COMS4721 Machine Learning for Data Science Homework 1

Po-Chieh Liu

UNI: pl2441

Problem 1.

Given N observations $(x_i, \dots, x_N)$, where $x_i \in \{0, 1, \dots, \infty\}$. Assume x_i are i.i.d and follow Poisson distribution with unknown parameter $\lambda > 0$, i.e. $p(X|\lambda) = \frac{\lambda^X}{X!} e^{-\lambda}$

(a) The joint likelihood of the data can be reduced to product of each observation probability function due to i.i.d assumption and can be expressed as following:

$$p(\lambda; x_1, \dots, x_N) = \prod_{i=1}^{N} p(\lambda; x_i) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

(b) The log likelihood function can be derived from (a) as following

$$L(\lambda) = p(\lambda; x_1, \dots, x_N) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

$$l(\lambda) = ln\big(L(\lambda)\big) = ln\left(\prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right) = ln\left(\frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!} e^{-N\lambda}\right)$$

$$= ln(\lambda) \sum_{i=1}^{N} x_i - ln\left(\prod_{i=1}^{N} x_i!\right) - N\lambda$$

Take partial derivative in regards to λ and set the equation to 0, we can derive the maximum likelihood estimator.

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{\sum_{i=1}^{N} x_i}{\lambda} - N = 0$$

$$\lambda_{mle} = \underset{\lambda}{argmax}\, l(\lambda) = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{X}$$

The maximum likelihood estimator of λ is the sample mean.

(c) Given selecting $p(\lambda) = gamma(a, b) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$. Assume the a and b are shape and rate of gamma distribution.

$$\lambda_{map} = \underset{\lambda}{argmax}\, p(\lambda|X) = \underset{\lambda}{argmax}\frac{p(X|\lambda)p(\lambda)}{\int p(X|\lambda)p(\lambda)d\lambda} = \underset{\lambda}{argmax}\, p(X|\lambda)p(\lambda)$$

$$= \underset{\lambda}{argmax}\big(ln\big(p(X|\lambda)p(\lambda)\big)\big)$$

$$= \underset{\lambda}{argmax}\left(ln(\lambda)\sum_{i=1}^{N} x_i - ln\left(\prod_{i=1}^{N} x_i!\right) - N\lambda + ln\big(p(\lambda)\big)\right)$$

$$= \underset{\lambda}{argmax}\left(ln(\lambda)\sum_{i=1}^{N} x_i - ln\left(\prod_{i=1}^{N} x_i!\right) - N\lambda + ln(b^a) + ln(\lambda^{a-1})\right.$$

$$\left. - b\lambda - \ln\big(\Gamma(a)\big)\right)$$

Take partial derivative in regards to λ and set the equation equal to zero, then we can achieve the maximum value location, which is λ_{map}:

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

$$\lambda_{map} = \frac{\sum_{i=1}^{N} x_i + a - 1}{N + b}$$

(d) Apply Bayes rule on $p(\lambda|X)$:

$$p(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{\int p(X|\lambda)p(\lambda)d\lambda} \propto p(X|\lambda)p(\lambda) = \prod_{i=1}^{N}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda} * \frac{b^a\lambda^{a-1}e^{-b\lambda}}{\Gamma(a)}$$

$$= \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!}e^{-N\lambda} * \frac{b^a\lambda^{a-1}e^{-b\lambda}}{\Gamma(a)} = \frac{b^a\lambda^{\left(\sum_{i=1}^{N} x_i\right)+a-1}e^{-(N+b)\lambda}}{\prod_{i=1}^{N} x_i! * \Gamma(a)}$$

$$= \frac{b^a\lambda^{\left(\sum_{i=1}^{N} x_i+a\right)-1}e^{-(N+b)\lambda}}{\prod_{i=1}^{N} x_i! * \Gamma(a)}$$

By observation, we can find the posterior function $\mathrm{p}(\lambda|\mathrm{X})$ is actually following gamma distribution with parameter $a' = \sum_{i=1}^{N} x_i + a$, and $\mathrm{b}' = N + b$.

