Date...../...../.....

Page.....

Archit Matta
am5500

Homework 1
5/2/21

(1) $$p(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

(a) Joint likelihood of $(x_1, x_2 \cdots, x_N)$ would be equal to the product of the individual probabilities for x_is.

$$\Rightarrow \left(\frac{\lambda^{x_1} e^{-\lambda}}{x_1!}\right) \cdot \left(\frac{\lambda^{x_2} e^{-\lambda}}{x_2!}\right) \cdots \left(\frac{\lambda^{x_N} e^{-\lambda}}{x_N!}\right)$$

$$\Rightarrow \boxed{\frac{\lambda^{x_1+x_2+\cdots+x_N}}{x_1! \cdot x_2! \cdots x_N!} \cdot e^{-N\lambda}}$$

(b) Maximum Likelihood Estimate for λ (λ_{ML})

$$\lambda_{ML} = \arg\max_{\lambda} \text{(Joint Likelihood)}$$

We can take log of the function & maximise it.

$$\lambda_{ML} = \arg\max_{\lambda} \left(\ln\left[\frac{\lambda^{x_1+x_2+\cdots x_N}}{x_1! x_2! \cdots x_N!} \cdot e^{-N\lambda}\right]\right)$$

Setting derivative w.r.t. λ as 0

$$\frac{d}{d\lambda}\left[(x_1+x_2+\cdots x_n)\ln\lambda + \ln\left(\frac{1}{x_1! x_2! \cdots x_n!}\right) - N\lambda\right] = 0$$

$$\Rightarrow \frac{(x_1+x_2+\cdots x_n)}{\lambda} - N = 0$$

$$\Rightarrow \boxed{\lambda_{ML} = \frac{x_1+x_2+\cdots+x_N}{N}}$$

Date...../...../.....

Page.....

(c) $p(\lambda) = \text{gamma}(a, b)$

$$p(\lambda) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$$

$$\lambda_{MAP} = ?$$

$$\lambda_{MAP} = \arg\max_{\lambda} (p[\lambda/X])$$

Taking and maximising log since it is monotonic

$$\lambda_{MAP} = \arg\max_{\lambda} (\ln(p[\lambda/X]))$$

$$p[\lambda/X] = \frac{p[X/\lambda] \cdot p[\lambda]}{\text{Constant}} \quad (\text{Baye's Rule})$$

Setting the derivative equal to 0

$$\Rightarrow \frac{d}{d\lambda}\left[\ln(p[X/\lambda]) + \ln(p[\lambda]) - \ln(\text{constant})\right] = 0$$

$$\Rightarrow \frac{d}{d\lambda}\left[\ln\left(\frac{\lambda^{(X_1+X_2+\ldots X_N)} e^{-N\lambda}}{X_1! X_2! \ldots X_N!}\right) + \ln\left[\frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}\right] - \ln(\text{constant})\right] = 0$$

$$\Rightarrow \frac{d}{d\lambda}\left[X \ln\lambda - N\lambda + (a-1)\ln\lambda - b\lambda\right] = 0$$

(Ignoring Terms without λ)

$$\Rightarrow \frac{X + a - 1}{\lambda} - (b + N) = 0$$

$$\Rightarrow \boxed{\lambda_{MAP} = \frac{X + a - 1}{b + N}} \quad \text{where } X = x_1 + x_2 + \ldots$$

Date...../...../.....

Page.....

