CSE512 Fall 2018 - Machine Learning - Homework 2

Name: Manideep Attanti
Solar Id: 112028167
Netid email: manideep.attanti@stonybrook.edu

Q1) 1.1.1, 1.1.2, 1.1.3

Q1)

(1.1.1) $P(x=k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$

$x = (x_1, x_2, \ldots, x_n)$

$$P(x \mid \lambda) = P(x_1 \mid \lambda) \cdot P(x_2 \mid \lambda) \cdots P(x_n \mid \lambda)$$

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

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

$$\log P(x \mid \lambda) = \log \left(e^{-n\lambda} \cdot \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \right)$$

$$= -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \log \lambda - \log\left(\prod_{i=1}^{n} x_i!\right)$$

$$= \left(\sum_{i=1}^{n} x_i\right) \log \lambda - n\lambda - \sum_{i=1}^{n} (\log x_i!)$$

(1.1.2) MLE for $\lambda = \hat{\lambda}$

$$\frac{\partial \log P(x \mid \lambda)}{\partial \lambda} = 0$$

$$\Rightarrow \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(1.1.3) Observed $x = (4, 5, 3, 5, 6, 9, 10)$

from 1.1.2 $\lambda_{MLE} = \frac{4+5+3+5+6+9+10}{7}$

$= 6$

Q1) 1.2.1

1.2.1) $$P(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

posterior distribution over λ

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

$P(x)$ is independent of λ

$$P(x \mid \lambda) = \prod_{i=1}^{n} P(x_i \mid \lambda) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right)$$

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

$$P(\lambda \mid x) = \frac{\dfrac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}}{P(x)}$$

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

$$P(x) = \int_0^{\infty} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \frac{\hat{\lambda}^{\sum_{i=1}^{n} x_i + \alpha - 1} \cdot e^{-\hat{\lambda}(\beta+n)}}{\prod_{i=1}^{n} x_i!} \, d\hat{\lambda}$$

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

Q1) 1.2.2

(1.2.2) MAP of λ

$$\log p(\lambda|x) = \alpha \log \beta - \lambda(n+\beta) + \left(\sum_{i=1}^{n} x_i + \alpha - 1\right) \log \lambda$$
$$- \log(P(x)) - \log \Gamma(\alpha)$$
$$- \log\left(\prod_{i=1}^{n} x_i!\right)$$

$$\frac{\partial \log P(\lambda|x)}{\partial \lambda} = -(n+\beta) + \frac{\left(\sum_{i=1}^{n} x_i + \alpha - 1\right)}{\lambda} = 0$$

$$\Rightarrow \boxed{\lambda = \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{n+\beta}}$$

Q1) 1.3.1

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

$$P(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$$\ln \eta = -2\lambda$$

$$\lambda = -\frac{\ln \eta}{2}$$

$$\log P(x|\lambda) = \log\left(\frac{\lambda^x}{x!} e^{-\lambda}\right)$$

$$= -\lambda + x \log \lambda - \log(x!)$$

$$\log P(x|\eta) = -\left(-\frac{\ln \eta}{2}\right) + x \log\left(-\frac{\ln \eta}{2}\right) - \log(x!)$$

$$= \frac{\ln \eta}{2} + x \log(-\ln \eta) - \log 2 - \log(x!)$$

$$\frac{\partial \log P(x|\eta)}{\partial \eta} = \frac{1}{2\eta} + x\left(\frac{-1}{\ln \eta}\right)\left(\frac{-1}{\eta}\right) = 0$$

$$\Rightarrow \frac{1}{\eta}\left(\frac{1}{2} + \frac{x}{\ln \eta}\right) = 0$$

$$\Rightarrow 2x + \ln \eta = 0$$

$$\Rightarrow \ln \eta = -2x$$

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

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

Q1) 1.3.2, 1.3.3

(1.3.2) Bias of $\hat{\eta} = E(\hat{\eta}) - \eta$ where $\eta = e^{-2\lambda}$

$$\cancel{E(\hat{\eta}) = \sum_{k=0}^{\infty} e^{-2X_k} \cdot \frac{\lambda^{X_k}}{X_k!} e^{-\lambda}}$$

$$E(\hat{\eta}) = \sum_{X=0}^{\infty} e^{-2X} \cdot \frac{\lambda^X}{X!} \cdot e^{-\lambda}$$

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

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

$$E(\hat{\eta}) = e^{-\lambda} \cdot e^{\lambda/e^2}$$

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

$$\text{Bias of } \hat{\eta} = e^{-\lambda(1 - 1/e^2)} - e^{-2\lambda}$$

(1.3.3) $\hat{\eta} = (-1)^X \Rightarrow E(\hat{\eta}) = \sum_{X=0}^{\infty} (-1)^X \frac{\lambda^X}{X!} e^{-\lambda}$

$$= e^{-\lambda} \left(1 - \frac{\lambda}{1!} + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} \dots \right)$$

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

$$\therefore \text{Bias of } \hat{\eta} = e^{-2\lambda} - e^{-2\lambda} = 0$$

$\therefore (-1)^X$ is an unbiased estimate of η

It's a bad estimator because it can take only values (−1) or (1) where as η is originally an exponential function in λ.

