ELEN 4903 Machine Learning

HW1

Name: Wenbo Song
UNI: ws2505

Problem 1

(a) Because $x_1, x_2 \dots . x_n$ are i.i.d, given that $p(x_i|\pi) = \pi^{xi}(1-\pi^{1-xi})$ the joint likelihood of (x_1, …,x_N) is shown as follow

$$\begin{aligned} p(x_1, x_2 \dots . x_n|\pi) &= p(x_1|\pi) \dots \dots . p(x_n|\pi) \\ &= \left(\pi^{x1}(1-\pi^{1-x1})\right) \dots \dots . . \left(\pi^{xN}(1-\pi^{1-xN})\right) \\ &= \pi^{\sum_{i=1}^{n} xi}(1-\pi)^{\sum_{i=1}^{n} 1-xi} \end{aligned}$$

(b) The maximum likelihood estimates $\hat{\pi}_{ML} = argmax(\pi^{\sum_{i=1}^{n} xi}(1-\pi)^{\sum_{i=1}^{n} 1-xi})$

To find $\hat{\pi}_{ML}$, we want to calculate the gradient of $\pi^{\sum_{i=1}^{n} xi}(1-\pi)^{\sum_{i=1}^{n} 1-xi} = 0$ and the solution is the $\hat{\pi}_{ML}$.

$$\nabla_\pi \pi^{\sum_{i=1}^{n} xi}(1-\pi)^{\sum_{i=1}^{n} 1-xi} = 0$$

$$\frac{\sum_{i=1}^{n} xi}{\pi} - \sum_{i=1}^{n} xi - N + \sum_{i=1}^{n} xi = 0$$

We get the solution

$$\pi = \frac{\sum_{i=1}^{n} xi}{N}$$

(c) We want to find $\Pr(\pi|X) = \frac{\Pr(X|\pi)\Pr(\pi)}{\Pr(X)}$

Then we have

$$\hat{\pi}_{map} = argmaxPr(\pi|X)$$

$$= argmax \frac{\Pr(X|\pi)\Pr(\pi)}{\Pr(X)}$$

$$= argmaxPr(X|\pi)\Pr(\pi)$$

$$= argmax \prod_{xi \in X} \Pr(Xi|\pi)\Pr(\pi)$$

We write the equation in log form

$$\begin{aligned} argmaxPr(\pi|X) &= argmaxlogPr(\pi|X) \\ &= argmzxlog \prod_{xi \in X} \Pr(Xi|\pi)\Pr(\pi) \\ &= argmax \sum logPr(xi|\pi) + \Pr(\pi) \end{aligned}$$

We want to find $\hat{\pi}_{map}$ which is the π that let

$$\frac{1}{\pi}\sum_{i=1}^{n} xi - \frac{1}{1-\pi}\sum_{i=1}^{n}(1-xi) + \frac{\alpha-1}{\pi} - \frac{\beta-1}{1-\pi} = 0$$

Solve the equation we have

$$\hat{\pi}_{map} = \frac{\sum xi + \alpha - 1}{n+\beta+\alpha-2}$$

(d) if we know that π is followed beta distribution, we have

$$P(\pi) = Beta(\pi|a,b) = \frac{\Gamma(a+b)}{\Gamma(\mathrm{a})\Gamma(\mathrm{b})}\pi^{a-1}(1-\pi)^{b-1}$$

from Bayes rule

$$P(\pi|x_{1,\ldots\ldots,}x_n) = \frac{P(x_{1,\ldots\ldots,}x_n|\pi)P(\pi)}{\int_0^1 P(x_{1,\ldots\ldots,}x_n|\pi)P(\pi)d\pi}$$

We can write $p(\pi|x) \propto p(x|\pi)p(\pi)$

Multiply the two we have

$$P(\pi|x1,\ldots\ldots xn) \propto \pi^{\sum_{i=1}^{n} xi+a-1} 1-\pi^{\sum_{i=1}^{n}(i-xi)+b-1}$$

We can recognize this as $P(\pi|x1,\ldots\ldots xn) = Beta(\sum_{i=1}^{n}(xi+a), \sum_{i=1}^{n}(1-xi)+b)$

So, it's a Beta distribution.

(e) The mean of π is $E[X] = \frac{1}{1+\frac{\beta}{\alpha}}$ $where\ \alpha = \sum_{i=1}^{n} xi + a\ , \beta = \sum_{i=1}^{n}(1-xi)+b$

So we can get that $E[X] = \frac{1}{1+\frac{\sum_{i=1}^{n}(1-xi)+b}{\sum_{i=1}^{n} xi+a}}$

The variance of π is $var(X) = E[(X-u)^2] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

