Machine Learning for Data Science
HW1

Xinyuan Cao
Uni: xc2461

February 13, 2019

1. (a) Given data $(x_1, \cdots, x_N)$, $x_i \overset{iid}{\sim} P(X|\lambda) = \frac{\lambda^X}{X!}e^{-\lambda}$ So the joint likelihood of the data is

$$P(x_1, \cdots, x_N|\lambda) = \prod_{i=1}^{N} P(x_i|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!}e^{-N\lambda}$$

(b)

$$\lambda_{ML} = \underset{\lambda}{\operatorname{argmax}}\, P(x_1, \cdots, x_N|\lambda) = \underset{\lambda}{\operatorname{argmax}}\, \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!}e^{-N\lambda}$$

$$\nabla_\lambda \prod_{i=1}^{N} P(x_i|\lambda) = 0$$

$$\frac{1}{\prod_{i=1}^{N} x_i!}(\sum_{i=1}^{N} x_i \lambda^{\sum_{i=1}^{N} x_i - 1} + \lambda^{\sum_{i=1}^{N}} x_i(-N)e^{-N\lambda}) = 0$$

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

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

Therefore we have the maximum likelihood estimate $\lambda_{ML} = \sum_{i=1}^{N} x_i/N$.

(c) Given the prior $p(\lambda) = gamma(a, b) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$, we have

$$\begin{aligned} P(\lambda|x_1, \cdots, x_N) &\propto P(x_1, \cdots, x_N|\lambda)P(\lambda) \\ &\propto \lambda^{\sum_{i=1}^{N} x_i} e^{-N\lambda}\lambda^{a-1}e^{-b\lambda} \\ &= \lambda^{\sum_{i=1}^{N} x_i + a - 1} e^{-(N+b)\lambda} \end{aligned}$$

$$
\begin{aligned}
\lambda_{MAP} &= \operatorname*{argmax}_{\lambda} P(\lambda|x_1, \cdots, x_N) \\
&= \operatorname*{argmax}_{\lambda} \lambda^{\sum_{i=1}^{N} x_i + a - 1} e^{-(N+b)\lambda} \times \text{constant} \\
&= \operatorname*{argmax}_{\lambda} (\sum_{i=1}^{N} x_i + a - 1) \log \lambda - (N+b)\lambda
\end{aligned}
$$

So we get

$$
\begin{aligned}
\nabla_{\lambda}[(\sum_{i=1}^{N} x_i + a - 1) \log \lambda - (N+b)\lambda] = 0 \\
\frac{\sum_{i=1}^{N} x_i + a - 1}{\lambda} - (N+b) = 0 \\
\lambda = \frac{\sum_{i=1}^{N} x_i + a - 1}{N+b}
\end{aligned}
$$

Here we derive the MAP estimate $\lambda_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{N+b}$.

(d) As we show in (c) we have

$$
P(\lambda|x_1, \cdots, x_N) \propto \lambda^{\sum_{i=1}^{N} x_i + a - 1} e^{-(N+b)\lambda}
$$

This is the format of Gamma distribution (after regularizing so that the sum of the probability is equal to 1). So we know that

$$
P(\lambda|x_1, \cdots, x_N) = \Gamma(\sum_{i=1}^{N} x_i + a, N+b)
$$

The posterior distribution of λ is the Gamma distribution with parameter $\sum_{i=1}^{N} x_i + a$ and $N+b$.

(e) According to the properties of Gamma distribution we have

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

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

Here we may find that when N goes to infinity, λ_{ML} and λ_{MAP} will converge to the same. So it shows that the larger the sample size is, the less influence the prior knowledge has. Also we find that given the prior Gamma(1,b), when $b \to 0^+$, λ_{MAP} goes to λ_{ML}. Besides we find the expectation of γ is not equal to MAP estimate, so this is a biased estimate.

2. From the lecture, we have $y \sim N(Xw, \sigma^2 I)$, $w_{RR} = (\lambda I + X^T X)^{-1} X^T y$, and we know $E(yy^T) = \sigma^2 I + (Xw)(Xw)^T$.

$$\begin{aligned}
E[w_{RR}] =& E[(\lambda I + X^T X)^{-1} X^T y] \\
=& \int [(\lambda I + X^T X)^{-1} X^T y] P(y|X, w) dy \\
=& (\lambda I + X^T X)^{-1} X^T E[y] \\
=& (\lambda I + X^T X)^{-1} X^T X w
\end{aligned}$$

