Michelle Chen (mc4571)
Machine learning HW1

Problem 1

(a) What is the joint likelihood of the data $(x_1, \ldots, x_n)$?

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

Joint Likelihood $\Rightarrow$ $L(\lambda; x) = \prod_{i=1}^{n} p(X_i \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$

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

$\uparrow$ $x_1! \cdot x_2! \cdot x_3! \ldots x_n!$

(b) MLE (λ_{ML}) for λ

$$\log(L(\lambda; x)) = \sum_{i=1}^{n} x_i \log\lambda - n\lambda$$

$$\frac{d}{d\lambda}\left[\sum_{i=1}^{n} x_i \log\lambda - n\lambda\right] = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0 \Rightarrow \frac{\sum_{i=1}^{n} x_i}{\lambda} = n$$

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

(c) Derive the maximum a posteriori (MAP) estimate λ_{map} for λ

$$\underbrace{P(\theta \mid X)}_{\text{posterior}} \propto \underbrace{P(X \mid \theta)}_{\text{likelihood}} \cdot \underbrace{P(\theta)}_{\text{prior}}$$

* Assume that prior for λ is $p(\lambda) = \frac{B^a}{\Gamma(a)} \lambda^{a-1} e^{-B\lambda}$ which is a gamma(a, B) prior

* Likelihood $= \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$ (from 1.a)

$$\lambda_{MAP} = \arg\max_{\lambda} \ln(p(\lambda \mid X))$$

$$= \arg\max_{\lambda} \ln\left[\frac{p(X \mid \lambda) \cdot p(\lambda)}{p(X)}\right] = \arg\max_{\lambda} \ln[p(X \mid \lambda)] + \ln[p(\lambda)] - \cancel{\ln[p(X)]}^{0}$$

$$= \arg\max_{\lambda} \sum_{i=1}^{n} x_i \log\lambda - n\lambda + \ln[\lambda^{a-1} e^{-B\lambda}]$$

$$\text{"} \quad \sum_{i=1}^{n} x_i \log\lambda - n\lambda + \ln[\lambda^{a-1}] + \ln[e^{-B\lambda}]$$

$$\text{"} \quad \sum_{i=1}^{n} x_i \log\lambda - n\lambda + (a-1)\ln\lambda - B\lambda$$

$$\frac{d}{d\lambda}\left[-(n+B)\lambda + \sum_{i=1}^{n} x_i \log\lambda + (a-1)(\ln\lambda)\right] = 0$$

$$-(n+B) + \frac{\sum_{i=1}^{n} x_i}{\lambda} + \frac{a-1}{\lambda} = 0$$

$$n+B = \frac{\sum_{i=1}^{n} x_i + a - 1}{\lambda} \rightarrow \boxed{\lambda_{MAP} = \frac{\sum_{i=1}^{n} x_i + a - 1}{n+B}}$$

d) Derive the posterior distribution of λ and identify this distribution

$$p(\lambda|X) \propto p(X|\lambda) \cdot p(\lambda)$$

posterior — Likelihood (joint) — prior $\rightarrow$ gamma

$$" \propto \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\prod_{i=1}^{n} x_i!} \cdot \frac{B^a}{\Gamma(a)} \lambda^{a-1} e^{-B\lambda}$$

$$" \propto \frac{e^{-(n+B)\lambda}\lambda^{\sum_{i}^{n} x_i + a-1}}{\prod_{i=1}^{n} x_i!}\left(\frac{B^a}{\Gamma(a)}\right)$$

$\ll$ but since posterior is a proportion of the product, we can get rid of constant coefficents $\gg$

$$p(\lambda|X) \propto e^{-(n+B)\lambda}\lambda^{\sum_{i=1}^{n} x_i + a-1}$$

$\therefore$ The distribution is $\text{Gamma}\Big(\underbrace{\sum_{i=1} x_i + a}_{\text{"}a\text{"}}, \underbrace{n+B}_{\text{"}B\text{"}}\Big)$

e) What is the mean and variance of λ under posterior distribution? Discuss how it relates to λ_{ML} and λ_{MAP}.

Gamma distribution $\left(\sum_{i=1}^{n} x_i + a, \; n+B\right)$

Mean $= E[\lambda] = \dfrac{\sum_{i=1}^{n} x_i + a}{n+B}$

variance $= var[\lambda] = \dfrac{\sum_{i=1}^{n} x_i + a}{(n+B)^2}$

The mean and variance of λ can help us to describe a distribution in which our point estimates λ_{MLE} and λ_{map} fall within.

$\lambda_{map} = \dfrac{\sum_{i=1}^{n} x_i + a - 1}{n+B}$ is the <u>mode</u> of the posterior distribution (Gamma),

where the mode takes on the form of $E[\underset{\substack{\uparrow \\ \text{random variable}}}{\lambda}] - \frac{1}{B}$.

Whereas λ_{MLE} is the value of the parameter that <u>maximizes the entire distribution</u>.