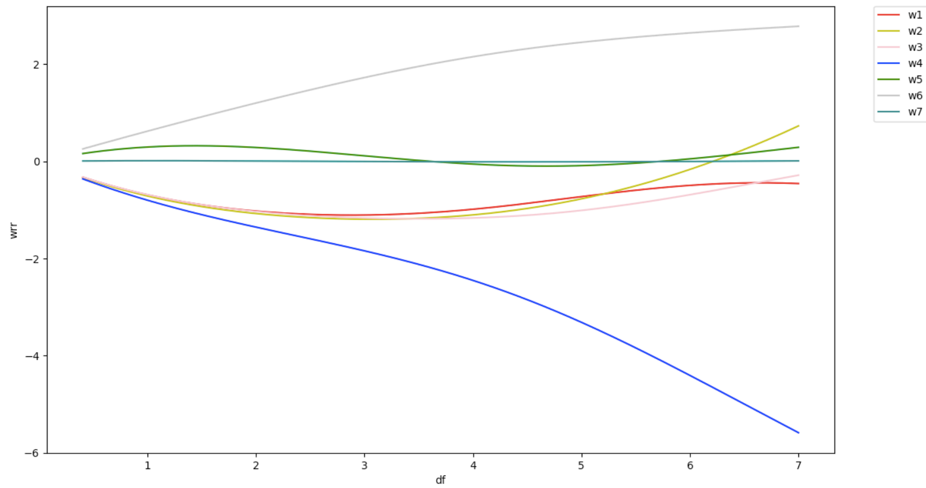
So we can get that $var(X) = \frac{(\sum_{i=1}^{n} xi+a)(\sum_{i=1}^{n}(1-xi)+b)}{(\sum_{i=1}^{n} xi+a+\sum_{i=1}^{n}(1-xi)+b)^2(\sum_{i=1}^{n} xi+a+\sum_{i=1}^{n}(1-xi)+b+1)}$

Relations: $\hat{\pi}_{ML}$ is unbiased but potentially has high variance, by contrast, $\hat{\pi}_{map}$ is biased but has a lower variance than $\hat{\pi}_{ML}$

Problem 2

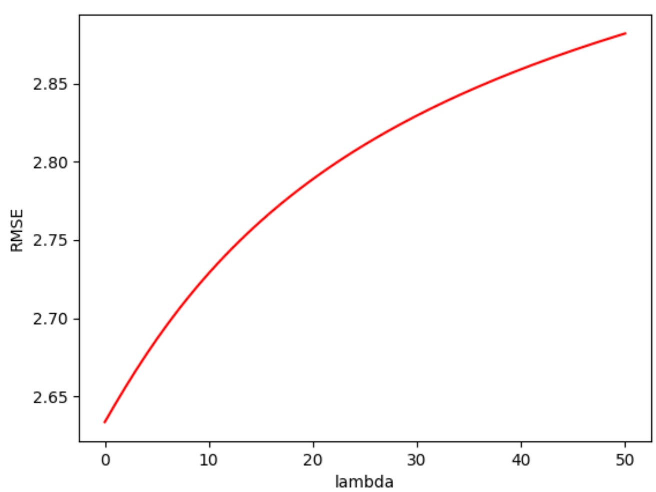
Part 1

(a) For λ = 0,1,2,3….,5000 the relation between wrr and df(λ) is shown as following graph



(b) We can see that w4 and w6 stand out than other features, which indicates that the w4, car weight and the w6, car year has more influence than the other features.

(c) The root mean squared error on the test set is shown as follow:

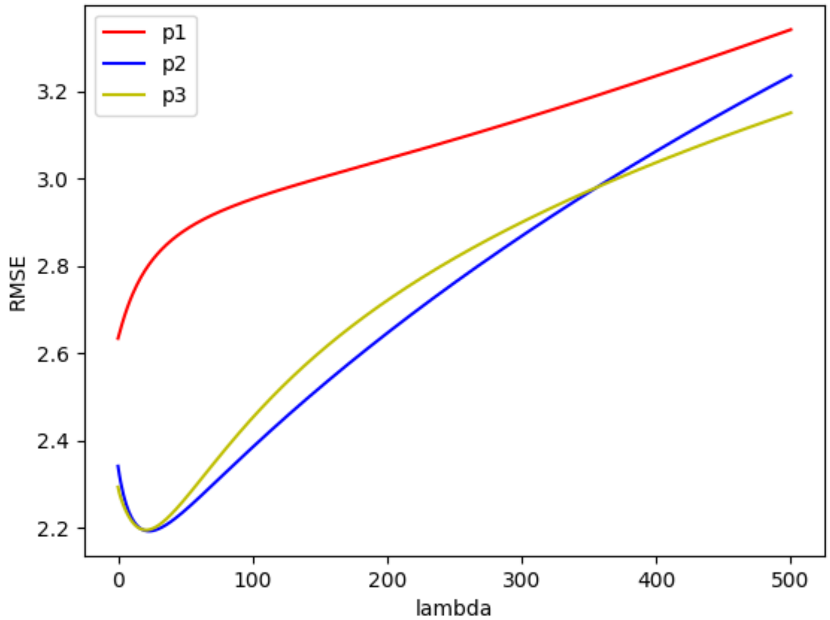


When λ decrease, we can find that eh RMSE decrease, which will lead to a better result, so in this case, we can choose the minimum lambda to get the optimal value.

When λ is 0, the regression became the least square regression. So in this case, we can choose the square regression to get a better performance.

Part2.

(d) The RMSE for p = 1,2,3 can be seen as follow:



Based on this plot, we can see that when $\lambda < 500$, p = 2 and p = 3 has very similar performance. Before $\lambda = 300$, we can choose p2, after $\lambda = 300$ before $\lambda = 500$, we can choose p3.

For λ, we can see that when $\lambda = 20$, all three regression, p1, p2, p3 has the relatively smallest values, so in this problem, we can choose λ to be 20.