1 CSE 512 Machine Learning - Homework 2

Name: Rohit Bhal
Solar ID: 112073893
NetID email address: rbhal@cs.stonybrook.edu / rohit.bhal@.stonybrook.edu

Q1 Parameter Estimation

1.1 MLE

Ans: (i) We have

$$P(X=k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \{0,1,2 \ldots\}$$

The log likelihood function of X given λ is:

$$\ln\left(P(X=k|\lambda)\right) = \ln\left(\frac{\lambda^k e^{-\lambda}}{k!}\right) \quad (\because \ln ab = \ln(a) + \ln(b))$$

$$= \ln \lambda^k + \ln e^{-\lambda} + - \ln(k!) \quad (\ln e = 1)$$

$$L(X) = k \ln \lambda - \lambda - \ln(k!)$$

(2) MLE for the given λ can be derived as:

$$\frac{\partial L(X)}{\partial \lambda} = 0$$

Now, $L(X=k|\lambda) = \sum_{k=1}^{N} (k \ln \lambda - \lambda - \ln(k!))$

$$\therefore \frac{\partial (L(X))}{\partial \lambda} = \sum_{k=1}^{N} \frac{k}{\lambda} - N = 0$$

$$\lambda = \frac{\sum_{k=1}^{N} k}{N}$$

$$\lambda = \frac{\sum_{k=1}^{N} k}{N}$$

(3) $$\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

Trip	1	2	3	4	5	6	7
Delay in minutes	4	5	3	5	6	9	10

$$\lambda = \frac{(4+5+3+5+6+9+10)}{7}$$

$$= \frac{42}{7} = 6$$

1.2 MAP

(1) $$P(\lambda|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

$$P(\lambda|x=k) = \frac{P(\lambda)\ P(x=k|\lambda)}{P(x=k)}$$

The posterior distribution over λ is:

$$P(\lambda|x=k) = \prod_{i=1}^{n} \frac{P(\lambda)\ P(x=k_i|\lambda)}{P(x=k_i)}$$

$$= \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \frac{(\lambda^{k_i} e^{-\lambda})}{(k_i!)}$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha) \prod_{i=1}^{n}(k_i!)} e^{(-\beta\lambda - \lambda\sum_{i=1}^{n}(1))} \lambda^{((\sum_{i=1}^{n} k_i) + \alpha - 1)}$$

$$P(\lambda | x=k) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{e^{-\lambda(\beta+n)}}{\prod_{i=1}^{n}(k_i!)} \lambda^{((\alpha-1)+\sum_{i=1}^{n} k_i)}$$

(2) MAP estimate for λ is

$$\frac{\partial \ln(P(\lambda | x=k))}{\partial \lambda} = 0$$

$$= \frac{\partial}{\partial \lambda} \ln\left(\frac{\beta^{\alpha}}{\Gamma(\alpha) \prod_{i=1}^{n}(k_i!)} \frac{e^{-\lambda(\beta+n)} \lambda^{((\alpha-1)+\sum_{i=1}^{n} k_i)}}{} \right) = 0$$

$$= \frac{\partial}{\partial \lambda} \left(\ln \beta^{\alpha} + \ln e^{-\lambda(\beta+n)} + \ln \lambda^{\alpha-1+\sum_{i=1}^{n} k_i} - \ln\left(\Gamma(\alpha) \prod_{i=1}^{n}(k_i!)\right) \right) = 0$$

$$= -(\beta+n) + \frac{\left(\alpha-1+\sum_{i=1}^{n} k_i\right)}{\lambda} = 0$$

$$\lambda(\beta+n) = (\alpha-1) + \sum_{i=1}^{n} k_i$$

$$\therefore \lambda = \frac{\left((\alpha-1) + \sum_{i=1}^{n} k_i\right)}{(\beta+n)}$$