(e) From (d), $\mathrm{p}(\lambda|\mathrm{X})$ is a gamma distribution with $\mathrm{a}' = \sum_{i=1}^{N} x_i + a$, and $\mathrm{b}' = N + b$. The mean and variance of gamma distribution are:

$$E[\lambda] = \frac{a'}{b'} = \frac{\sum_{i=1}^{N} x_i + a}{N + b}$$

$$Var(\lambda) = \frac{a'}{b'^2} = \frac{\sum_{i=1}^{N} x_i + a}{(N + b)^2}$$

Compare those two estimators, the λ_{mle} is mainly derived from observations. In

contrast, λ_{map} is obtained from both observations and selected prior distribution (gamma function in this case). The differences between λ_{mle} and λ_{map} can be thought as the interactions or balances between observations and prior assumptions. And λ_{map} is trying to maximize the product of prior probability and maximum likelihood in the same time.

Problem 2.

Given $(x_i, y_i), i = 1, \dots, n$, where $\mathrm{x} \in \mathrm{R}^{\mathrm{d}}, y \in R$. Assume $\mathrm{y_i} \sim N(x_i^T w, \sigma^2)$ and i.i.d. Apply Ridge regression to estimate w and from course slide we know $\mathrm{w_{LS}} = (X^T X)^{-1} X^T y$ and $\mathrm{w_{RR}} = (\lambda I + X^T X)^{-1} X^T y$. Note, only unknown here is y.

Derive expectation of $\mathrm{w_{RR}}$:

$$\begin{aligned}\mathrm{E}[\mathrm{w_{RR}}] &= E[(\lambda I + X^T X)^{-1} X^T y] = E[(\lambda I + X^T X)^{-1} (\boldsymbol{X^T X})(\boldsymbol{X^T X})^{-1} X^T y] \\ &= E[(\lambda I + X^T X)^{-1} (X^T X) w_{LS}] = (\lambda I + X^T X)^{-1} (X^T X) E[w_{LS}] \\ &= (\lambda I + X^T X)^{-1} X^T X w\end{aligned}$$

For deriving variance of $\mathrm{w_{RR}}$, we apply variance property $\mathrm{Var}(\mathrm{a} * \mathrm{Y}) = \mathrm{a}^2 Var(Y)$. Also from course slide we know that $\mathrm{Var}[\mathrm{w_{LS}}] = \sigma^2 (X^T X)^{-1}$ and $w_{RR} = (\lambda (X^T X)^{-1} + I)^{-1} w_{LS}$. Derive variance of $\mathrm{w_{RR}}$:

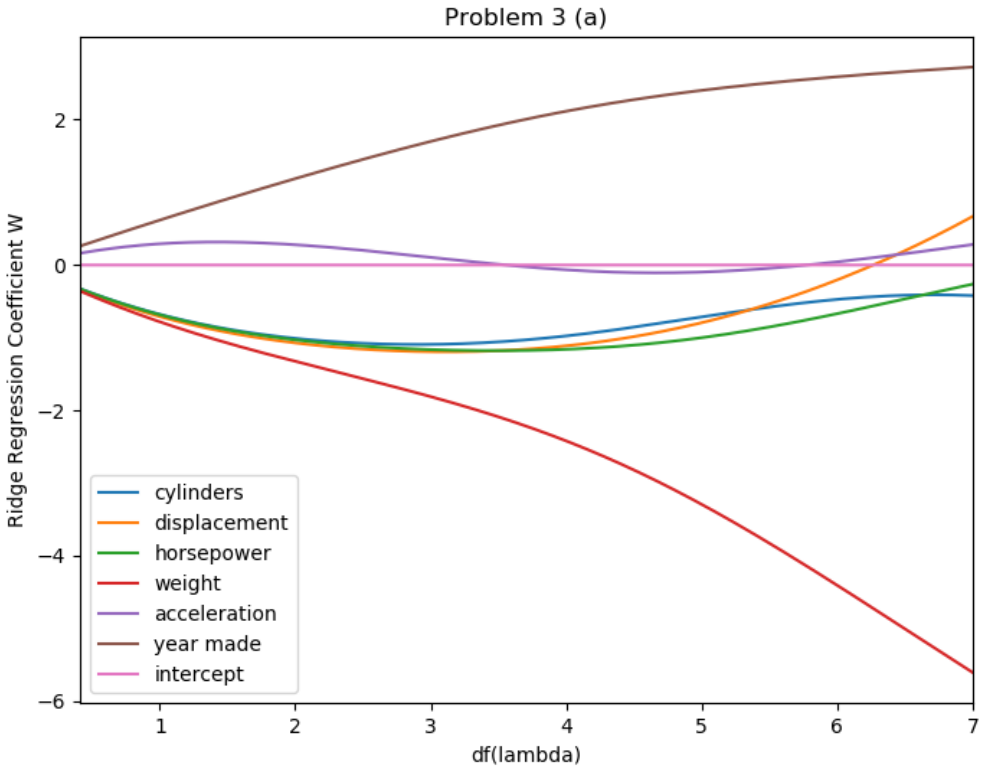
$$\begin{aligned}\mathrm{Var}[w_{RR}] &= Var[(\lambda (X^T X)^{-1} + I)^{-1} w_{LS}] \\ &= (\lambda (X^T X)^{-1} + I)^{-1} Var[w_{LS}] [(\lambda (X^T X)^{-1} + I)^{-1}]^T \\ &= (\lambda (X^T X)^{-1} + I)^{-1} \sigma^2 (X^T X)^{-1} [(\lambda (X^T X)^{-1} + I)^{-1}]^T \\ &= \sigma^2 (\lambda (X^T X)^{-1} + I)^{-1} (X^T X)^{-1} [(\lambda (X^T X)^{-1} + I)^{-1}]^T = \sigma^2 Z (X^T X)^{-1} Z^T\end{aligned}$$