Problem 2

Given $y_i \overset{iid}{\sim} N(x_i^T w, \sigma^2)$ and $w_{RR} = (\lambda I + X^T X)^{-1} X^T y$

Calculate $E[w_{RR}]$, $var[w_{RR}]$

we know $E[w_{LS}]$ $\downarrow$ $(X^TX)^{-1}X^Ty$

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

$$= E[(\lambda I + X^TX)^{-1} \underbrace{(X^TX)(X^TX)^{-1}}_{=I} X^T y]$$

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

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

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

$$= (\lambda(X^TX)^{-1} + I)^{-1} (X^TX)^{-1} X^TX \, w = (\lambda(X^TX)^{-1} + I)^{-1} w$$

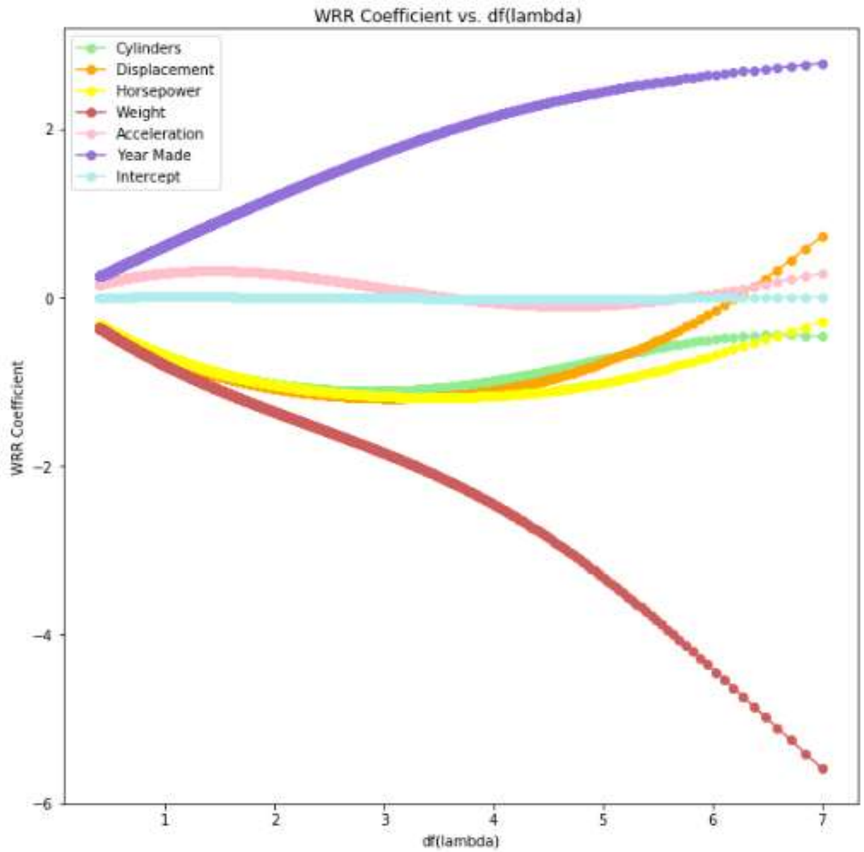
$$var[w_{RR}] = var[(\lambda(X^TX)^{-1} + I)^{-1} w_{LS}]$$

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

$$var(Z w_{LS}) = Z \, var(w_{LS}) Z^T$$

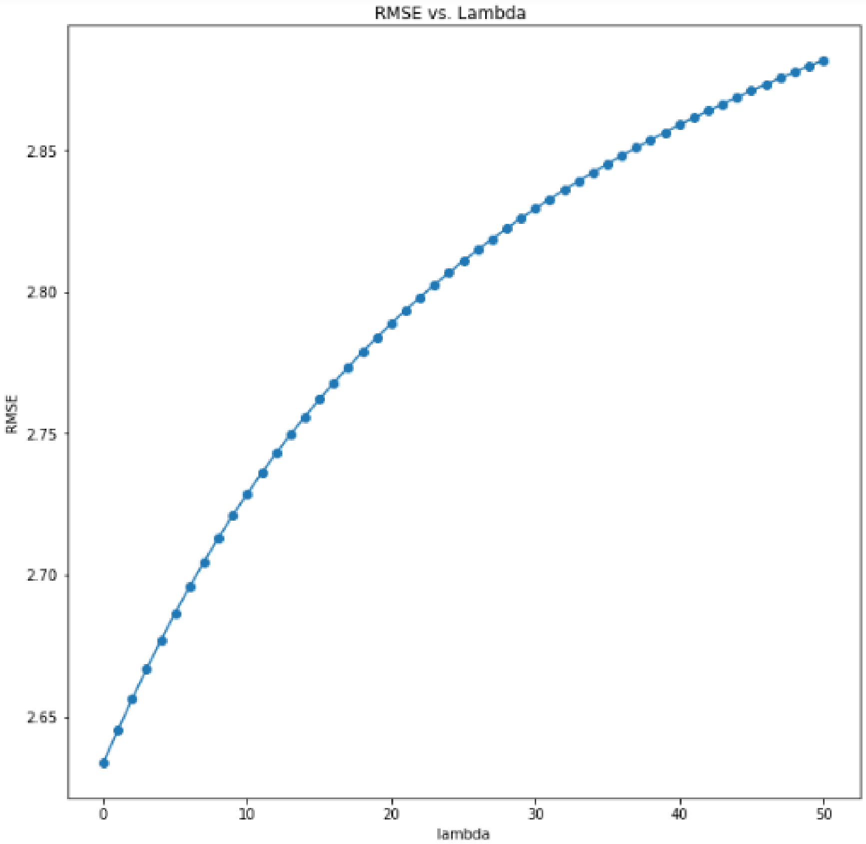
$$= Z \sigma^2 (X^TX)^{-1} Z^T = \boxed{\sigma^2 (I + \lambda(X^TX)^{-1})^{-1} (X^TX^{-1}) ((I + \lambda(X^TX)^{-1})^{-1})^T}$$