(d) $$p(\lambda/x) = \frac{p(x/\lambda) \cdot p(\lambda)}{\int_{R^N} p(x/\lambda) \cdot p(\lambda)\, d\lambda}$$

$$\Rightarrow p(\lambda/x) \propto p(x/\lambda) \cdot p(\lambda)$$

$$\Rightarrow p(\lambda/x) \propto \left(\frac{\lambda^{x_1+x_2+\cdots x_N}}{x_1!\, x_2! \cdots x_N!} e^{-N\lambda}\right) \cdot \left(\frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}\right)$$

$$\Rightarrow p(\lambda/x) \propto \frac{b^a}{(x_1!\, x_2! \cdots x_N!)\Gamma(a)} \cdot \lambda^{(x_1+x_2+\cdots+x_N)+(a-1)} \cdot e^{-(N+b)\lambda}$$

We can see this looks like a Gamma Distribution
The gamma distribution can be obtained by multiplying constants.
The parameters of the distribution would be -

$$\alpha = x_1 + x_2 + \cdots + x_N + a$$
$$\beta = N + b$$

Thus, the posterior distribution of λ is gamma (α, β).

(e) Mean of Gamma $= \alpha/\beta$, Variance of Gamma $= \alpha/\beta^2$

$$\text{Mean} = \frac{x_1 + x_2 + \cdots x_N + a}{(N+b)}$$

$$\text{Variance} = \frac{x_1 + x_2 + \cdots + x_N + a}{(N+b)^2}$$

We can see the following —

(2) Mean $= \lambda_{MAP} + \frac{1}{b+N}$

(3) Variance $= \frac{(\lambda_{MAP} + \frac{1}{b+N})}{b+N}$

(2) Mean $= ((\lambda_{ML} \times N) + a) / N+b$

(4) Variance $= \frac{(\lambda_{ML} \times N) + a}{(b+N)^2}$

Date...../...../.....

Page.....

② (a) $w_{RR} = (\lambda I + X^T X)^{-1} X^T y$

$E[w_{RR}] = E[(\lambda I + X^T X)^{-1} X^T y]$

$E[w_{RR}] = (\lambda I + X^T X)^{-1} X^T E[y]$

$E[y] = Xw$ since $y \sim N(Xw, \sigma^2 I)$

$\Rightarrow$ $\boxed{E[w_{RR}] = (\lambda I + X^T X)^{-1} X^T X w}$

$Var[w_{RR}] = E[w_{RR} w_{RR}^T] - E[w_{RR}] E[w_{RR}]^T$

let $(\lambda I + X^T X)^{-1} = A$

$\Rightarrow Var[w_{RR}] = E[A X^T y y^T X A^T] - A X^T X w w^T X^T A^T$

$= A X^T (\sigma^2 I + X w w^T X^T) X A^T - A X^T X w w^T X^T X A^T$

$= A X^T \sigma^2 I X A^T + A X^T X w w^T X^T X A^T - A X^T X w w^T X^T X A^T$

$Var[w_{RR}] = \sigma^2 (\lambda I + X^T X)^{-1} X^T X ((\lambda I + X^T X)^{-1})^T$

Right Multiply Both sides by $(\lambda I + X^T X)^T (X^T X)^{-1}$

↳ This is equal to $(\lambda (X^T X)^{-1} + I)^T$

$Var[w_{RR}] (\lambda (X^T X)^{-1} + I)^T = \sigma^2 (\lambda I + X^T X)^{-1}$

Right Multiply by $((\lambda (X^T X)^{-1} + I)^{-1})^T$

$Var[w_{RR}] = \sigma^2 (\lambda I + X^T X)^{-1} ((I + \lambda (X^T X)^{-1})^{-1})^T$

Left Multiply Both sides by $(\lambda I + X^T X)(X^T X)^{-1}$

$(\lambda (X^T X)^{-1} + I) Var[w_{RR}] = \sigma^2 (X^T X)^{-1} ([I + \lambda (X^T X)^{-1}]^{-1})$

Date ... / ... / ...

Page ...