1.3 Estimator Bias

(1) $\eta = e^{-2\lambda}$

$\ln \eta = (-2\lambda) \ln e$

$-\frac{1}{2}(\ln \eta) = \lambda$

$$\therefore P(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{\left(-\frac{1}{2}\ln \eta\right)^k e^{-\left(-\frac{1}{2}\log \eta\right)}}{k!}$$

Now, $\ln P(\lambda) = \ln\left(-\frac{1}{2}\ln \eta\right)^k + \ln e^{\frac{1}{2}\log \eta} - \ln k!$

$= k(\ln -\ln 2 - \ln 2) + \frac{1}{2}\ln \eta - \ln k!$

$\frac{\partial \ln P(\lambda)}{\partial \lambda} = 0$

$= k\left(\frac{1}{-\ln \eta}\,\frac{-1}{\eta}\right) + \frac{1}{2\eta} = 0$

$\frac{-1}{2\eta} = \frac{k}{\eta^* \ln \eta}$

$\ln \eta = -2k$

$\eta = e^{-2k} \quad (k = x)$

$\therefore \hat{\eta} = e^{-2k}$

$\Longrightarrow \hat{\eta}$ is the maximum likelihood estimate of λ.

(2) Bias of $\hat{\eta}$ = Expected Value - True Value

$$Ex(\hat{n}) = \sum P(\hat{\eta})\,\hat{\eta}$$

$$\hat{\eta} = e^{-2x}$$

$$\therefore Ex(\hat{n}) = \sum_{i=0}^{\infty} e^{-2x_i}\,\frac{e^{-\lambda}(\lambda)^{x_i}}{(x_i!)}$$

$$= e^{-\lambda} \sum_{i=0}^{\infty} \frac{(e^{-2}\lambda)^{x_i}}{(x_i!)}$$

$$= e^{-\lambda}\, e^{e^{-2}\lambda}$$

$$= e^{(\lambda/e^2 - \lambda)} = e^{\lambda(1/e^2 - 1)}$$

$$\therefore \text{Bias of } \hat{\eta} = e^{\lambda\left(\frac{1}{e^2} - 1\right)} - e^{-2\lambda}$$

$$= \left(e^{-\lambda\left(1 - \frac{1}{e^2}\right)} - e^{-2\lambda}\right)$$

(3) Expectation of an unbiased estimator $f(x)$ is equal to the true mean.

$$\therefore E(f(x)) = e^{-2\lambda}$$

$$\sum_{i=0}^{\infty} f(x)\,\frac{\lambda^{x_i} e^{-\lambda}}{(x_i!)} = e^{-2\lambda}$$

$$\sum_{x=0}^{\infty} f(x) \frac{\lambda^{x_i}}{(x_i!)} = e^{-\lambda}$$

Now, we know that

$$e^{x} = \sum_{k=0}^{\infty} \frac{x^{k}}{k!} = \sum_{k=0}^{\infty} \frac{x^{k}}{k!}$$

$$\sum_{i=0}^{\infty} f(x) \frac{\lambda^{x_i}}{(x_i!)} = \sum_{k=0}^{\infty} \frac{(-\lambda)^{k}}{k!}$$

Now, Both RHS & LHS can be equal iff

$$f(x) = (-1)^{x}$$

$$\Rightarrow f(x) = (-1)^{x}.$$

For $x = 100$, $f(x) = 1$.

$x = 101$, $f(x) = -1$.

Now, we can see that the $f(x)$ is a bad estimator of η for the above x values.

The deviation from true value is very large.