$$\begin{aligned}
Var[w_{RR}] =& E[(w_{RR} - E[w_{RR}])(w_{RR} - E[w_{RR}])^T] \\
=& E[w_{RR} w_{RR}^T] - E[w_{RR}] E[w_{RR}]^T \\
=& E[(\lambda I + X^T X)^{-1} X^T y y^T X (\lambda I + X^T X)^{-T}] \\
&- (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\lambda I + X^T X)^{-T} \\
=& (\lambda I + X^T X)^{-1} X^T E[yy^T] X (\lambda I + X^T X)^{-T} \\
&- (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\lambda I + X^T X)^{-T} \\
=& (\lambda I + X^T X)^{-1} X^T \sigma^2 I X (\lambda I + X^T X)^{-T} \\
&+ (\lambda I + X^T X)^{-1} X^T (Xw)(Xw)^T X (\lambda I + X^T X)^{-T} \\
&- (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\lambda I + X^T X)^{-T} \\
=& (\lambda I + X^T X)^{-1} X^T \sigma^2 I X (\lambda I + X^T X)^{-T} \\
&+ (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\lambda I + X^T X)^{-T} \\
&- (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\lambda I + X^T X)^{-T} \\
=& (\lambda I + X^T X)^{-1} X^T \sigma^2 I X (\lambda I + X^T X)^{-T} \\
=& \sigma^2 (\lambda I + X^T X)^{-1} X^T X (\lambda I + X^T X)^{-T} \\
=& \sigma^2 [(X^T X)(I + \lambda (X^T X)^{-1})]^{-1} X^T X [(X^T X)(I + \lambda (X^T X)^{-1})]^{-T} \\
=& \sigma^2 (I + \lambda (X^T X)^{-1})^{-1} (X^T X)^{-1} X^T X (X^T X)^{-T} (I + \lambda (X^T X)^{-1})^{-T} \\
=& \sigma^2 (I + \lambda (X^T X)^{-1})^{-1} (X^T X)^{-1} X^T X (X^T X)^{-1} (I + \lambda (X^T X)^{-1})^{-T} \\
=& \sigma^2 (I + \lambda (X^T X)^{-1})^{-1} (X^T X)^{-1} (I + \lambda (X^T X)^{-1})^{-T}
\end{aligned}$$

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

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

(a) The result is shown in Figure 1.

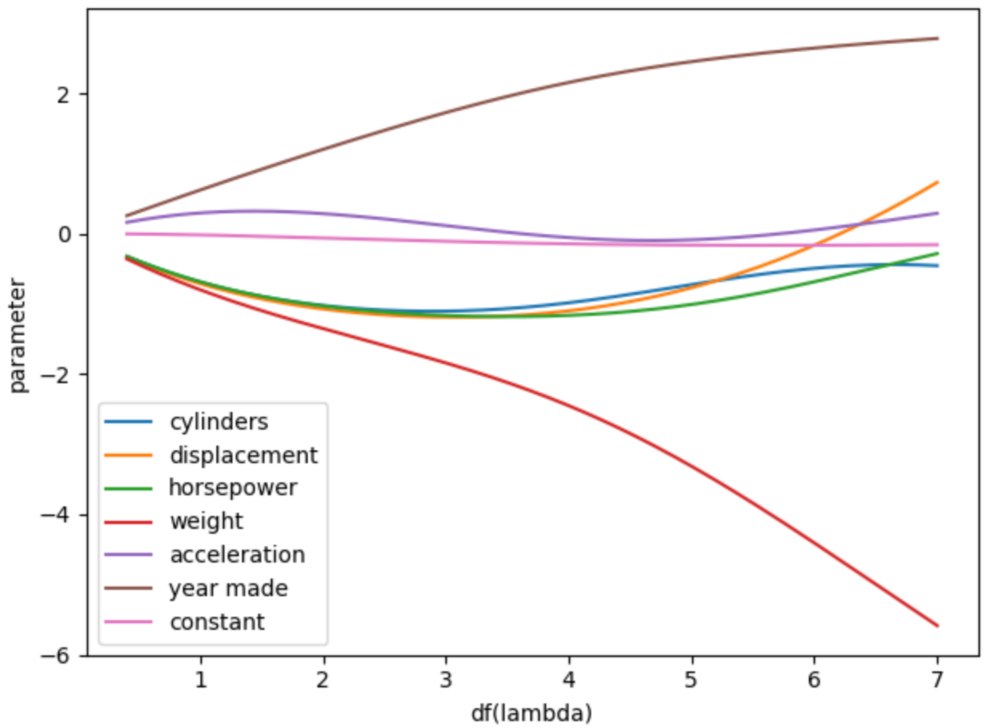

Figure 1: Q3(a) parameters with respect to df(λ)

(b) From Figure 1 we find that two features **weight** and **year made** stand out over the others. It shows that these two features heavily affect y (miles per gallon). They are the most important features relevant to y. Also from the figure we know that **weight** negatively affect y, and **year made** positively affect y. That is to say the larger weight is, the smaller y will be, and the larger year made is, the larger y will be.

(c) The result is shown in Figure 2. We find that RMSE increases when λ increases. So to get minimum RMSE, we should let λ to be as small as possible (let $\lambda = 0$). That is to say, under this circumstance, we should choose least square instead of ridge regression to get smaller RMSE.

