Name : PRASHAM D. SHETH

UNI : pds2136

Problem - 1

(a) $X_1 \dots X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$

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

$$L(\lambda; X_1 \dots X_n) = p(X_1 \dots X_n | \lambda)$$

$$= \prod_{i=1}^{N} P(X_i|\lambda)$$

$$= \boxed{\prod_{i=1}^{N} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}}$$

(b) $$\hat{\lambda}_{ML} = \underset{\lambda}{\text{argmax}} \prod_{i=1}^{N} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}$$

$$= \underset{\lambda}{\text{argmax}} \log\left(\prod_{i=1}^{N} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}\right)$$

$$= \underset{\lambda}{\text{argmax}} \left[\log \lambda \cdot \sum_{i=1}^{N} x_i - N\cancel{\log}\lambda - \sum_{i=1}^{N} \log(x_i!)\right]$$

$$= \underset{\lambda}{\text{argmax}} \left[\underbrace{(\log\lambda)\left(\sum_{i=1}^{N} x_i\right) - N\cancel{\log}\lambda}_{L_1}\right]$$

$$L_1 = (\log\lambda)\cdot\left(\sum_{i=1}^{N} x_i\right) - N\cancel{\log}\lambda$$

$$\frac{\partial L_1}{\partial \lambda} = \frac{\sum_{i=1}^{N} x_i}{\lambda} - N$$

$$\frac{\partial L_1}{\partial \lambda} = 0$$

$$\therefore \frac{\sum_{i=1}^{N} x_i}{\lambda} = N$$

$$\therefore \boxed{\hat{\lambda}_{ML} = \frac{\sum_{i=1}^{N} x_i}{N}}$$

(c) Prior distribution of Gamma is assumed for λ.

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

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} \log(p(\lambda | x_1 \dots x_N))$$

$$= \arg\max_{\lambda} \left[\log(p(x_1 \dots x_N | \lambda)) + \log p(\lambda)\right]$$

$$= \arg\max_{\lambda} \left[\log\left(\prod_{i=1}^{N} p(x_i|\lambda)\right) + \log p(\lambda)\right]$$

$$= \arg\max_{\lambda} \left[\sum_{i=1}^{N} \log p(x_i|\lambda) + \log p(\lambda)\right]$$

$$= \arg\max_{\lambda} \left[\sum_{i=1}^{N} \left(\frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}\right) + \log\left(\frac{b^a \cdot \lambda^{a-1} \cdot e^{-b\lambda}}{\Gamma(a)}\right)\right]$$

$$= \arg\max_{\lambda} \left[\log\lambda \cdot \left(\sum_{i=1}^{N} x_i\right) - N\cdot\lambda + (\log\lambda)(a-1) - b\lambda\right]$$

$$\therefore \hat{\lambda}_{map} = \underset{\lambda}{\arg\max} \left[\log \lambda \left[\sum_{i=1}^{N} x_i + a - 1 \right] - \lambda (N+b) \right]$$

$$L_1 = \log \lambda \left(\sum_{i=1}^{N} x_i + a - 1 \right) - \lambda (N+b)$$

$$\frac{\partial L_1}{\partial \lambda} = 0$$

$$\therefore \frac{\sum_{i=1}^{N} x_i + a - 1}{\lambda} - (N+b) = 0$$

$$\therefore \boxed{\hat{\lambda}_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{N+b}}$$

(d) for posterior distribution of λ, we use Bayes' rule.

$$P(\lambda | X_1 \dots X_N) = \frac{P(X_1 \dots X_N | \lambda) \cdot P(\lambda)}{P(X_1 \dots X_n)}$$

Here, the denominator normalizes the numerator & doesn't depend on λ.

$$\therefore P(\lambda | X_1 \dots X_n) \propto P(X_1 \dots X_n | \lambda) \cdot P(\lambda)$$

$$\propto \frac{\lambda^{\sum_{i=1}^{N} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{N} (x_i!)} \cdot \frac{b^a \cdot \lambda^{a-1} \cdot e^{-b\lambda}}{\Gamma(a)}$$

$$\propto \lambda^{\left(\sum_{i=1}^{N} x_i + a - 1\right)} \cdot e^{-\lambda(n+b)}$$

From the above form we can identify that posterior distribution of λ must be a Gamma with parameters $\left(\sum_{i=1}^{N} x_i + a, N+b\right)$

$$\therefore \quad P(\lambda \mid X_1 \dots X_N) = \text{Gamma}\left(\sum_{i=1}^{N} x_i + a, N+b\right)$$

(e) We know, $E[X] = \frac{a}{b}$ and $Var[X] = \frac{a}{b^2}$

when $X \sim \text{Gamma}(a, b)$

Now, as

$$P(\lambda \mid X_1 \dots X_n) \sim \text{Gamma}\left(\sum_{i=1}^{N} x_i + a, N+b\right)$$

Mean value of λ = $E[\lambda] = \dfrac{\sum_{i=1}^{N} x_i + a}{n+b}$

Variance of λ = $Var[\lambda] = \dfrac{\sum_{i=1}^{N} x_i + a}{(N+b)^2}$

→ This clearly shows that the mean value of λ under the posterior is greater than $\hat{\lambda}_{MAP}$.

