ELEN 4720: Machine learning for Signals, Information & Data

Homework-1

Date - 10th October 2021

Name- Rakshith Kamath

UNI- RK3165

Problem-1 -

a) $(x_1, x_2, x_3 \ldots\ldots, x_N)$ iid $\dfrac{\lambda^x e^{-\lambda}}{x!}$

Since each of the observations are independent of each other, the joint probability will be product of the individual probabilities.

$$\therefore \quad p(X_1, X_2, \ldots X_N | \lambda) = \prod_{i=1}^{N} p(x_i | \lambda)$$

$$p(x_1, x_2, x_3, \cdots x_N | \lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

b) Maximum likelihood will be to find the value of λ for which the joint probability we derived would be maximum.

This can be defined by a function as follows-

$$\lambda_{ML} = \arg\max_{\lambda} \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Taking logarithm doesn't change the value of λ_{ML} in this case.

$$\lambda_{ML} = \arg\max_{\lambda} \left[\log \left(\prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right) \right]$$

$$= \arg\max_{\lambda} \left[\sum_{i=1}^{N} x_i \log \lambda - \sum_{i=1}^{N} \lambda \log e - \sum_{i=1}^{N} \log(x_i!) \right]$$

The last term doesn't have a λ term & can be ignored for arg max calculation.

$$\lambda_{ML} = \arg\max_{\lambda} \left[\log \lambda \sum_{i=1}^{N} x_i - N\lambda \right]$$

$$\text{consider } f = \log \lambda \sum_{i=1}^{N} x_i - N\lambda$$

To get λ_{ML}, $\frac{\partial f}{\partial \lambda} = 0$

Applying the same,

$$\frac{1}{\lambda_{ML}} \sum_{i=1}^{N} x_i - N = 0$$

$$\boxed{\lambda_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

c) λ_{MAP} uses posterior distribution to get the optimal values of λ.

It can be defined as follows-

$$\lambda_{MAP} = \arg\max_{\lambda} \ln \left[p(\lambda|X) \right]$$

$$= \arg\max_{\lambda} \ln\left[\frac{p(X|\lambda)\cdot p(\lambda)}{p(X)}\right]$$

$\because$ $p(X)$ doesn't depend on λ, it can be ignored in the calculations. Substituting the other values, we get

$$\lambda_{MAP} = \arg\max_{\lambda}\left[\ln p(X|\lambda) + \ln p(\lambda)\right]$$

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

$$= \arg\max_{\lambda}\left\{\ln(\lambda)\cdot\sum_{i=1}^{N}x_i - \sum_{i=1}^{N}\lambda\ln e - \sum_{i=1}^{N}\ln(x_i!) + \ln b^a + (a-1)\ln\lambda - b\lambda\ln e - \ln(\Gamma[a])\right\}$$

ignoring terms without λ again,

$$= \arg\max_{\lambda}\underbrace{\left\{\ln(\lambda)\sum_{i=1}^{N}x_i - N\lambda + (a-1)\ln\lambda - b\lambda\right\}}_{f}$$

To get λ_{MAP}, $\frac{\partial f}{\partial \lambda} = 0$

i.e,

$$\sum_{i=1}^{N} x_i \cdot \left(\frac{1}{\lambda}\right) - N + \frac{a-1}{\lambda} - b = 0$$

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

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

d) The Bayes rule states that

$$p(\lambda|x) = \frac{p(x|\lambda) \cdot p(\lambda)}{p(x)}$$

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

$$\propto \frac{\lambda^x e^{-N\lambda}}{x!} \cdot \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma[a]}$$

where $X = \sum_{i=1}^{N} x_i$

Ignoring the non-λ terms, we get

$$p(\lambda|x) \propto [\lambda^x e^{-N\lambda} \lambda^{a-1} e^{-b\lambda}]$$

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

Let $x+a$ be u & $b+N$ be v

$$p(\lambda|x) \propto \lambda^{u-1} e^{-\lambda t}$$

we can multiply constants to this

$$p(\lambda|x) \propto \frac{u^v \lambda^{u-1} e^{-\lambda t}}{\Gamma[u]}$$

This shows that $p(\lambda|x)$ is proportional to gamma (u, v).

i.e, $p(\lambda|x)$ has a distribution of gamma$\left(\sum_{i=1}^{N} x_i + a, b+N\right)$

e) Since we know the distribution is gamma, we can say it's mean & variance as follows

$$E[X] = \frac{a}{b} \quad \& \quad Var[X] = \frac{a}{b^2} \quad \text{for } X \sim \text{Gamma}(a,b)$$

To get it's mean

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

& Variance

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

$$E[\lambda] = \frac{\sum_{i=1}^{N} x_i + a}{N+b} \qquad \lambda_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \lambda_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{N+b}$$

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

$$\boxed{\lambda_{MAP} = E[\lambda] - \frac{1}{N+b}}$$

$\therefore \lambda_{MAP} < E[\lambda]$

If $E[\lambda]\Big|_{a,b=0} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\boxed{E[\lambda]\Big|_{a,b=0} = \lambda_{ML}}$$

This says that if a & b values are 0, we get λ_{ML}.