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

Problem 3.

Note, for all 4 questions in problem 3, I used normalized data to make plot. For features (dimensions), the values were subtracted mean and divided by the variance. For predictor, y, the values were subtracted mean only.

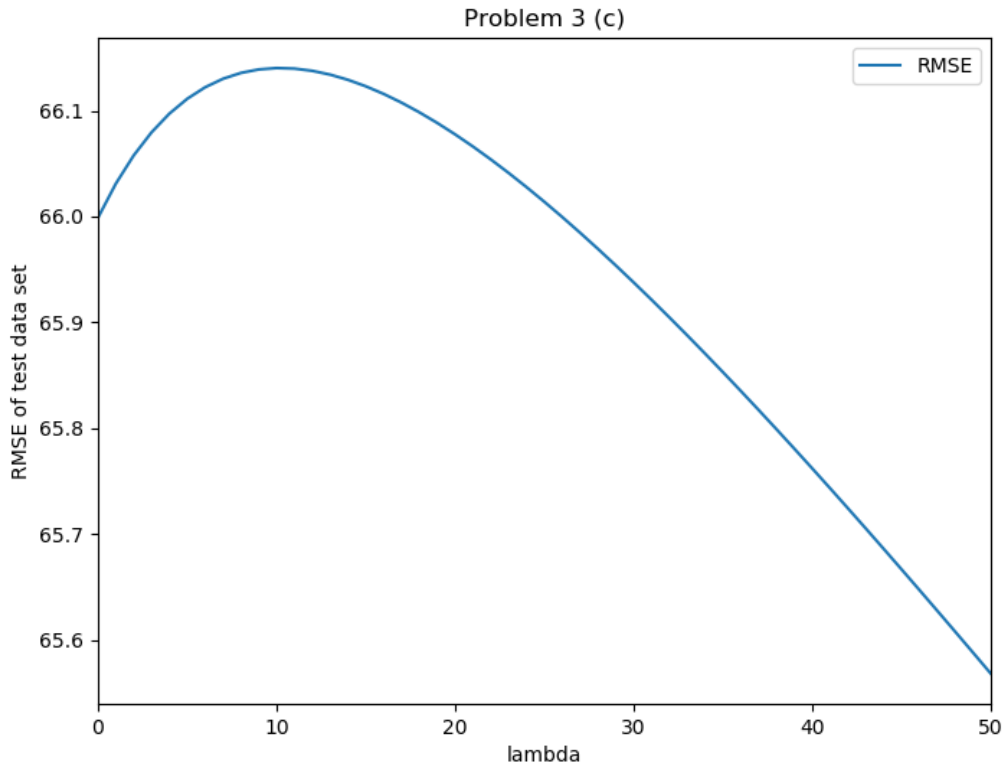
(a) The plot is attached:



