CSE512 Fall 2018 - Machine Learning - Homework 2

Your Name: Astitv Nagpal

Solar ID: 112008011

NetID email address: anagpal@cs.stonybrook.edu

Names of people whom you discussed the homework with: Ayush Garg

QUESTION - 1 - PARAMETER ESTIMATION

(1.1) - MLE

① Given, $X \sim$ Poisson Distribution with parameter λ.

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

To find the log-likelihood of the function, we take log on both sides & summation for the function.

$$\log(F(x)) = \sum_{k=1}^{\infty} \log\left(\frac{\lambda^k e^{-\lambda}}{k!}\right)$$

$$= \sum_{\text{all } k} \left[\log(\lambda^k) + \log(e^{-\lambda}) - \log(k!)\right]$$

$$= \sum_{\text{all } k} \left[k \cdot \log(\lambda) - \lambda - \log(k!)\right]$$

$$= \sum_{\text{all } k} k \cdot \log(\lambda) - \sum_{\text{all } k} \lambda - \sum_{\text{all } k} \log(k!)$$

$$\boxed{\text{log-Likelihood} = (\log \lambda) \sum_{\text{all } k} k - n\lambda - \sum_{\text{all } k} \log(k!)}$$ answer.

② Computing the MLE.

From our previous answer, we got the log-likelihood func.

To calculate the MLE we can take the partial differential w.r.t. 'λ' & solve for MLE.

$$= \frac{\partial \left[\log \lambda \sum_{all\,k} k - n\lambda - \sum_{all\,k} \log(k!) \right]}{\partial \lambda} = 0$$

$$= \frac{1}{\lambda} \sum_{all\,k} k - n - 0 = 0$$

$$\therefore \boxed{\hat{\lambda}_{MLE} = \frac{\sum_{all\,k} k}{n}}$$

③ Computing the MLE for λ using the observed X.

→ We have $\hat{\lambda}_{MLE} = \frac{\sum_{all\,k} k}{n}$, &

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

$\therefore$ we get $\hat{\lambda}_{MLE} = \frac{(4+5+3+5+6+9+10)}{7} = \frac{42}{7} = 6$

$$\therefore \boxed{\hat{\lambda}_{MLE} = 6}.$$

(1.2) - MAP

① To compute the posterior distribution over λ.

From previous examples we have.

$$\hat{\lambda}_{MLE} = \frac{\sum_{all\,k} k}{n} \quad \& \quad P(X=k/\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\therefore \text{likelihood function} \rightarrow \prod_{i=1}^{n} \frac{\lambda^{k_i} \cdot e^{-\lambda}}{k_i!} = \frac{\lambda^{\sum_{i=1}^{n} k_i} \cdot e^{-\sum \lambda}}{\prod_{i=1}^{n} k_i!}$$

∴ Likelihood function is

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

→ Given to us prior follows Gamma Distribution.

→ Using Bayes' formula we get,

$$P(x \mid Data) = \frac{P(Data \mid x) \cdot P(X = k)}{P(Data)}$$; For some Data

$$= \frac{(Likelihood)(Prior)}{P(Data)}$$

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

$$\boxed{\therefore \text{Posterior Dist.} = \frac{\lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} \cdot e^{-\lambda(n+\beta)} \cdot \beta^{\alpha}}{\prod_{i=1}^{n} x_i! \; \Gamma(\alpha)}}$$

Scanned by CamScanner

2 To derive MAP estimate of λ over Posterior Dist.

→ We have

$$\text{Post. Dist} = \frac{\lambda^{\sum x_i + \alpha - 1} \cdot e^{-\lambda(n+\beta)} \cdot \beta^{\alpha}}{\prod_{i=1}^{n} x_i! \cdot \Gamma(\alpha)}$$

As per the hint instead of solving the complete equation we observe that the distribution depends on :

$$\propto \lambda^{\sum x_i + \alpha - 1} \cdot e^{-\lambda(n+\beta)} \quad \text{—— (i)}$$

∴ Solving for eq (i) & taking log we get.

$$= \left(\sum x_i + \alpha - 1\right) \log \lambda + (-\lambda)(n+\beta)$$

Now taking derivative & equating to zero we get.

$$= \frac{\partial \left[\left(\sum x_i + \alpha - 1\right)(\log \lambda) + (-\lambda)(n+\beta)\right]}{\partial \lambda} = 0$$

$$= \frac{\sum x_i + \alpha - 1}{\lambda} - (n+\beta) = 0$$

$$\therefore \boxed{\hat{\lambda}_{MAP} = \frac{\sum_{k=1}^{n} x_i + \alpha - 1}{n+\beta}}$$

(1.3) → ESTIMATOR BIAS

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

Given that $X \sim$ Poisson(λ).