→ $\hat{\lambda}_{ML}$ and this mean value of λ under the posterior would be equal when a and b would be zero.

Problem - 2 :-

→ $y_i \overset{iid}{\sim} N(x_i^T \cdot w, \sigma^2)$

Using data we already have approximated W_{RR} as

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

→ We also know that

$$W_{LS} = (X^T X)^{-1} X^T y$$

$$E[W_{LS}] = w$$

$$Var[W_{LS}] = \sigma^2 (X^T X)^{-1}$$

→ As, $W_{RR} = (\lambda I + X^T X)^{-1} X^T y$

we can write it as

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

$$= (\lambda I + X^T X)^{-1} (X^T X) \cdot W_{LS}$$

From this we can calculate the expected value for W_{RR}

$$\therefore E[W_{RR}] = (\lambda I + X^T X)^{-1} (X^T X) \cdot E[W_{LS}]$$

$$= (\lambda I + X^T X)^{-1} (X^T X) \cdot W$$

Now, for variance of w_{RR},

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

$$= \left((X^TX)(\lambda \cdot (X^TX)^{-1} + I)\right)^{-1} (X^TX)\, w_{LS}$$

$$= (\lambda (X^TX)^{-1} + I)^{-1} \cdot (X^TX)^{-1} \cdot (X^TX) \cdot w_{LS}$$

$$= (\lambda (X^TX)^{-1} + I)^{-1} \cdot w_{LS}$$

lets say $(\lambda (X^TX)^{-1} + I)^{-1} = Z$

$$\therefore \quad w_{RR} = Z \cdot w_{LS}$$

$$Var[w_{RR}] = Z \cdot Var[w_{LS}] \cdot Z^T$$

$$= Z \cdot \sigma^2 (X^TX)^{-1} \cdot Z^T$$

$$= \sigma^2 \cdot Z (X^TX)^{-1} \cdot Z^T$$

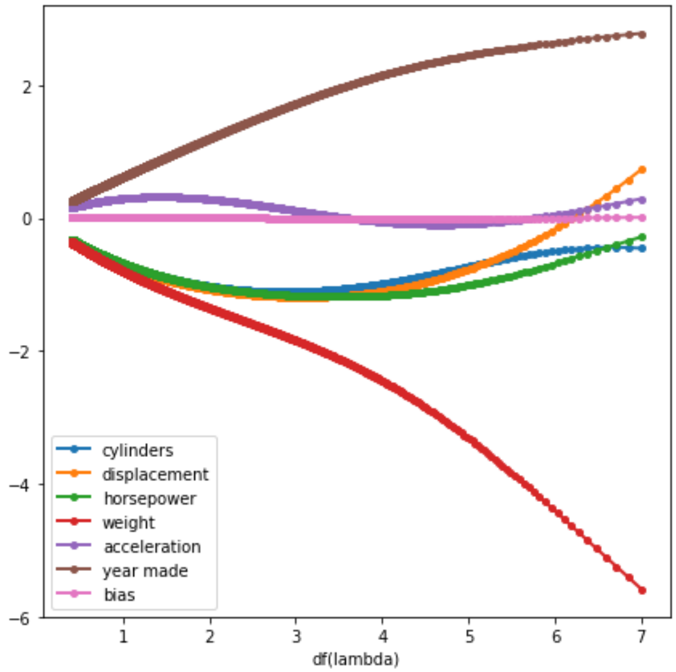
$$\therefore \quad E[w_{RR}] = (\lambda I + X^TX)^{-1} X^TX\, w$$

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

where $Z = (I + \lambda (X^TX)^{-1})^{-1}$

Problem – 3:

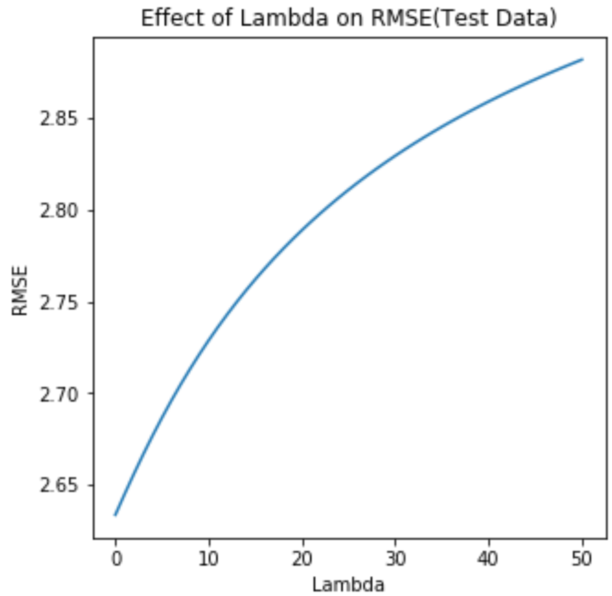
(a) *For _ = 0; 1; 2; 3; …; 5000, solve for WRR. (Notice that when lambda = 0, WRR = WLS.) In one figure, plot the 7 values in WRR as a function of df(lambda). You will need to call a built in SVD function to do this as discussed in the slides. Be sure to label your 7 curves by their dimension in X.*