2. Ridge Regression & LOOCV

$$\underset{w,b}{\text{minimize}} \quad \lambda \|w\|^2 + \sum_{i=1}^{n} (w^T x_i + b - y_i)^2$$

Ans (2.1) $\bar{w} = [w; b]$ $\bar{x} = [x; I_n^T]$, $\bar{I} = [I_k, 0_k; 0_k^T, 0]$

$$C = (\bar{x}\bar{x}^T + \lambda \bar{I}), \quad \bar{d} = \bar{x} y$$

To Prove:

$$\bar{w} = C^{-1} d$$

$$L(w) = \lambda \|w\|^2 + \sum_{i=1}^{n} (w^T x_i + b - y_i)^2$$

$$= \lambda w^T w + \sum_{i=1}^{n} (w^T x_i + b - y_i)^2$$

$$\frac{\partial L(w)}{\partial w} = (\lambda + \lambda) w + 2 \sum_{i=1}^{n} (\bar{w}^T \bar{x}_i + b - y_i) \bar{x}_i = 0$$

$$-2\lambda w = 2 \sum_{i=1}^{n} (\bar{w} \bar{x}_i^T + b - y_i) x_i^T$$

$$-\lambda w = w \bar{X}$$

$$L(w) = \lambda \bar{w}^T \bar{w} + \sum_{i=1}^{n} (\bar{w} \bar{x}_i^T + - y_i)^2$$

$$\frac{\partial L(w)}{\partial w} = (\lambda + \lambda) \bar{w} + 2 \sum_{i=1}^{n} (\bar{w} \bar{x}_i^T - y_i) \bar{x}_i = 0$$

$$= 2\lambda \bar{w} + 2 \sum_{i=1}^{n} (\bar{w} \bar{x}_i^T \bar{x}_i - \bar{x}_i y_i) = 0$$

$$= \lambda \bar{w} + \bar{w} \bar{X}^T \bar{X} - \bar{X} y$$

$$\bar{w} (\bar{X}\bar{X}^T + \lambda \bar{I}) = \bar{X} y$$

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

$$\therefore \quad \bar{\omega} = C^{-1}d$$

(2.2) $\quad C = (\bar{X}\bar{X}^T + \lambda\bar{I})$

Now, $\bar{X}\bar{X}^T = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\, x_{ji}$ for $n \times m$ matrix

If we remove x_k row from $\bar{X}$, we get

$$\bar{X}_{-k}\bar{X}_{-k}^T = \sum_{\substack{i=1 \\ i\neq k}}^{n}\sum_{\substack{j=1 \\ j\neq k}}^{m} x_{ij}\, x_{ji}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\, x_{ji} - \sum_{i=k}\sum_{j=k} x_{ij}\, x_{ji}$$

$$= (\bar{X}\bar{X}^T - x_k x_k^T)$$

$$C_{(i)} = (\bar{X}_{-i}\bar{X}_{-i}^T + \lambda\bar{I}) \qquad (x_i \text{ is the data/row removed from } \bar{X})$$

$$= (\bar{X}\bar{X}^T - x_i x_i^T + \lambda\bar{I})$$

$$C_{(i)} = (C - x_i x_i^T)$$

Similarly, $d = \bar{X}y$

$$d_{(i)} = (\bar{X}y - x_i y_i)$$

2.3 $C_{(i)} = (C - x_i x_i^T)$

$C_{(i)}^{-1} = (C - x_i x_i^T)^{-1}$

Now, we know that :

$$(A + UV^T)^{-1} = A^{-1} - \frac{(A^{-1} U V^T A^{-1})}{(1 + V^T A^{-1} U)}$$ (Sherman Morrison formula)

We can see that,

$A = C, \quad U = -x_i, \quad V^T = x_i^T$

$$\therefore (C - x_i x_i^T)^{-1} = C^{-1} - \left(\frac{C^{-1}(-x_i) x_i^T C^{-1}}{1 + x_i^T C^{-1}(-x_i)} \right)$$

$$C_{(i)}^{-1} = C^{-1} + \left(\frac{(C^{-1} x_i x_i^T C^{-1})}{1 - (x_i^T C^{-1} x_i)} \right)$$

