HW01
Yawen Han
UNI: yh3069
Feb 13, 2018

Problem 1

(a)

$X_1, X_2 \dots X_N$ have joint likelihood denoted as

$$p_\lambda(x_1, x_2, \dots x_N) = p(x_1, x_2, \dots x_N|\lambda) = p(x_1|\lambda) \cdot p(x_2|\lambda) \dots p(x_N|\lambda)$$
$$= \prod_{i=1}^{N} p(x_i|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

(b)

According to the joint likelihood in (a), the log likelihood will be:

$$l(\lambda) = \log\, p_\lambda(x_1, x_2, \dots x_N) = \log \prod_{i=1}^{N} p(x_i|\lambda) = \log \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$
$$= \sum_{i=1}^{N} \log\left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right) = \sum_{i=1}^{N} (x_i \log\lambda - \lambda - \log x_i!)$$
$$= \log\lambda \sum_{i=1}^{N} x_i - N\lambda - \sum_{i=1}^{N} \log x_i!$$

Find the maximum by making the derivative of $l(\lambda)$ be 0:

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

that is,

$$\widehat{\lambda_{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{X}$$

(c)

The prior distribution $\mathrm{p}(\lambda) = \mathrm{gamma}(\mathrm{a}, \mathrm{b}) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$, then the posterior:

$$\mathrm{p}(\lambda|\mathrm{X}) = \frac{p(X|\lambda)p(\lambda)}{p(X)} \propto p(X|\lambda)p(\lambda)$$

thus the MAP estimation λ_{MAP},

$$\lambda_{MAP} = \arg\max_\lambda p(X|\lambda)p(\lambda) = \arg\max_\lambda \log p(X|\lambda)p(\lambda)$$
$$= \arg\max_\lambda \log \prod_{i=1}^{N} p(x_i|\lambda)\ p(\lambda)$$

$$= \arg\max_{\lambda} \sum_{i=1}^{N} \log p(x_i|\lambda)\ p(\lambda)$$

Calling the objective $\mathcal{L}(\lambda)$, then,

$$\text{set } \mathcal{L}(\lambda) = \sum_{i=1}^{N} \log p(x_i|\lambda)\ p(\lambda)$$

$$= \log\lambda \sum_{i=1}^{N} x_i - N\lambda - \sum_{i=1}^{N} \log x_i! + \log p(\lambda)$$

$$= \log\lambda \sum_{i=1}^{N} x_i - N\lambda - \sum_{i=1}^{N} \log x_i! + a\log b + (a-1)\log\lambda - \mathrm{b}\lambda - \log\Gamma(a)$$

Find the maximum by making the derivative of $\mathcal{L}(\lambda)$ be 0:

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

that is,

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

(d)

The prior distribution $\mathrm{p}(\lambda) = \mathrm{gamma}(\mathrm{a}, \mathrm{b}) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$, then the posterior:

$$\mathrm{p}(\lambda|\mathrm{X}) = \frac{p(X|\lambda)p(\lambda)}{p(X)} \propto p(X|\lambda)p(\lambda)$$

$$= \left[\prod_{i=1}^{N} p(x_i|\lambda)\right]\left[\frac{b^a\lambda^{a-1}e^{-b\lambda}}{\Gamma(a)}\right]$$

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

$$\propto \lambda^{a-1+\sum x_i} e^{-(b+N)\lambda}$$

We recognize this as $\mathrm{p}(\lambda|\mathrm{X}) = gamma(a + \sum_{i=1}^{N} x_i\,, b + N)$, which is also a gamma distribution. The posterior distribution of λ stays the same class of function as the prior distribustion.

(e)

According to the property of gamma distribution, the mean and variance for the posterior is,

$$\mathrm{E}[\lambda|\mathrm{X}] = \frac{a + \sum_{i=1}^{N} x_i}{b + N}, Var(\lambda|X) = \frac{a + \sum_{i=1}^{N} x_i}{(b+N)^2}$$

The mean of λ under this posterior is nearly the same as the MAP estimate found in question(a). $\lambda_{ML}, \lambda_{MAP}$ *and* E[λ|X] have the similar expression, and λ_{ML} maximize the likelihood and λ_{MAP} maximize the posterior. As N grows, both numerators and denominators in the expressions above become increasingly more similar, which means that they approach each other for the large dataset. In other words, large data diminishes the importance of prior knowledge when estimating λ.