(b) From the plot, we can observe that dimensions “year made” and “weight” are the two most stand out dimensions. When the degree of freedom is large, i.e. λ is small, the “year made” dimension has highly positive correlation (w) with miles per gallon usage, and weight has highly negative correlation (w). Those make sense to me because new build cars have novel technologies may help improve oil usage efficiency. And heavy weight car need more power to move which might reduce the oil usage efficiency. When degree of freedom decreases, i.e. λ is large, the model is in favor to minimize the penalty term. All weight coefficients are decreased to lower the penalty. We can observe all weights are close to 0 at degree equals to 0. For other features, we can observe Intercept is not changing when degree of freedom changed. Acceleration features is not sensitive to degree of freedom, however,

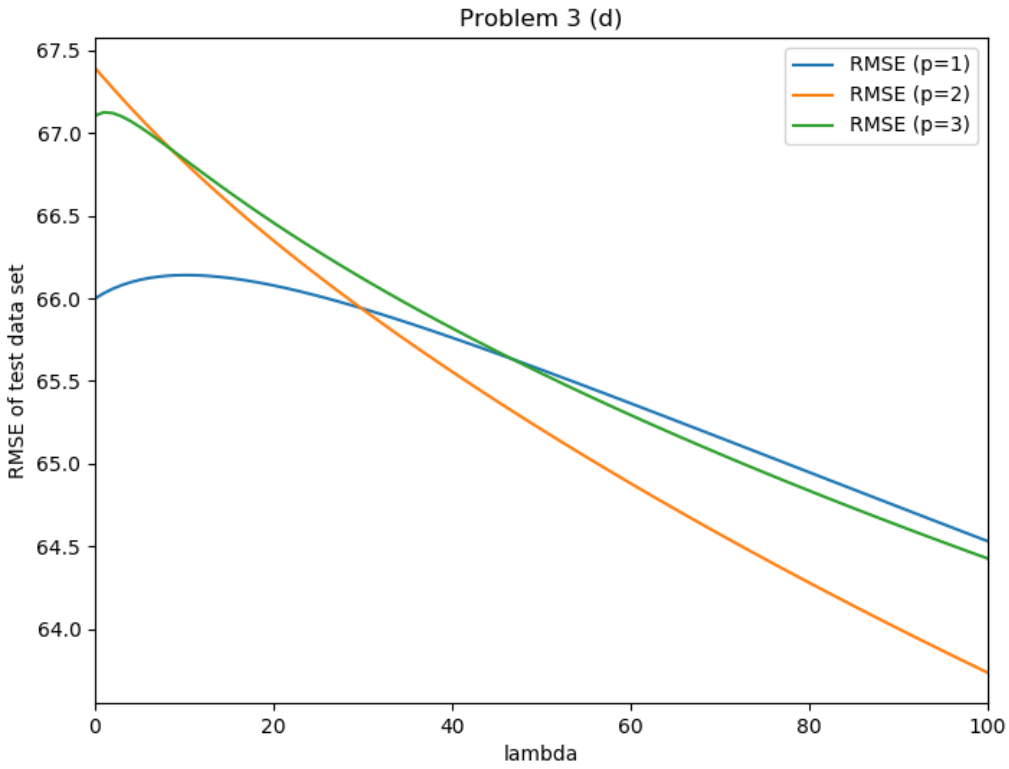
acceleration weight sign changed from positive to negative around 3.5 degree of freedom. The rest three features share similar fashion, but only displacement weights changed from negative to positive around degree of freedom at 6.5.

(c) The plot is attached:



From the plot, we can observe before lambda greater around 30, the RMSE are greater than least squares model. When lambda keeps increasing to 50, the RMSE keep decreasing from 66 to around 65.6. From this plot, I will pick Ridge regression model with lambda equals 50 due to the lowest RMSE.

(d) The plot is attached:



Based on the plot and different purposes I will choose different setting. For best predicting result, I will choose p equals to 3 and lambda equals to 100 due to lowest RMSE. For better model interpretation, I will choose p equals to 1 and lambda equals to 100 due to model simplicity.

Note that the best RMSE between each p values are not very huge, the RMSE approximately reduced 1% from 65 to 64 by changing p from 1 to 3.

Also, looks like the RMSE keep decreasing when increasing lambda, we should try larger lambda to see if we can construct model with lower RMSE.