2.4 To Prove :

$$\bar{\omega}_{(i)} = \bar{\omega} + (C^{-1} \bar{x}_i) \frac{(-y_i + \bar{x}_i^T \bar{\omega})}{(1 - \bar{x}_i^T C^{-1} \bar{x}_i)}$$

$$\bar{\omega}_{(i)} = C_{(i)}^{-1} d_{(i)} = \left[C^{-1} + \left(\frac{(C^{-1} \bar{x}_i \bar{x}_i^T C^{-1})}{1 - (\bar{x}_i^T C^{-1} \bar{x}_i)} \right) \right] (d - \bar{x}_i y_i)$$

$$= C^{-1}d - C^{-1}\bar{x}_i y_i + \frac{(C^{-1}\bar{x}_i \bar{x}_i^T C^{-1})}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)} d - \frac{(C^{-1}\bar{x}_i \bar{x}_i^T C^{-1})(\bar{x}_i y_i)}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)}$$

$$= \bar{\omega} + C^{-1}\bar{x}_i \left(-y_i + \frac{(C^{-1}\bar{x}_i \bar{x}_i^T C^{-1})d}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)} - \frac{(\bar{x}_i^T C^{-1}\bar{x}_i y_i)}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)} \right)$$

$$= \bar{\omega} + C^{-1}\bar{x}_i \left(\frac{-y_i + y_i \bar{x}_i^T C^{-1}\bar{x}_i + \bar{x}_i^T C^{-1}d - \bar{x}_i^T C^{-1}\bar{x}_i y_i}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)} \right)$$

$$\bar{\omega}_{(i)} = \bar{\omega} + C^{-1}\bar{x}_i \frac{(-y_i + \bar{x}_i^T \bar{\omega})}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)}$$

Hence proved.

2.5 To Prove

$$\omega_{(i)}^T \bar{x}_i - y_i = \frac{(\bar{\omega}^T x_i - y_i)}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)}$$

$$\omega_{(i)}^T \bar{x}_i - y_i = \left[\bar{\omega} + C^{-1}\bar{x}_i \frac{(-y_i + \bar{x}_i^T \bar{\omega})}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)} \right] \bar{x}_i - y_i$$

$$= \left[\frac{\bar{\omega}\bar{x}_i - \bar{\omega}\bar{x}_i^T C^{-1}\bar{x}_i \bar{x}_i + (-C^{-1}\bar{x}_i y_i)\bar{x}_i + C^{-1}x_i \bar{x}_i^T \bar{\omega}\bar{x}_i}{(1-\bar{x}_i^T C^{-1}\bar{x}_i)} \right] - y_i$$

$$= \frac{\left(\bar{w}^T \bar{x}_i - \bar{w}\bar{x}_i^T C^{-1} \bar{x}_i \bar{x}_i - C^{-1}\bar{x}_i y_i \bar{x}_i + C^{-1}\bar{x}_i \bar{x}_i^T \bar{w}\bar{x}_i - y_i + \bar{x}_i^T C^{-1}\bar{x}_i y_i\right)}{\left(1 - \bar{x}_i^T C^{-1}\bar{x}_i\right)}$$

$$= \frac{\left(\bar{w}^T \bar{x}_i - y_i\right)}{\left(1 - \bar{x}_i^T C^{-1}\bar{x}_i\right)}$$

Hence proved

2.6

$$\text{LOOCV} : \sum_{i=1}^{n} \left(\bar{w}_{(i)}^T \bar{x}_i - y_i\right)^2$$

$$\left(w_{(i)}^T \bar{x}_i - y_i\right) = \frac{\left(\bar{w}^T \bar{x}_i - y_i\right)}{\left(1 - \bar{x}_i^T C^{-1}\bar{x}_i\right)}$$

C is a matrix of (k, k)

To calculate inverse of C, we need $O(k^3)$ time.

Also, to calculate $\bar{w}^T x_i$, we need $O(k)$ time.

To calculate $\bar{x}_i^T C^{-1} \bar{x}_i$, we need $O(k^2)$ time, but this is calculated n times.