Problem 2- (20 points)

a) Given- Data - (x_i, y_i), $i = 1, 2, \cdots n$. $x \in R^d$ & $y \in R$

Model - $y_i \overset{iid}{\sim} N(x_i^T w, \sigma^2)$

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

Mean of w_{RR} ($E[w_{RR}]$)-

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

Since in this given model, we assume that we know the X matrix & Y matrix is the one that is random. we get

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

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

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

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

$$\boxed{E[w_{RR}] = Z w} \qquad \text{where } Z = \left(\lambda (X^T X)^{-1} + I\right)^{-1}$$

Variance of w_{RR} ($Var(w_{RR})$)

$$Var(w_{RR}) = E\left[(w_{RR} - E[w_{RR}])(w_{RR} - E[w_{RR}])^T\right]$$

$$Var(w_{RR}) = E[w_{RR} w_{RR}^T] - E[w_{RR}] E[w_{RR}]^T \longrightarrow (1)$$

w.r.t $w_{RR} = (\lambda I + X^T X)^{-1} X^T y = \left((X^T X)^{-1} \lambda + I\right)^{-1} (X^T X)^{-1} X^T y$

$$w_{RR} = Z (X^T X)^{-1} X^T y$$

& $E[yy^T] = \sigma^2 I + X w w^T X^T$

Applying these results to (1)

$$Var(w_{RR}) = Z (X^T X)^{-1} X^T E[yy^T] X (X^T X)^{-1} Z^T - Z w w^T Z^T$$

$$= Z (X^T X)^{-1} X^T \left[\sigma^2 I + X w w^T X^T\right] X (X^T X)^{-1} Z^T - Z w w^T Z^T$$

$$= \sigma^2 \left[Z (X^T X)^{-1} X^T X (X^T X)^{-1} Z^T \right] + Z (X^T X)^{-1} X^T X \, w w^T X^T X (X^T X)^{-1} Z^T$$

$$- Z w w^T Z^T$$

$$\boxed{\text{Var}(x) = \sigma^2 \; Z \; (X^T X)^{-1} Z^T}$$

b) $$w_{RR} = (\lambda I + X^T X)^{-1} X^T y \quad , \quad w_{LS} = (X^T X)^{-1} X^T y$$

$$w_{RR} = (\lambda I + X^T X)^{-1} (X^T X) \underbrace{(X^T X)^{-1} X^T y}_{w_{LS}}$$

$$= (\lambda I + X^T X)^{-1} X^T X \; w_{LS}$$

$$= \left[(X^T X) \left(\lambda (X^T X)^{-1} + I \right) \right]^{-1} X^T X \; w_{LS}$$

$$= \left[\lambda (X^T X)^{-1} + I \right]^{-1} (X^T X)^{-1} X^T X \; w_{LS}$$

$$\boxed{w_{RR} = \left[\lambda (X^T X)^{-1} + I \right]^{-1} w_{LS}}$$

w.r.t $X = USV^T$

$$X^TX = VSU^T USV^T$$

$$X^TX = VS^2V^T \quad , (X^TX)^{-1} = VS^{-2}V^T$$

Substituting that in the above equation, we get

$$w_{RR} = \left[\lambda VS^{-2}V^T + I\right]^{-1} w_{LS}$$

$$\boxed{w_{RR} = V\left[\lambda S^{-2} + I\right]^{-1} V^T w_{LS}}$$

It can be re-written as

$$w_{RR} = V M V^T w_{LS} \quad , \text{ where } M = \frac{S_{ii}^2}{\lambda + S_{ii}^2}$$

Problem-3 (Coding) - (30 points)

a)

(a) For $\lambda = 0, 1, 2, 3, \ldots, 5000$, solve for w_{RR}. (Notice that when $\lambda = 0$, $w_{\text{RR}} = w_{\text{LS}}$.) In one figure, plot the 7 values in w_{RR} as a function of $df(\lambda)$. You will need to call a built in SVD function to do this as discussed in the slides. Be sure to label your 7 curves by their dimension in x.[2]