Left Multiplying both sides by $(\lambda (X^T X)^{-1} + I)^{-1}$

$$\text{Var}[w_{RR}] = \sigma^2 \left[\lambda (X^T X)^{-1} + I\right]^{-1} (X^T X)^{-1} \left[\left[\lambda (X^T X)^{-1} + I\right]^{-1}\right]^T$$

Let $Z = (I + \lambda (X^T X)^{-1})^{-1}$

$$\boxed{\text{Var}[w_{RR}] = \sigma^2 Z (X^T X)^{-1} Z^T}$$

(b)

$$w_{RR} = (\lambda I + X^T X)^{-1} X^T y$$
$$w_{LS} = (X^T X)^{-1} X^T y$$

$$X = USV^T$$

$$w_{RR} = (\lambda I + X^T X)^{-1} (X^T X) \overbrace{(X^T X)^{-1} X^T y}^{w_{LS}}$$
$$w_{RR} = \left[(X^T X)\left[\lambda (X^T X)^{-1} + I\right]\right]^{-1} X^T X \, w_{LS}$$
$$w_{RR} = \left[\lambda (X^T X)^{-1} + I\right]^{-1} (X^T X)^{-1} (X^T X) \, w_{LS}$$
$$w_{RR} = \left[\lambda (X^T X)^{-1} + I\right]^{-1} w_{LS}$$

Using $X = USV^T$, we get $(X^T X)^{-1} = V S^{-2} V^T$

$$w_{RR} = \left(\lambda (VS^{-2}V^T) + I\right)^{-1} w_{LS}$$
$$w_{RR} = V (\lambda S^{-2} + I)^{-1} V^T w_{LS}$$
$$\boxed{w_{RR} = V M V^T w_{LS}}$$

where M is a diagonal matrix with $M_{ii} = \frac{S_{ii}^2}{\lambda + S_{ii}^2}$

$$\boxed{w_{RR} = V (\lambda S^{-2} + I)^{-1} V^T w_{LS}}$$

Problem 3

a) Here's a brief look at the obtained Wrr values.

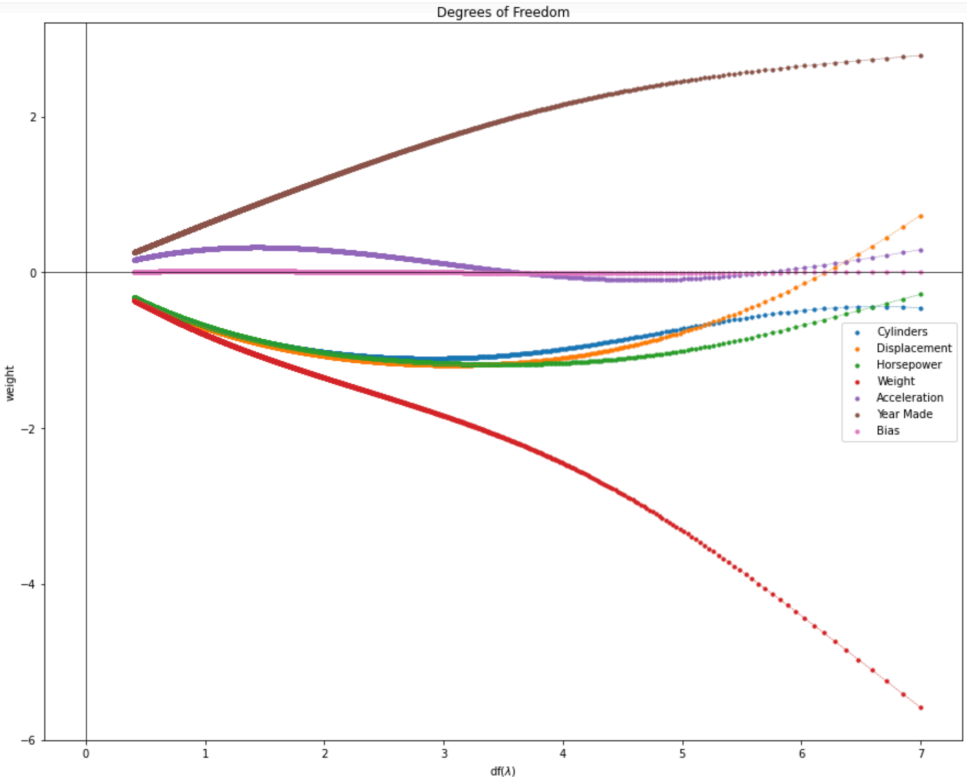
In [5]:
```
#Printing the W for Ridge Regression for all Lambdas
pd.DataFrame(np.asarray(wrr).reshape(5001,7))
```