∴ we need to calculate for.

$\frac{\partial(P(X=\lambda))}{\partial \hat{\eta}}$; this can be re-written as $\frac{\frac{\partial(P(X=\lambda))}{\partial(\lambda)}}{\frac{\partial(\eta)}{\partial(\lambda)}}$

$$= \frac{\frac{\partial\left[\frac{\lambda^k e^{-\lambda}}{k!}\right]}{\partial \lambda}}{\frac{\partial(e^{-2\lambda})}{\partial \lambda}} = \frac{\frac{k\lambda^{k-1} \cdot e^{-\lambda} - \lambda^k \cdot e^{-\lambda}}{k!}}{-2e^{-2\lambda}}$$

$$= \frac{\lambda^k e^{-\lambda}\left[\frac{k}{\lambda} - 1\right]}{k!\cdot(-2)(e^{-2\lambda})} = \frac{\lambda^k e^{\lambda}\left[\frac{k}{\lambda} - 1\right]}{-2\,k!} = 0$$

$$\therefore \frac{k}{\lambda} - 1 = 0 \Rightarrow \boxed{k = \lambda}.$$

Hence, $\hat{\eta} = e^{-2X}$ for $X = k$ is the MLE of η.

(2) To show that bias of $\hat{\eta}$ is $e^{-(1-1/e^2)\lambda} - e^{-2\lambda}$.

We know, bias $= E(\hat{\theta}(x)) - \theta(x)$.

Here $X \sim$ poisson distribution.

we know → $E[f(x)] = \sum_{x=1}^{n} f(x) \cdot P(x)$

∴ we get.

$$\text{Bias} = \left[\sum_{k=0}^{\infty} e^{-2k} \cdot \frac{\lambda^{k} e^{-\lambda}}{k!} \right] - e^{-2\lambda}$$

$$= e^{-\lambda} \left(\sum_{\text{all } k} \frac{e^{-2k} \cdot \lambda^{k}}{k!} \right) - e^{-2\lambda}$$

$$= e^{-\lambda} \left(\sum_{\text{all } k} \frac{(\lambda/e^{2})^{k}}{k!} \right) - e^{-2\lambda} \longrightarrow (i)$$

We know, $e^{x} = \sum_{\text{all } k} \frac{x^{k}}{k!} \longrightarrow (ii)$

In eq (i) let $\alpha = \lambda/e^{2}$ ∴ we get.

$$= e^{-\lambda} \left(\sum_{\text{all } k} \frac{\alpha^{k}}{k!} \right) - e^{-2\lambda}$$

substituting eq^n (ii) here we get.

$$= e^{-\lambda} \cdot e^{\alpha} - e^{-2\lambda}$$

∴ now replacing α with λ/e^{2}

$$= e^{-\lambda} \cdot e^{\lambda/e^{2}} - e^{-2\lambda} \Rightarrow e^{\lambda/e^{2} - \lambda} - e^{-2\lambda}$$

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

Hence Proved

3) To prove that $(-1)^x$ is unbiased.

We know Bias $= E[\hat{p}(n)] - p(n)$

If Bias $= 0$, that means it is unbiased.

$$\Rightarrow \sum_{x=0}^{\infty} g(n) \frac{\lambda^x e^{-\lambda}}{\lambda!} - e^{-2\lambda}$$

given $g(n) = (-1)^x$ $\therefore$ we get $\sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x e^{-\lambda}}{\lambda!} - e^{-2\lambda}$

$$\Rightarrow e^{-\lambda}\left(\sum_{\text{all } x} \frac{(-1)^x \lambda^x}{x!}\right) - e^{-2\lambda}$$

$$\Rightarrow e^{-\lambda}\underbrace{\left[1 - \frac{\lambda}{1!} + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \cdots\right]}_{\downarrow} - e^{-2\lambda}$$

We know $\Rightarrow e^{-\lambda}$ $\therefore$ we get.

$$e^{-\lambda} \cdot e^{-\lambda} - e^{-2\lambda} \Rightarrow e^{-2\lambda} - e^{-2\lambda} = \boxed{0 = Bias}.$$

Hence it is unbiased.

$\Rightarrow (-1)^x$ is not a good estimator as

(i) It has a factor of $(-1)^x$. So for odd values of X it will give negative values & we can have only positive values here.