Problem 2

Before we calculate the expectation and variance of the ridge estimator,

We know that $\mathbb{E}[w_{ML}] = \mathbb{E}[(X^TX)^{-1}X^Ty] = (X^TX)^{-1}X^T\mathbb{E}[y] = (X^TX)^{-1}X^TXw = w$

$$\mathrm{Var}[w_{ML}] = \mathbb{E}[w_{ML}w_{ML}^T] - \mathbb{E}[w_{ML}]\mathbb{E}[w_{ML}]^T = \mathbb{E}[(X^TX)^{-1}X^Tyy^TX(X^TX)^{-1}] - ww^T$$

$$= (X^TX)^{-1}X^T\mathbb{E}[yy^T]X(X^TX)^{-1} - ww^T$$

$$= (X^TX)^{-1}X^T(\sigma^2I + Xww^TX^T)X(X^TX)^{-1} - ww^T = \sigma^2(X^TX)^{-1}$$

Then we can use the result above $\mathbb{E}[w_{ML}] = w, and\ \mathrm{Var}[w_{ML}] = \sigma^2(X^TX)^{-1}$ to calculate the expectation and variance of ridge estimator.

The expectation of the ridge estimator:

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

$$= \mathbb{E}[(\lambda I + X^TX)^{-1}(X^TX)(X^TX)^{-1}X^Ty]$$

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

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

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

The variance of the ridge estimator:

$$\mathrm{Var}[w_{RR}] = \mathrm{Var}[(\lambda I + X^TX)^{-1}X^Ty]$$

$$= \mathrm{Var}[(\lambda I + X^TX)^{-1}(X^TX)(X^TX)^{-1}X^Ty]$$

$$= \mathrm{Var}[\{(X^TX)^{-1}(\lambda I + X^TX)\}^{-1}(X^TX)^{-1}X^Ty]$$

$$\text{set } Z = \{(X^TX)^{-1}(\lambda I + X^TX)\}^{-1} = [\lambda(X^TX)^{-1} + I]^{-1}$$

$$then, \mathrm{Var}[w_{RR}] = Var[Z(X^TX)^{-1}X^Ty]$$

$$= Var[Zw_{ML}]$$

$$= Z\, Var[w_{ML}]\, Z^T$$

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

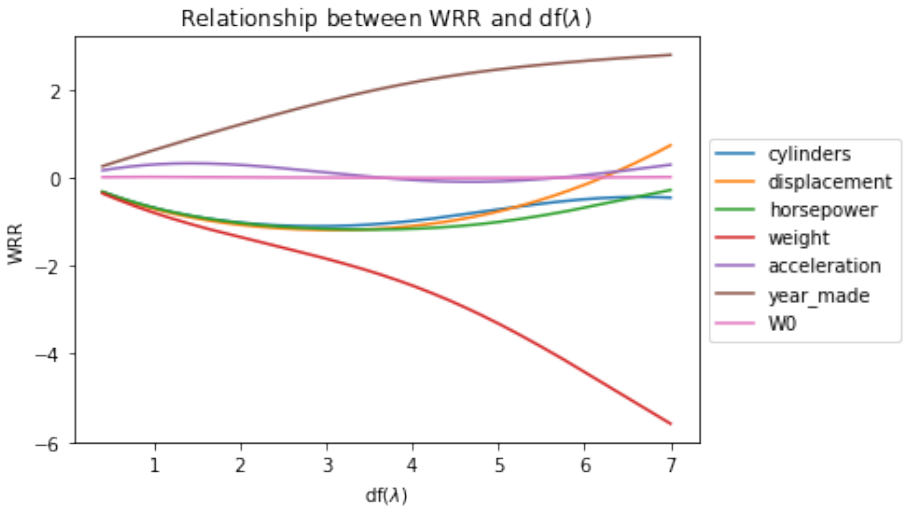
$$= \sigma^2 Z (X^T X)^{-1} Z^T$$

where, $Z = \{(X^T X)^{-1}(\lambda I + X^T X)\}^{-1} = [\lambda (X^T X)^{-1} + I]^{-1}$

Problem 3

(a)

Plot the 7 values in w_{RR} as a function of df(λ) below, each curve is labeled based on its dimension:

