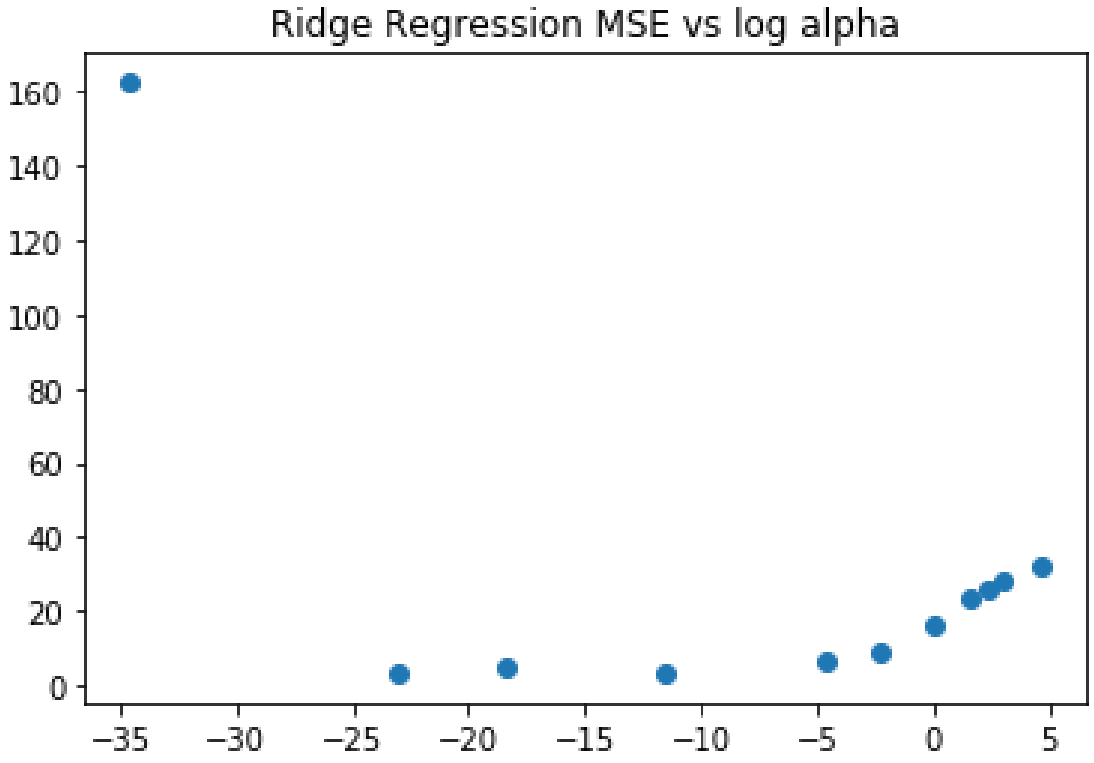

Ridge Regression (n =10)

After we got the polynomial models with different degree (=n), we need to check which one has the lowest MSE among those. We got n=8 as the optimal model, and found what are the coefficients of the model. And our formulation is $f(x) = 1.900 * 10^{-0.1}x - 2.848 * 10^{-04}x^2 - 3.347 * 10^{-06}x^3 + 1.989 * 10^{-08}x^4 - 4.887 * 10^{-11}x^5 + 6.293 * 10^{-14}x^6 - 4.170 * 10^{-17}x^7 + 1.124 * 10^{-20}x^8$



As you can see from the last graph, if n is too large, it might be overfitted.

3. **Use 5 fold cross validation to select the optimal regularization parameter λ. Plot the cross validation curve and report the optimal λ.**