(ii) As the value of X increases our value nears 0. But here we get for even X values as 1. Which is the other extreme side.

Hence, it is not a good estimator in this case.

QUESTION - 2 - RIDGE REGRESSION & LOOCV

(2.1) Given, $C = \bar{X}\bar{X}^T + \lambda \bar{I}$ | $d = \bar{X} y$

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

We have to minimize $\lambda \|W\|^2 + \sum_{i=1}^{n} (W^T x_i + b - y_i)^2$ — (i)

We know, b will be added to $W^T x_i$ &

$\|W\|^2 = W^T W$ & $\sum_{i=1}^{n} [A^T B_i - y_i]^2 = [A^T B - Y]^T [A^T B - Y]$

∴ using the above properties in eqⁿ (i) we get,

$= \lambda W^T W + [\bar{X}^T W - Y]^T [\bar{X}^T W - Y] \Rightarrow \lambda W^T W + W^T \bar{X} \bar{X}^T W - W^T \bar{X} Y - \bar{Y} \bar{X}^T X + Y^T Y$

To minimize we will take partial derivative w.r.t. W.

We know,

$\frac{\partial (X^T A X)}{\partial X} = 2AX$ & $\frac{\partial (XA)}{\partial X} = A$, using the propert.

In ~~$\partial (W^T$~~ Our eqⁿ, $W^T \bar{X} Y$ & $\bar{Y} \bar{X}^T X$ are scalar values & are equal. hence we can take one as a common.

∴ The eqⁿ becomes:

$W^T \bar{X} \bar{X}^T W - 2 W^T \bar{X} Y + Y^T Y$ — (ii)

Taking Derivative of eqⁿ (ii) we get.

$$\frac{\partial (W^T \bar{X} \bar{X}^T W - 2 W^T \bar{X} Y + Y^T Y)}{\partial (W)} = 2 \bar{X} \bar{X}^T \bar{W} - 2 \bar{X} Y + 0 = 0$$

$$= \cancel{2} \bar{X} \bar{X}^T \bar{W} - \cancel{2} \bar{X} Y = 0$$

Taking W common we get,

$$[\bar{X}\bar{X}^T + \lambda \bar{I}]\bar{W} - \bar{X}Y = 0$$

$$\therefore \bar{W} = [\bar{X}\bar{X}^T + \lambda \bar{I}]^{-1} \cdot \bar{X}Y$$

$$\boxed{\bar{W} = C^{-1} \cdot d} \rightarrow \text{Hence Proved}$$

(2.2). Express C_i in terms of C & x_i . d_i in terms of d & x_i.

We have, $C = \bar{X}\bar{X}^T + \lambda \bar{I}$; We know $\bar{X}\bar{X}^T = \sum_{\text{all } i} x_i x_i^T$

$$\therefore C_i = \bar{X}_i \bar{X}_i^T + \lambda \bar{I} \quad \text{—— (i)}$$

and $\bar{X}_i \bar{X}_i^T = \bar{X}\bar{X}^T - x_i x_i^T$; replacing in eqn (i)

$C_i = \bar{X}\bar{X}^T - x_i x_i^T + \lambda \bar{I}$; we have $C = \bar{X}\bar{X}^T + \lambda \bar{I}$

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

Similarly $d = \bar{X}y$

$d_i = \bar{x}_i y_i$ and $\bar{x}_i y_i = \bar{X}Y - x_i y_i$, replacing it

$$d_i = \bar{X}Y - x_i y_i$$

$$\boxed{d_i = d - x_i y_i}$$

(2.3) Express C_i^{-1} in terms of C^{-1} & x_i.

From Previous example we got → $C_i = C - \bar{X}_i \bar{x}_i^T$

By comparing with Sherman-Morisson's equation we get;

$A = C$

$u = -x_i$

$v^T = x_i^T$

∴ Applying them in the formula we get:

$$C_i^{-1} = C^{-1} - \frac{C^{-1}(-x_i)(x_i)^T C^{-1}}{1 + (x_i)^T C^{-1}(-x_i)}$$

$$\boxed{C_i^{-1} = C^{-1} + \frac{C^{-1} x_i (x_i)^T C^{-1}}{1 - (x_i)^T C^{-1}(x_i)}}$$