Out[5]:

	0	1	2	3	4	5	6
0	-0.456268	0.730185	-0.284646	-5.585580	0.289570	2.781397	0.010157
1	-0.445729	0.577783	-0.344522	-5.409677	0.251099	2.763334	0.008127
2	-0.441314	0.445755	-0.399202	-5.250282	0.216899	2.746404	0.006362
3	-0.441431	0.330230	-0.449221	-5.104973	0.186365	2.730448	0.004816
4	-0.444922	0.228265	-0.495063	-4.971812	0.159004	2.715338	0.003450
5	-0.450930	0.137581	-0.537159	-4.849218	0.134409	2.700969	0.002236
6	-0.458814	0.056385	-0.575889	-4.735887	0.112239	2.687253	0.001152
7	-0.468090	-0.016752	-0.611593	-4.630732	0.092206	2.674118	0.000179
8	-0.478388	-0.082984	-0.644570	-4.532839	0.074067	2.661502	-0.000698
9	-0.489422	-0.143252	-0.675083	-4.441430	0.057613	2.649351	-0.001492
10	-0.500972	-0.198333	-0.703368	-4.355839	0.042665	2.637622	-0.002212

Here's the graph between Degrees of Freedom and Weight

In [6]:
```
#Plotting Degrees of Freedom vs Weight
wrr_np = np.asarray(wrr)
svd_np = np.asarray(svd)
plt.figure()
dimensions = ["Cylinders", "Displacement", "Horsepower", "Weight", "Acceleration", "Year Made", "Bias"]
for i in range(wrr_np[0].shape[0]):
    plt.plot(svd_np, wrr_np[:,i], linestyle = 'dashed', linewidth = 0.5)
    plt.scatter(svd_np, wrr_np[:,i], s = 10, label = dimensions[i])
plt.axhline(y = 0, color = 'black', linewidth = 0.7)
plt.axvline(x = 0, color = 'black', linewidth = 0.7)
plt.xlabel('df($\lambda$)')
plt.ylabel('weight')
plt.title('Degrees of Freedom')
plt.legend()
plt.show()
```



b)