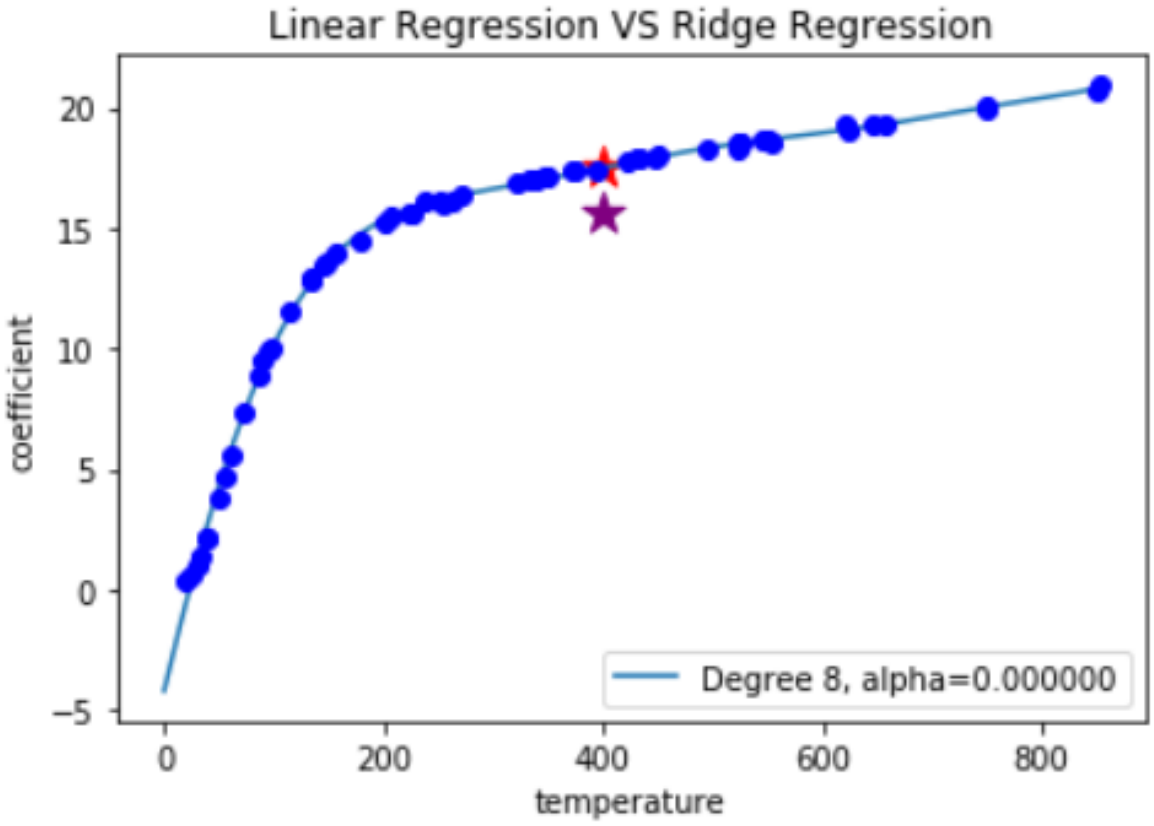
Optimal alpha = 1e-10

Since the MSE of the model with degree 8 is the lowest, I chose 8 as an optimal order, and the optimal α = 1e-10 After checking the 5-fold cross validation, we found that the optimal α is 1e-10, at the order 8.

4. **Predict the coefficient at 400 degree Kelvin using both models. Comment on how would you compare the accuracy of predictions.**

```
Closest check point(less than 400):  393.32 actual with the check point(less than 40
0):  17.339
Error with Linear Regression:  12.006125930047062
Closest check point(more than 400):  422.02 actual with the check point(more than 40
0):  17.765
Error with Linear Regression:  10.574254526183191
Closest check point(less than 400):  393.32 actual with the check point(less than 40
0):  17.339
Error with Poly + Ridge Regression:  0.6284766013925802 %
Closest check point(more than 400):  422.02 actual with the check point(more than 40
0):  17.765
Error with Poly + Ridge Regression:  0.2856874187884384 %
```

With the temperature(400K), Linear Regression predicted coefficient is 15.4037117878581 96
Polynomial Regression with ridge regression predicted coefficient is 17.510210595344432



As we can see through the graph, we can see that Polynomial Regression method has better performance than Linear Regression

Problem 4

1. **Find the closed form solution for $\hat{\beta}(\lambda)$ and its distribution.**