(2.4). We have $\bar{w} = C^{-1}d$

∴ $\bar{w}_i = (C_i^{-1})(d_i)$, from previous examples we have values of C_i^{-1} & d_i, by substituting those values, we get:

$$= \left[C^{-1} + \frac{C^{-1} x_i x_i^T C^{-1}}{1 - x_i^T C^{-1} x_i}\right]\left[d - x_i y_i\right]$$

$$= C^{-1}d - C^{-1}x_i y_i + \frac{C^{-1}x_i x_i^T C^{-1} d - C^{-1}x_i x_i^T C^{-1} x_i y_i}{1 - x_i^T C^{-1} x_i}$$

$$= \bar{w} - \frac{(C^{-1}x_i y_i)(1 - x_i^T C^{-1} x_i) + C^{-1}x_i x_i^T C^{-1} d - C^{-1}x_i x_i^T C^{-1} x_i y_i}{1 - \cancel{X}_i^T C^{-1} x_i}$$

$$= \bar{w} - \frac{C^{-1}x_i y_i + \cancel{C^{-1}x_i y_i x_i^T C^{-1} x_i} + C^{-1}x_i x_i^T C^{-1} d - \cancel{C^{-1} x_i x_i^T C^{-1} x_i y_i}}{1 - x_i^T C^{-1} x_i}$$

Here we can see that $C^{-1}x_i y_i x_i^T C^{-1} x_i$ & $C^{-1} x_i x_i^T C^{-1} x_i y_i$ have the same dimensions $(k \times 1)$ ~~hence~~ & all the values are jumbled but same. Hence, they can be cancelled.

$$= \bar{w} - \frac{C^{-1} x_i y_i + C^{-1} x_i x_i^T C^{-1} d}{1 - x_i^T C^{-1} x_i}$$

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

$$= \boxed{\bar{w} + (C^{-1} x_i) \cdot \frac{-y_i + x_i^T \bar{w}}{1 - x_i^T C^{-1} x_i}} \rightarrow \text{Hence Proved.}$$

2.5. We have value of $\bar{w}_i$ from the previous examples, hence RHS becomes:

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

We know $(A+B)^T = A^T + B^T$ & $(AB)^T = B^T A^T$, using them.

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

Here, Since the denominator is a Scalar value hence transpose on it doesn't make sense.

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

$$= \cancel{\bar{w}^T + (-y_i^T + x_i^T \bar{w}) C^{-1} x_i}$$

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

$$= \bar{w}^T x_i + \frac{(-y_i^T + \bar{w}^T x_i)(x_i^T C^{-1^T} x_i)}{1 - x_i^T C^{-1} x_i} - y_i$$

Since C is a square matrix we have $C^{-1^T} = C^{-1}$ order.

$$= \frac{\bar{w}^T x_i - \bar{w}^T x_i \cancel{x_i^T} C^{-1} x_i - y_i^T x_i^T C^{-1} x_i + \bar{w}^T \cancel{x_i x_i^T} C^{-1} x_i}{1 - x_i^T C^{-1} x_i} - y_i$$

$$= \frac{\bar{w}^T x_i - y_i^T \cancel{x_i^T} C^{-1} x_i - y_i + \cancel{y_i x_i^T} C^{-1} x_i}{1 - x_i^T C^{-1} x_i} = \boxed{\frac{\bar{w}^T x_i - y_i}{1 - w_i^T C^{-1} x_i}}$$

Hence Proved

(2.6) To find the time complexity we will see all the running times.

- For Matrix inversion $O(k^3)$
- For Matrix dot product $O(k^2)$
- To calculate LOOCV we are running linearly $O(n)$
- To calculate for 'm' values of lamda we are running $O(m)$.

Now, we are running m times & computing inverse & dot product & LOOCV.

$\therefore$ We have : $O[m(k^3 + k^2 + n^2 + n)]$

$= O[mk^3 + mk^2 + mn^2 + mn]$

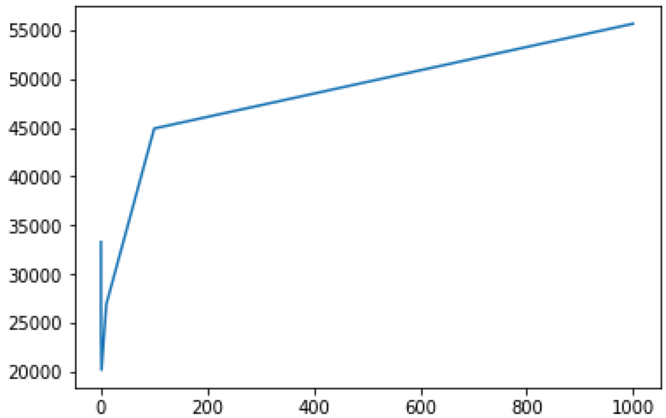
Here we can omit mk^2, mn^2, mn as they are less that mk^3.