Maximum Time complexity = $\text{Max}\left(O(k^3), O(nk^2)\right)$

If we use the naive way of calculating LOOCV error, we will be calculating C^{-1} for n time.

Time Complexity = $O(nk^3)$.

3. Programming

3.2

2. $\lambda = 0.77582246$ acheives the best performance

The objective value for this λ is

21217.73885524

The regularization term ($\lambda \|w\|_2$) is

10127.32674793

The RMSE for the above λ is using LOOCV

≈ 1.95950942

The sum of square of errors is:

11090.4121

The root mean square error for the data:

1.053110

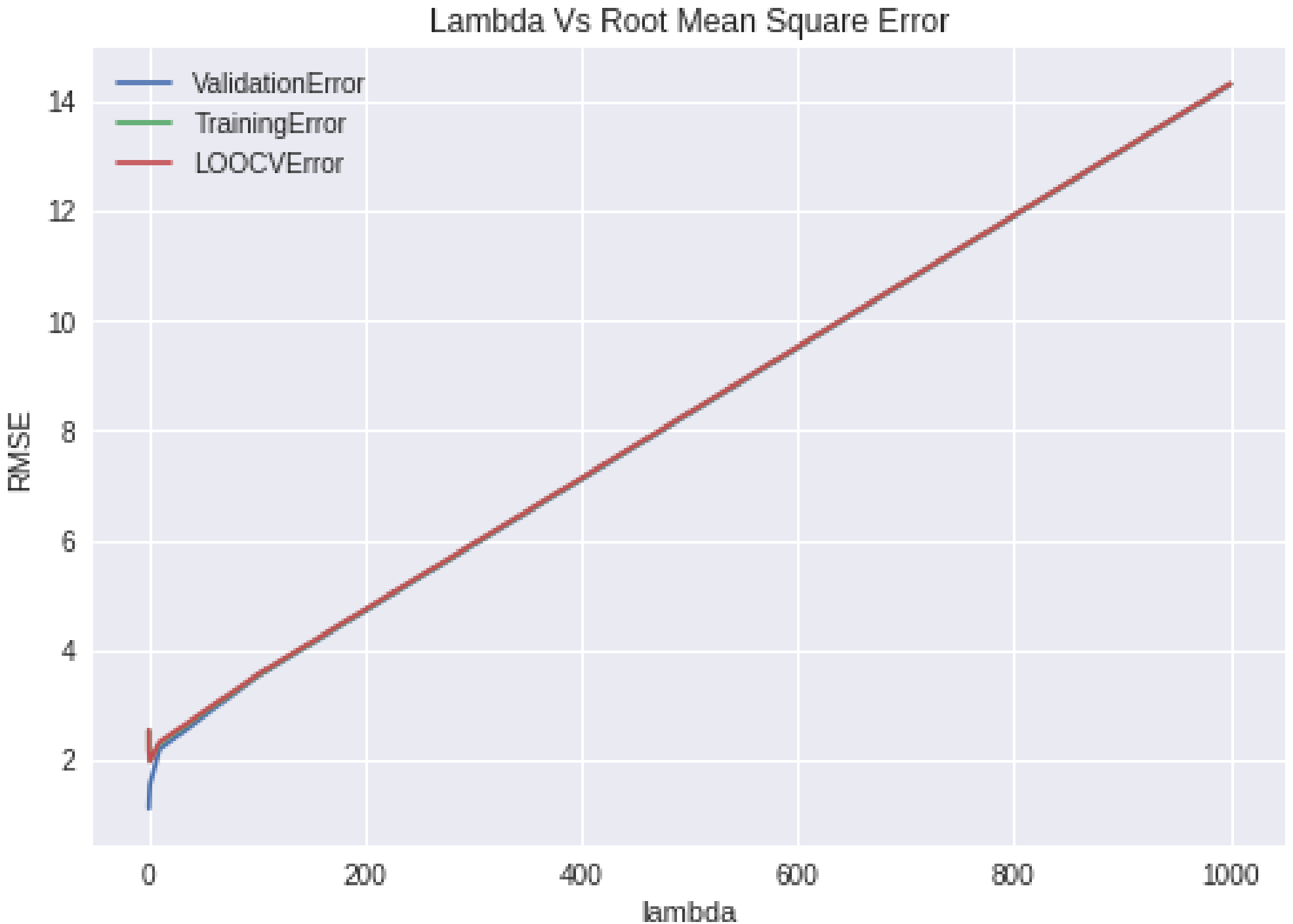
3. The top 10 most important features & the top 10 least important features are very hard to define.
If I have to choose, I would choose the features according to the significance it has to the obj. function.