(b) *Two dimensions clearly stand out over the others. Which ones are they and what information can we get from this?*

- We can clearly see that "weight" and "year made" clearly stand out over others. The dominance of these 2 show the importance of both the values in determining the dependent variable (i.e.) the miles per gallon for that car. If we reduce the degrees of freedom for the model, by increasing the value of lambda, at a certain point when df(lambda) is nearly 4, we see that the coefficient for acceleration becomes almost zero. The X-axis of the above plot shows the Degree of freedom which essentially is the function of lambda.
- Another key thing is that the increase in value of "year made" increases the miles per gallon which follows the intuitive way of understanding that the newer cars would have higher mileage than the older ones where as with increase in "weight" of the car, the mileage always decreases. These 2 things remain the same at any value of lambda (as the curves for both the features don't cross the line for 0 at any point in the plot)

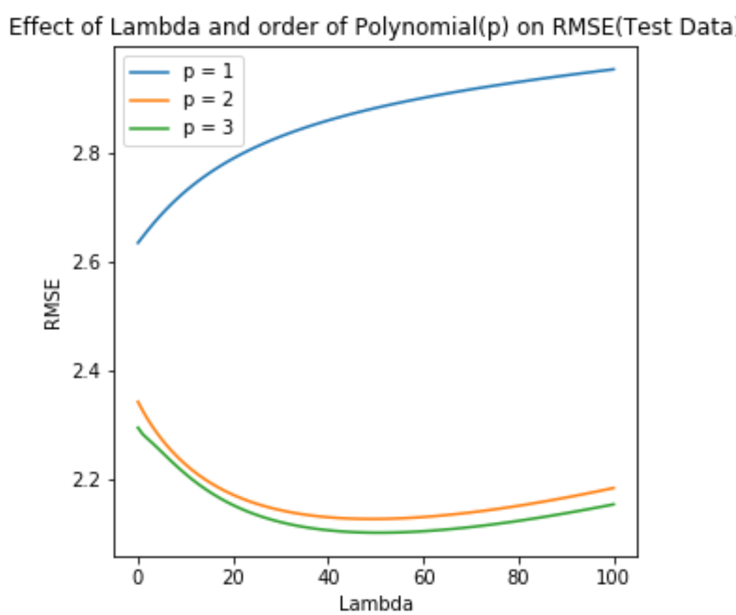
(c) For lambda = 0; …; 50, predict all 42 test cases. Plot the root mean squared error (RMSE) on the test set as a function of lambda —not as a function of df(lambda). What does this figure tell you when choosing for this problem (and when choosing between ridge regression and least squares)?



- From this plot we can see the RMSE value increases with increase in Lambda in this case. So, the RMSE value for lambda = 1 is less than RMSE value for lambda = (i+1).
- Also, from the above statement in this case, we can say that the RMSE for lambda = 0 would be the minimum.
- Thus, in this case we would prefer least square (same as ridge regression with lambda = 0) over the ridge regression.

(d) In one figure, plot the test RMSE as a function of lambda = 0; …; 100 for p = 1; 2; 3. Based on this plot, which value of p should you choose and why? How does your assessment of the ideal value of lambda change for this problem?

- Below figure shows the plot for RMSE as a function of lambda when p = 1, 2 3.



- From the plot we can see that
    - RMSE values for p = 1 increases with increase in value of lambda
    - RMSE values for p = 2, 3 decrease with increase in lambda (initially) while they after a certain lambda starts on increasing with increase in lambda.
- The RMSE values for p = 3 are the minimum and hence we can say that the model with p = 3 perform the best on the test set and hence, we will select the value of p as 3.
- In the case of p = 1, as the minimum RMSE was obtained at the point where lambda = 0, we would have selected that value of lambda had we selected p to be 1. For p = 2 the minimum is no longer corresponding to the value of lambda = 0. Now the minimum value of RMSE can be approximately seen to be obtained at lambda = 42. Similar is the case for p = 3. Hence, with introduction of p^{th} order polynomial terms we would require to have need for regularization and hence the ideal value for lambda no longer remains 0.