Hence, running time complexity is $\underline{O(mk^3)}$.

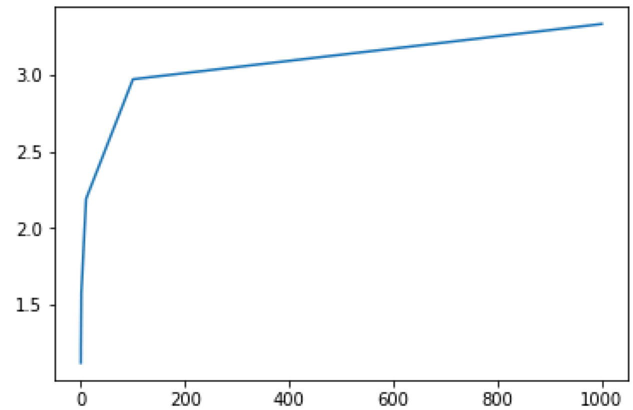
Scanned by CamScanner

Question 3

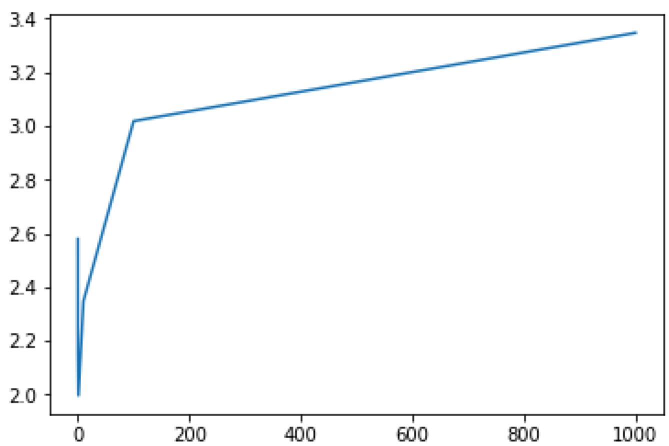
Lamda vs LOOCV Errors



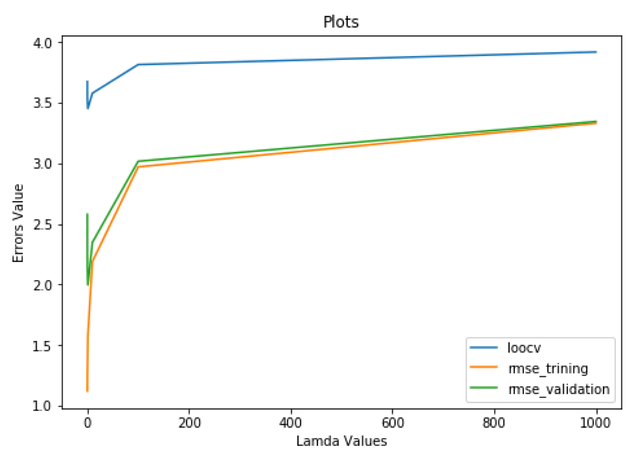
Lamda vs RMSE_Training Errors



Lamda vs RMSE_Validation Errors



Plotting LOOCV error values as 1/8th of the original to show on the



Minimum Lamda = 1

Objective Function Value = 17200.94056872298

LOOCV Error Training Data = 20189.93000694269

RMSE Error Training Data = 1.5780360753182014

Value of Regularization term = 0.0028123108309766136

The Features make sense as the least important are almost all near to zero hence are not impacting the final answer. Whereas the most important ones are having high values and hence are impacting the final outcome.

-> Final error

```
10 Most Important Features

1 : W-514 at index-513
2 : W-2908 at index-2907
3 : W-2954 at index-2953
4 : W-585 at index-584
5 : W-2347 at index-2346
6 : W-135 at index-134
7 : W-1728 at index-1727
8 : W-2305 at index-2304
9 : W-1845 at index-1844
10 : W-2276 at index-2275

10 Least Important Features

1 : W-2367 at index- 2366
2 : W-2431 at index- 2430
3 : W-2611 at index- 2610
4 : W-1754 at index- 1753
5 : W-1961 at index- 1960
6 : W-1315 at index- 1314
7 : W-2582 at index- 2581
8 : W-2384 at index- 2383
9 : W-1063 at index- 1062
10 : W-2005 at index- 2004
```

112008011_1.csv a few seconds ago by Astitv Nagpal 112008011	1.99726