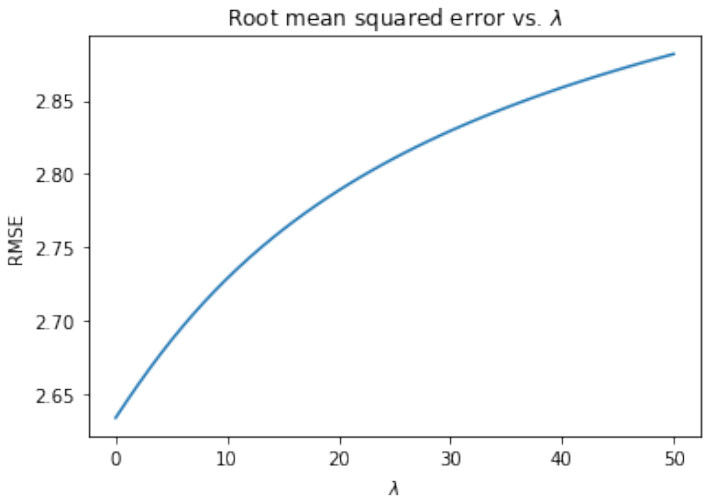


(b)

Two dimensions are clearly stand out over the others: “year_made” and “weight”. These two curves have greater change compared to others, meaning that the coefficients for these two features have more shrink towards 0 as λ goes from 0 to infinity. The value of lambda determines the importance of this penalty term. When lambda is zero, the result will be same as conventional regression; when the value of lambda is large, the coefficients will approach zero. Thus, with the increasing of the penalty term λ, the coefficients for “year_made” and “weight” are shrink a lot compared to other features.

(c)

Plot the root mean squared error (RMSE) on the test set as a function of λ:

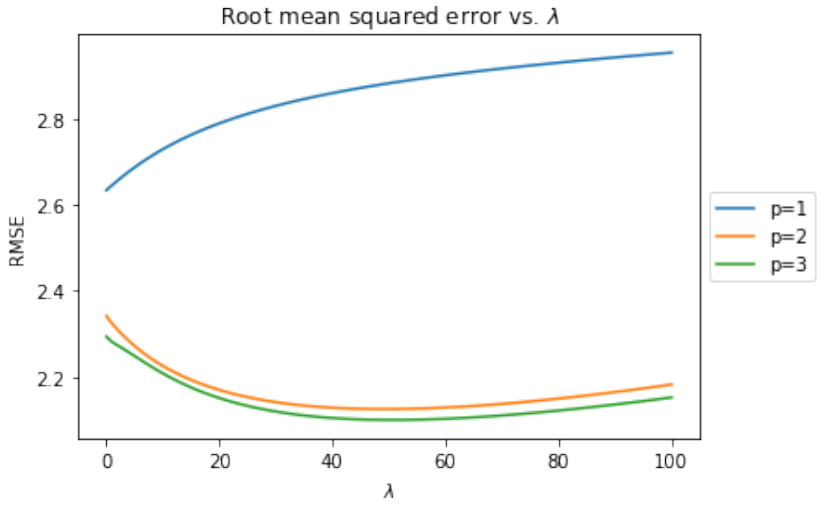


The figure shows that the root mean squared error (RMSE) on the test-set keeps increasing as λ increases. In other word, when penalizing larger on the sum of squares of the coefficients, the difference between predicted target and true target (RMSE) gets larger, indicating a poorer fit.

Ridge keeps all variables and shrinks the coefficients towards zero. In the plot, when λ values get small, that is unregularized. As RMSE shows a significant increase as λ grows, a better fit can be achieved by picking a small value for λ. Therefore, the resulting plot indicates that the unregularized full model does pretty well in this case. The least squares can give us a goof fit without regularizing the coefficients.

(d)

Plot the root mean squared error (RMSE) on the test-set as a function of $\lambda = 0, \ldots 100$ for p = 1,2,3 :



The plot above shows that

1) for p=1, RMSE has a significant increase as λ grows;
2) for p=2 and p=3, RMSE decreases at first and increases later as λ grows;
3) The similar changing behavior is observed for p=2 and p=3, and the difference between their RMSE value is relatively small.

According to the observations above, p=2 and p=3 provide a better fit on the test data with smaller RMSE values. Considering the similar changing behavior as well as the relative small difference between their RMSE values for p=2 and p=3, I will choose the 2nd-order polynomial regression model. In other word, the 2nd-order polynomial regression model can not only provide a good fit on the test data, but also decrease the complexity of the model.

For the 2nd-order polynomial regression model I choose, the ideal choice for λ is [40,60], as this range achieves the lowest stationary of the plot. What's more, with λ getting bigger from 1 to 3, the complexity of the model increases, and the ideal value for λ also getting bigger from 0 to [40,60]. This is because we need more penalization on coefficients to avoid overfitting of the complex model.