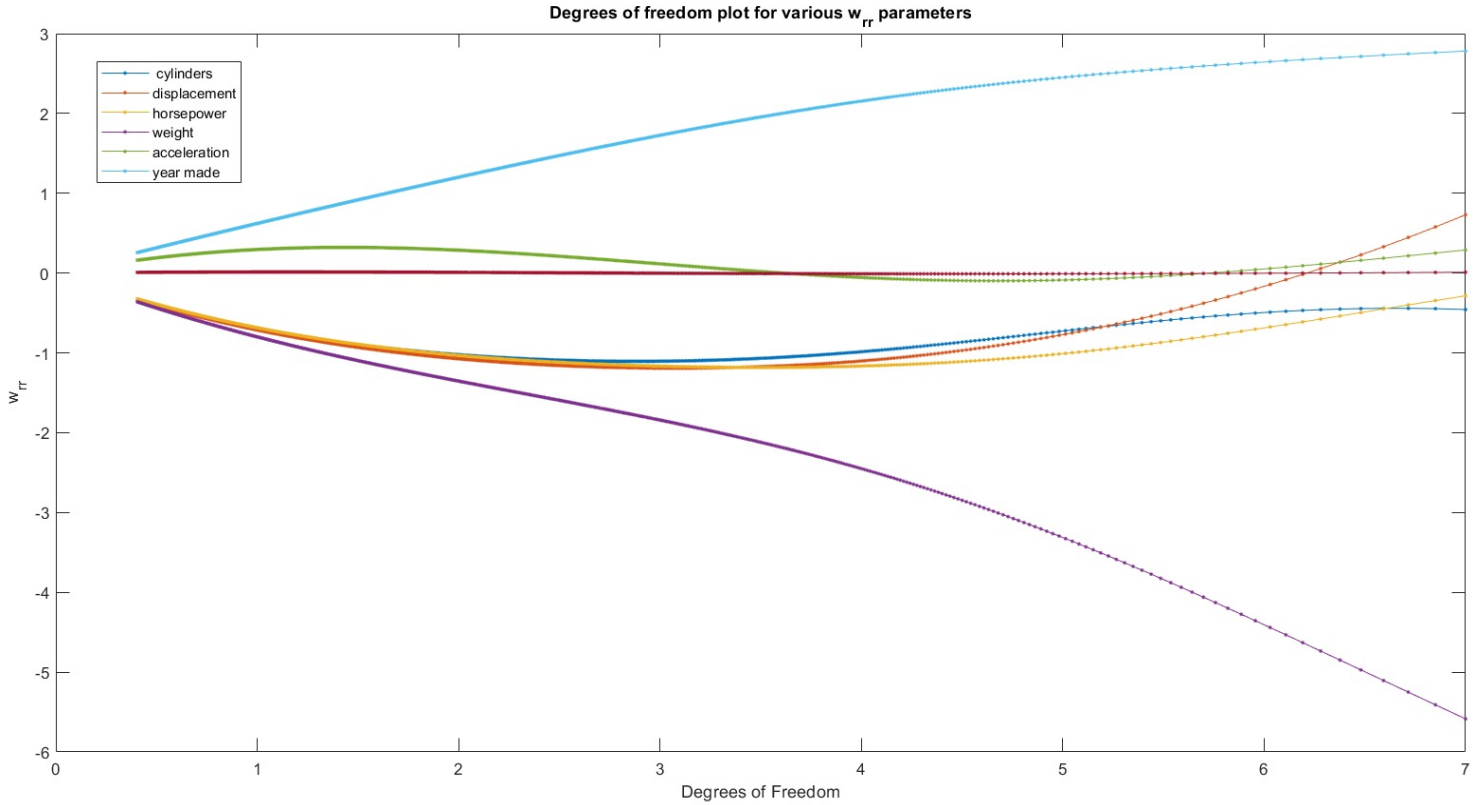

Figure-1

---

b)

(b) Two dimensions clearly stand out over the others. Which ones are they and what information can we get from this?

We clearly see that 'weight' & 'year-made' parameters have deviated the most, hence are key features for the estimation of miles per gallon of a car.

At $\lambda = 0$, i.e, the Least squares method the above 2 parameters have the highest absolute

value weights which suggests it's importance. Also, as λ value increases, it's degree of freedom decreases rapidly, thus furthering it's importance.

The information above says that as the 'year made' parameter increases, the miles per gallon increases, which says us that newer cars have higher & better mileage than older cars. The 'weight' parameter of car says that as the weight of the car increases, the miles per gallon decreases & is inversely proportional. This is true for all values of λ.