   $\hat{\beta}(\lambda) = argmin_\beta[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_2$
   To Find the optimal $\hat{\beta}(\lambda)$, we need to differentiate the function w.r.t. β
   $\frac{\partial\hat{\beta}(\lambda)}{\partial\beta} = argmin_\beta[-2X^T(y - \beta^T X) + 2\lambda\beta]$
   by setting this equation as 0, we can get
   $\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$
   To find the distribution, since ϵ is following Gaussian, y is also Gaussian. and $\hat{\beta}$ is linear transformation in y, so $\hat{\beta}$ has a Gaussian distribution as well.
   And Mean and Variance of $\hat{\beta}$ is following:
   $E[\hat{\beta}] = (X^TX + \lambda I)^{-1}X^TX\beta^*$
   $Var[\hat{\beta}] = \sigma^2(X^TX + \lambda I)^{-1}X^TX[(X^TX + \lambda I)^{-1}]^T$
   Thus, the distribution of the Ridge Regression estimator is:
   $\hat{\beta}\ N((X^TX + \lambda I)^{-1}X^TX\beta^*, \sigma^2(X^TX + \lambda I)^{-1}X^TX[(X^TX + \lambda I)^{-1}]^T)$

2. **Calculate the bias $E[x^T\hat{\beta}]$as a function of λ, and some fixed test point x.**

   $Bias = E[x^T\hat{\beta}] - x^T\beta^*$, which is
   $Bias = E[x^T((X^TX + \lambda I)^{-1}X^Ty] - x^T\beta^*$
   and $y = X^T\beta^*$, so
   $Bias = x^T((X^TX + \lambda I)^{-1}X^TX\beta^* - x^T\beta^*$
   $Bias = x^T[I - \lambda(X^TX + \lambda I)^{-1}]\beta^* - x^T\beta^*$
   Therefore,
   $Bias = -\lambda x^T((X^TX + \lambda I)^{-1})\beta^*$

3. **Calculate the variance term.**

   Let $w = (X^TX + \lambda I)^{-1}X^TX$. Then,
   $Var(x^T\hat{\beta}) = x^Tw\sigma^2(X^TX)w^Tx$
   After arrangement, we got
   $Var(x^T\hat{\beta}) = \sigma^2x^T(X^TX + \lambda I)^{-1}X^TX[(X^TX + \lambda I)^{-1}]^Tx$

4. **Use the results from parts (b) and (c) and the bias-variance decomposition to analyze the impact of λ in the squared error. Specifically, which term dominates when λ is small, and large, respectively?**

   $SquaredError = Bias^2 + Variance + noise$, which is
   $SquaredError = [-\lambda x^T((X^TX + \lambda I)^{-1})\beta^*]^2 + \sigma^2x^T(X^TX + \lambda I)^{-1}X^TX[(X^TX + \lambda I)^{-1}]^Tx$
   When λ is large, Bias is the dominant term, while if λ is small, Variance term is dominant.

Rubric

Problem 1

(a) 10 points. Reasonable numerical and graphical explanation can get full points.
(b) If the final result is not correct but partial of the work is correct **(-5)**.
(c) If the answer is not correct, but partial explanation is reasonable **(-2)**.
(d) Same as **(c)**.

Problem 2

Please be gentle when grading this problem. As students need to separate the dataset randomly, the results can be very different.

1**. [**5 **Points]** Correct Answer, 1 for each.

2**. [**10 **Points]** Try the problem.(5) Code can be run(3), result is not bad, accuracy is larger than 80% (2).

3**. [**10 **Points]** Try the problem.(5) Code can be run(3), result is not bad, accuracy is larger than 80% (2).

4**. [**5 **Points]** AUC curve looks similar to the sample. The x-axis have multiple correct answers such as max depth, max number of leaf nodes, total number of nodes.(5)

Problem 3

Even though the answer is not correct, if their logic is correct, give 5 points for each questions.
(1) 10 points. *Reasonable explanation*
Fitted model equation, and reasonable number of MSE is good to get the full credit.
(2) 10 points. *Reasonable explanation*
If their choice of degree is not reasonable (ex: n = 9 , 10, 1, 2). **(-5)**
Wrong formulation of ridge regression. **(-5)**
(3) 10 points. *Reasonable explanation*
Bad cross validation curve **(-5)**.
Wrong optimal λ **(-5)**
(4) 10 points. *Reasonable explanation*
Wrong prediction of coefficient for models **(-5)**
If they have correct logic to compare the accuracy of predictions give them **(+5)**.

Problem 4

If they try any questions in problem 4, give them 1 point for each (total 4). If their answers are correct, give them full credit (2.5 point for each)

(1) 2.5 points (2) 2.5 points (3) 2.5 points (4) 2.5 points