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



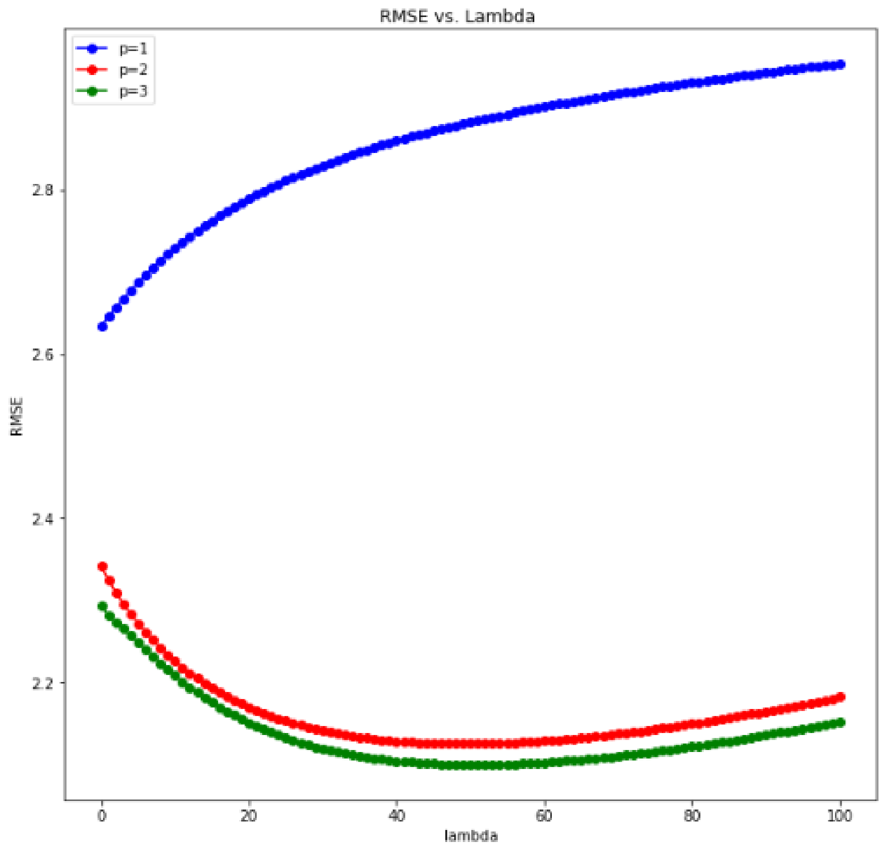
Part b)

The two dimensions that stand out the most over the others are 'Weight' and 'Year Made'. We can glean that these two variables, as the df(lambda) decreases as a result of our lambda increasing and causing the denominator of df(lambda) to get very large, this causes our wrr coefficients to go toward 0. It seems that Year Made and Weight are the two variables whose wrr coefficient does not go to 0 as quickly as the other variables. As a result, the two dimensions above that stand out have the most effect on the dependent variable, mpg for the car.



Insights:

We can see from the graph that at lower values of lambda, we experience low RMSE's. However, as the lambda increases, the RMSE increases, albeit at a decreasing rate. Since the RMSE is minimized when lambda = 0, we are effectively solving a least squares problem. Therefore, we might as well just use a least squares regression method to solve for/estimate w.



Part 2d)

We should choose the p = 3 technique, which involves the transformation and standardization of X to include squared and cubic terms. The general trend of this line decreases when lambda is in between [0,40] and then slowly increases after that, which may signal potential overfitting when lambda is large and the model is more complex. Since p=3 line overall results in a RMSE measure of lambda = 0 through 100, we want to choose this one since it has the least amount of model error. The lambda for which the RMSE is minimized is when lambda = 40.

Observations about the other lines include that the p=2 line also performs surprisingly well compared to p=3, but has a slightly higher RMSE across all lambdas. In stark contrast to these two, p=1 line performs the worse with an overall increasing RMSE across all lambdas, with the minimum RMSE occuring at lambda = 0. This implies that this model does not do terribly well at predicting our DV values. ¶

As for the optimal lambda value for the p=3 lines, which is 40, this value differs from our previous lambda choice, which was lambda = 0 for the p = 1 plot from earlier.

The moral of the story is that the transformation and standardization of our data can help us to achieve a much better model with less error.