---

c)

(c) For $\lambda = 0, \ldots, 50$, predict all 42 test cases. Plot the root mean squared error (RMSE)[3] on the test set as a function of λ—*not* as a function of $df(\lambda)$. What does this figure tell you when choosing λ for this problem (and when choosing between ridge regression and least squares)?

For the plot below we see that as the λ values increases the RMSE values increases as well. This shows us that regularization by ridge regression isn't exactly helping the case & the solution obtained by least square, i.e, at $\lambda=0$, is the best solution in this particular case.

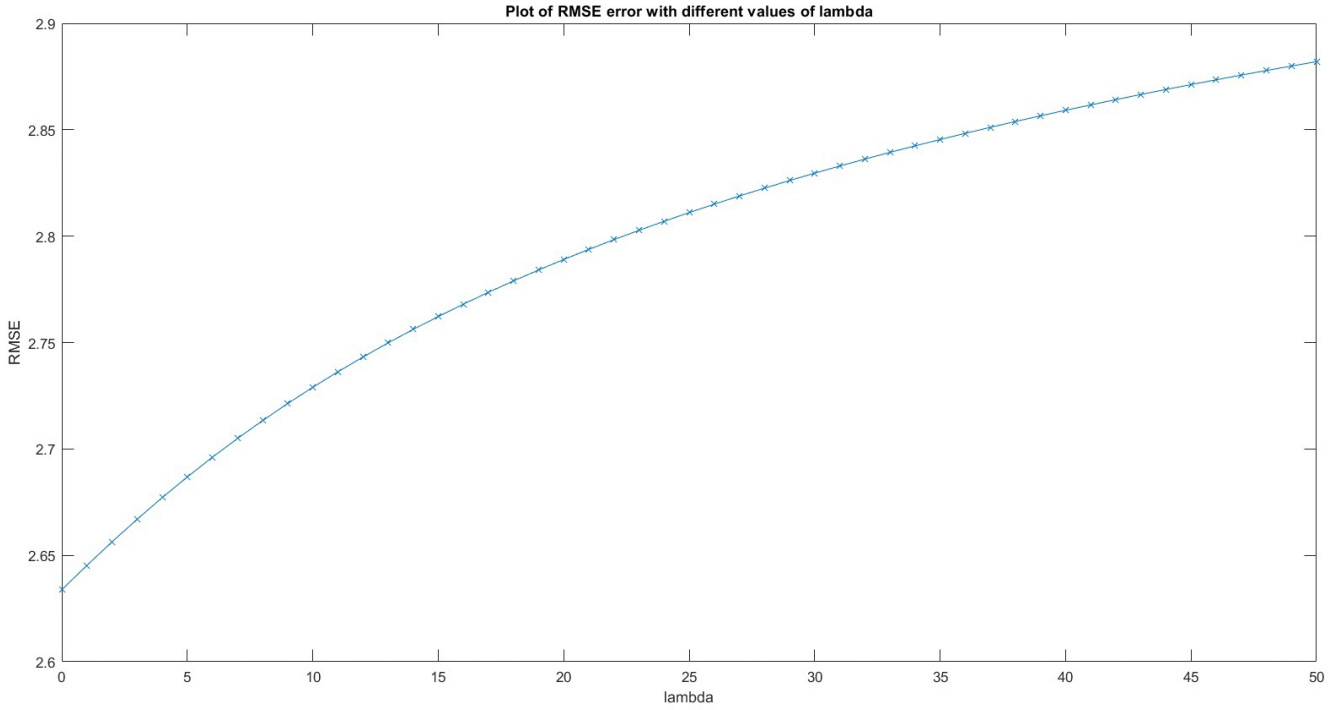

Figure - 2

---

d)

(d) In one figure, plot the test RMSE as a function of $\lambda = 0, \ldots, 100$ for $p = 1, 2, 3$. Based on this plot, which value of p should you choose and why? How does your assessment of the ideal value of λ change for this problem?

Based on the plot, we see that for $p=3$ we get the lowest root mean square error. At $\lambda=51$ & $p=3$ we get the least error of 2.10.

For $p=1$, as λ increases, the RMSE value also increases thus suggesting us that ridge

regression isn't the way to proceed since the error is getting larger compared to least square method. Hence, the best value for λ is 0.

For $p=2$, we see that as λ increases, there's a decrease in the error till $\lambda = 49$, where we get the least error 2.12, before it starts to increase again. This suggests us that if we are to use $p=2$ for model estimation, then the corresponding λ value for the best results is $\lambda = 49$.

Hence, we see that introduction of higher order polynomial terms would require us to regularise the result since $\lambda \neq 0$.