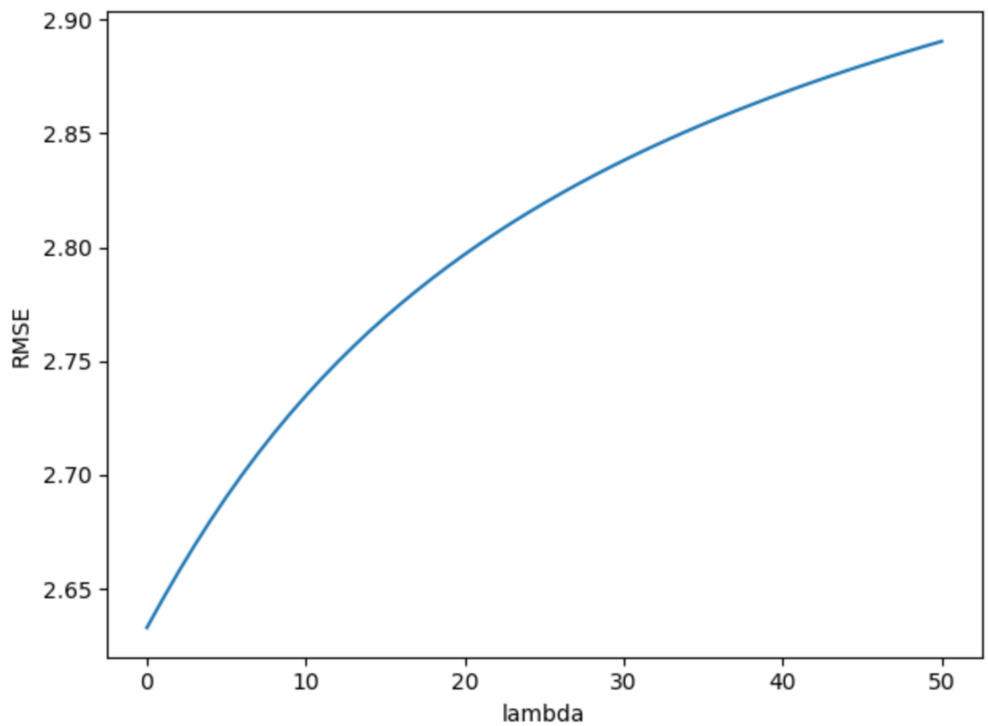

Figure 2: Q3(c) root mean squared error with respect to λ

(d) As shown in Figure3, we find the RMSE can be much smaller when we learn a 2nd/3rd-order polynomial regression comparing to 1st-order one. And for p = 2 or 3, we find RMSE reaches the lowest point when λ is around 50. This is different from Q3(c) where $\lambda = 0$ reaches the minimum RMSE. This is maybe because the first order polynimial regression underfits the data. Since the model itself cannot describe the data well, and the larger λ is, the worst it becomes. But for second or third order ones, the regression model can fit the data much better, and so here ridge regression with proper λ can have better result. So here I will choose $p = 3$ and $\lambda = 50$.

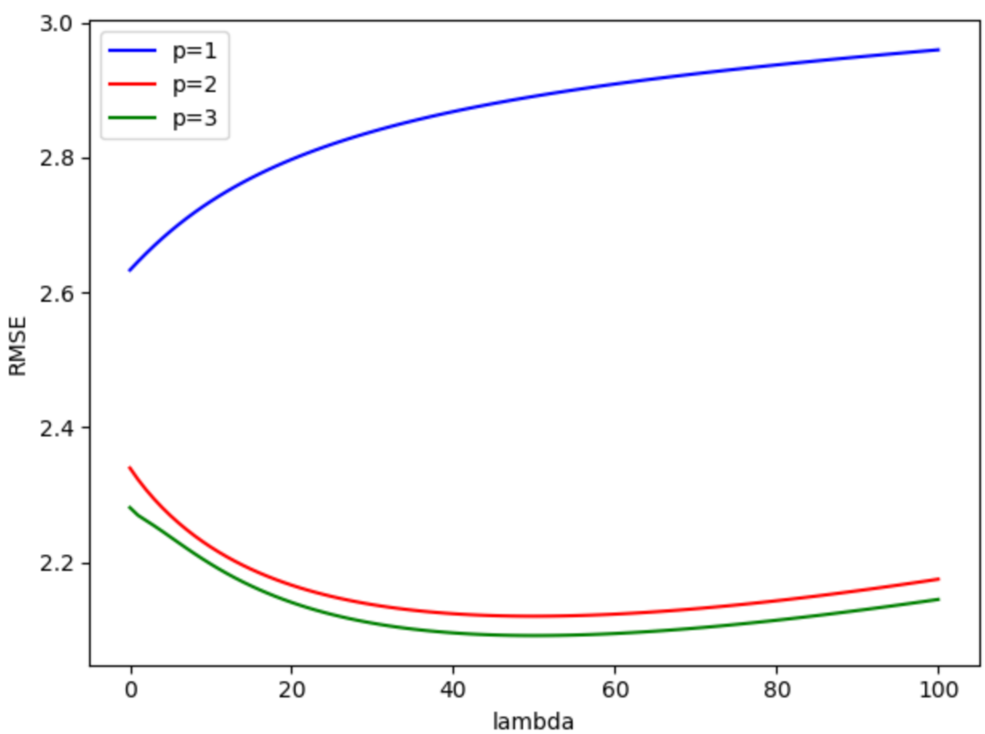

Figure 3: Q3(d) root mean squared error with respect to λ under order p=1,2,3