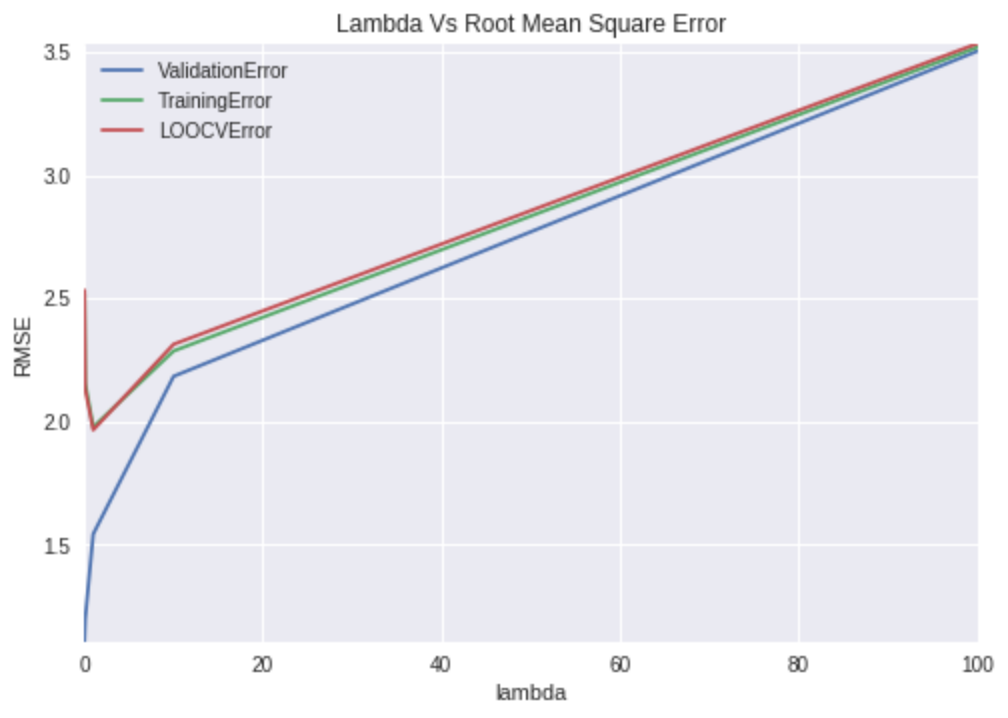
The top 10 most important features:

1. Age Years
2. Wine Price
3. Rare
4. Delightful
5. Pleasing
6. Area
7. Region
8. Versatile
9. Affordable
10. Smells

Top 10 least impt. features:

1. tar
2. Ava
3. pineapple pears
4. Grape
5. Leaner
6. Sonoma County
7. Tiny
8. Oil
9. Home
10. Rocks


Lambda Vs Root Mean Square Error
ValidationError
TrainingError
LOOCVError
RMSE
14
12
10
8
6
4
2
0
200
400
600
800
1000
lambda


Lambda Vs Root Mean Square Error
ValidationError
TrainingError
LOOCVError
RMSE
lambda
3.5
3.0
2.5
2.0
1.5
0
20
40
60
80
100