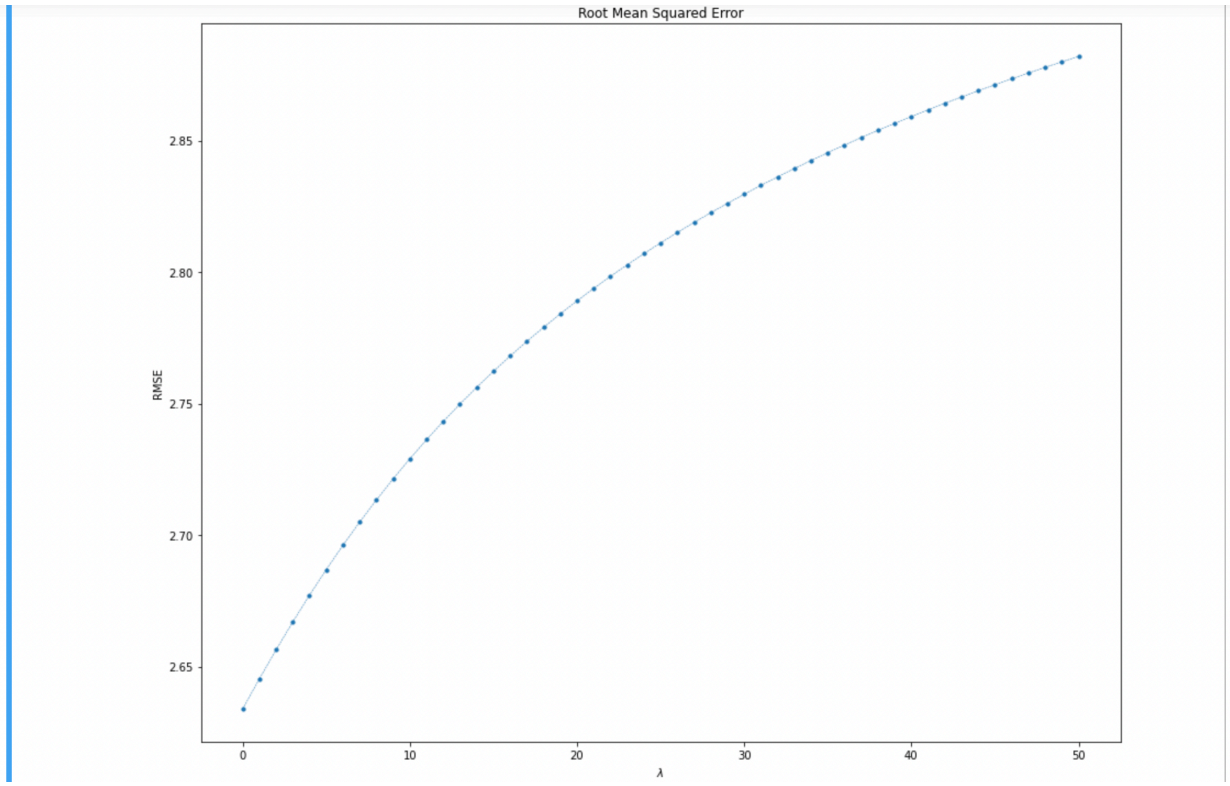
Here are the 2 dimensions that stand out

The 2 dimensions that stand out are 'Year Made' and 'Weight'. 'Year Made' has a high positive weight and thus indicates that newer cars have higher Miles per Gallon. 'Weight' has a high negative weight and thus indicates that lighter cars have higher Miles per Gallon.

c)

Here's the graph between RMSE and Lambda

In [9]:
```
#Plotting RMSE vs Lambda
plt.plot(lamda[:51], rmse, linestyle = 'dashed', linewidth = 0.5)
plt.scatter(lamda[:51], rmse, s = 10)
plt.xlabel('$\lambda$')
plt.ylabel('RMSE')
plt.title('Root Mean Squared Error')
plt.show()
```



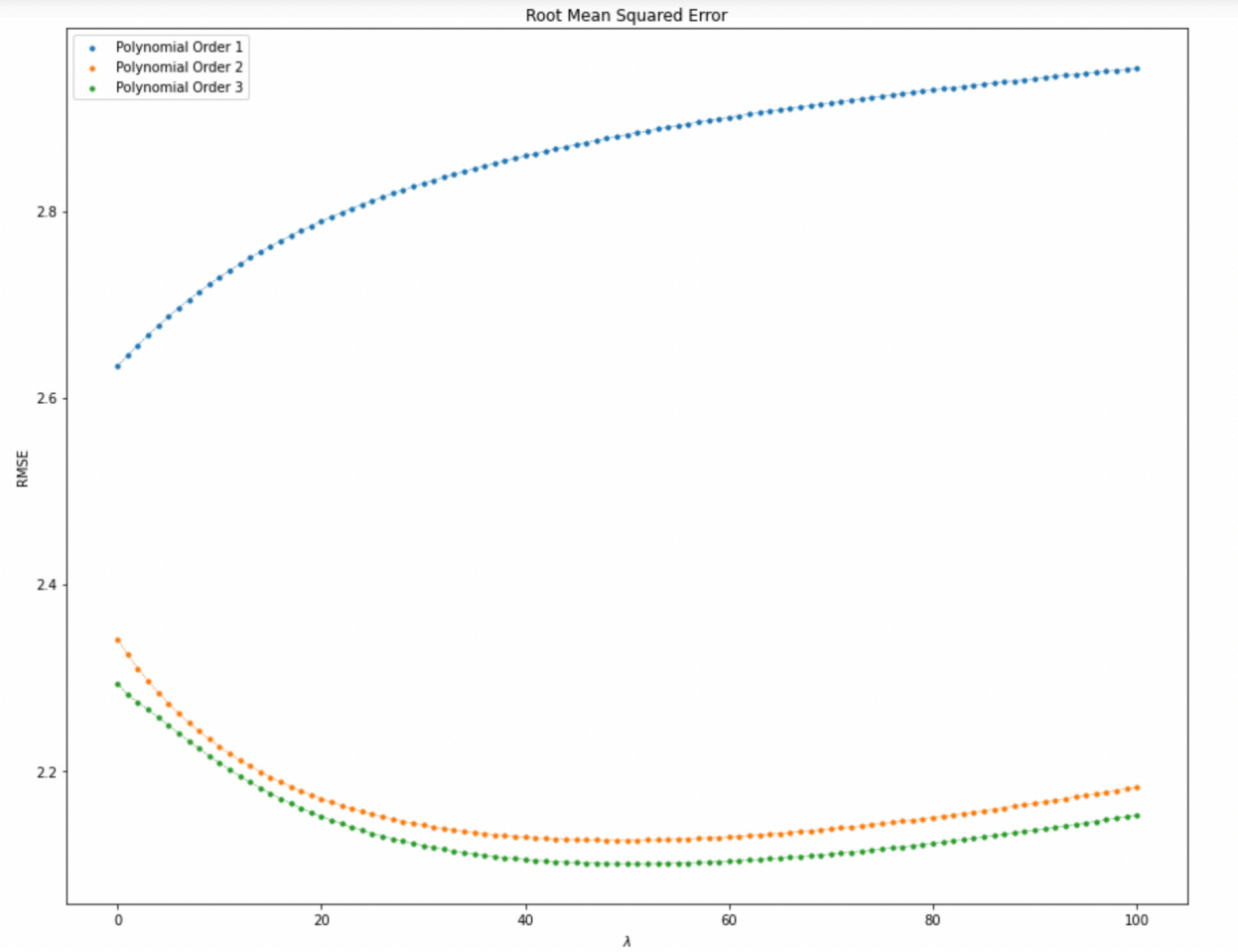
Here's what we observe by looking at this figure

Based on this figure, we can see that the optimal Lambda for this problem having the lowest RMSE is 0. We also know that for Lambda = 0, W for Ridge Regression is the same as W for Least Squares. Hence, based on this graph we should choose Least Squares for this problem.

d)

Here's the graph between RMSE and Lambda

```
#Plotting RMSE vs Lambda for different Polynomial Orders
plt.plot(lamda[:101], rmse_1, linestyle = 'dashed', linewidth = 0.5)
plt.scatter(lamda[:101], rmse_1, s = 10, label = 'Polynomial Order 1')
plt.plot(lamda[:101], rmse_2, linestyle = 'dashed', linewidth = 0.5)
plt.scatter(lamda[:101], rmse_2, s = 10, label = 'Polynomial Order 2')
plt.plot(lamda[:101], rmse_3, linestyle = 'dashed', linewidth = 0.5)
plt.scatter(lamda[:101], rmse_3, s = 10, label = 'Polynomial Order 3')
plt.xlabel('$\lambda$')
plt.ylabel('RMSE')
plt.title('Root Mean Squared Error')
plt.legend()
plt.show()
```



Here's what we observe by looking at this figure

Based on this plot, we should choose a Polynomial of Order 3 (p = 3) as it has the least RMSE overall. Our assessment of the ideal value of Lambda also changes to 51 from our initial value of 0 as the lowest RMSE value for p = 2 occurs at this point.