Q2) 2.1

Q2 (2.1) $\bar{w} = [w; b]$, $\bar{x} = [x; 1_n^T]$

$\bar{I} = [I_k, 0_k; 0_k^T, 1]$

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

$\lambda \|w\|^2 = \lambda(\|\bar{w}\|^2 - b^2)$

$= \bar{w}^T \lambda \bar{I} \bar{w} - \lambda b^2$

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

$= (\bar{w}^T \bar{x} - y^T)(\bar{x}^T \bar{w} - y)$

$= \bar{w}^T \bar{x} \bar{x}^T \bar{w} - \bar{w}^T \bar{x} y - y^T \bar{x}^T \bar{w} - y^T y$

$= \bar{w}^T \bar{x} \bar{x}^T \bar{w} - 2 y^T \bar{x}^T \bar{w} - y^T y$

$\therefore$ error $\equiv \bar{w}^T \lambda \bar{I} \bar{w} - \lambda b^2 + \bar{w}^T \bar{x} \bar{x}^T \bar{w} - 2 y^T \bar{x}^T \bar{w} - y^T y$

$$\frac{\partial \, \text{error}(\bar{w})}{\partial \bar{w}} = 2\lambda \bar{I} \bar{w} + 2 \bar{x} \bar{x}^T \bar{w} - 2 \bar{x} y$$

(by using the identities $\frac{\partial (a^T x)}{\partial x} = a$ & $\frac{\partial (x^T A x)}{\partial x} = (A + A^T) x$

$\Rightarrow (\lambda \bar{I} + \bar{x} \bar{x}^T) \bar{w} - \bar{x} y = 0$

$\Rightarrow C \bar{w} - d = 0$ (from question)

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

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

Q2) 2.2

(2.2) $C = \overline{X}\,\overline{X}^T + \lambda \overline{I}$

$C_{(i)} = \overline{X}_{(i)} \overline{X}_{(i)}^T + \lambda \overline{I}$

$$\overline{X} = \begin{bmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{n1} \\ \vdots & & \vdots & & \vdots \\ x_{1k} & \cdots & x_{ik} & \cdots & x_{nk} \\ 1 & \cdots & 1 & \cdots & 1 \end{bmatrix} \quad (k+1)\times n$$

$$\overline{X}^T = \begin{bmatrix} x_{11} & \cdots & x_{1k} & 1 \\ \vdots & & & \\ x_{i1} & \cdots & x_{ik} & 1 \\ \vdots & & & \\ x_{n1} & \cdots & x_{nk} & 1 \end{bmatrix} \quad n\times(k+1)$$

As we can see in $\overline{X}\,\overline{X}^T$ each of $(k+1)\times(k+1)$ matrix's terms are effected by marked column & row, it can be seen that

$$\overline{X}\,\overline{X}^T = \overline{X_{(i)}}\,\overline{X_{(i)}}^T + \overline{x_i}\,\overline{x_i}^T$$

$$\therefore \quad C_{(i)} = C - \overline{x_i}\,\overline{x_i}^T$$

$d = \overline{X} y \qquad d_{(i)} = \overline{X}_{(i)} y_{(i)}$

we can see for the resultant d each element of $\overline{x_i}$ multiplies with $y_{(i)}$. hence

$$~~d=~~ \; \overline{X} y = \overline{X}_{(i)} y_{(i)} + \overline{x_i}\, y_i$$

where y_i is (1×1) just a value of i^{th} label in vector ~~vector~~

$$\Rightarrow \quad d_{(i)} = d - \overline{x_i}\, y_i$$

Q2) 2.3 and 2.4

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

$C_{(i)}^{-1}$ from equation $(A+uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$

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

$$= C^{-1} + \frac{C^{-1}\bar{x_i}\,\bar{x_i}^T C^{-1}}{1-\bar{x_i}^T C^{-1}\bar{x_i}}$$

(2.4) $\bar{w}(i) = C_{(i)}^{-1} d(i)$

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

where y_i is a (1×1) matrix containing 1 value

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

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

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

Q2) 2.5

(2.5) $\overline{w}_{(i)}^T \overline{x_i} - y_i$

$= \overline{x_i}^T (w_{(i)}) - y_i$

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

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

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

$$= \frac{\overline{w}^T \overline{x_i} - y_i}{1 - \overline{x_i}^T C^{-1} \overline{x_i}}$$

We used $\overline{w}^T \overline{x_i}$ & $\overline{x_i}^T \overline{w}$ as same because they yield same value which is (1×1) matrix or a scalar value.

Q2) 2.6

(2.6) Complexity of multiplying to matrices of type $(m \times n)$ and $(n \times p)$ ~~has~~ is $O(mnp)$

Complexity of calculating $\overline{w} = O(k^3)$

$\because$ all matrix ~~po~~ multiplications lead to a maximum of $O(nk)$ complexity

$\because$ one dimension is '1' in all multiplications

$\therefore$ For usual way without using 2.5 the complexity will be $O(nk^3)$.

This is because for all n, ~~x_i~~ removals w_i must be calculated.

Using (2.5) we need to calculate $\overline{w}$ and C^{-1} only once. Once calculated based on multiplications in (2.5) maximum complexity will be $O(k^2)$ for one error

$\therefore$ for n elements it will be $O(nk^2)$

$\therefore$ Overall complexity $\equiv O(\max(k^3, nk^2))$

~~If n is very large than~~

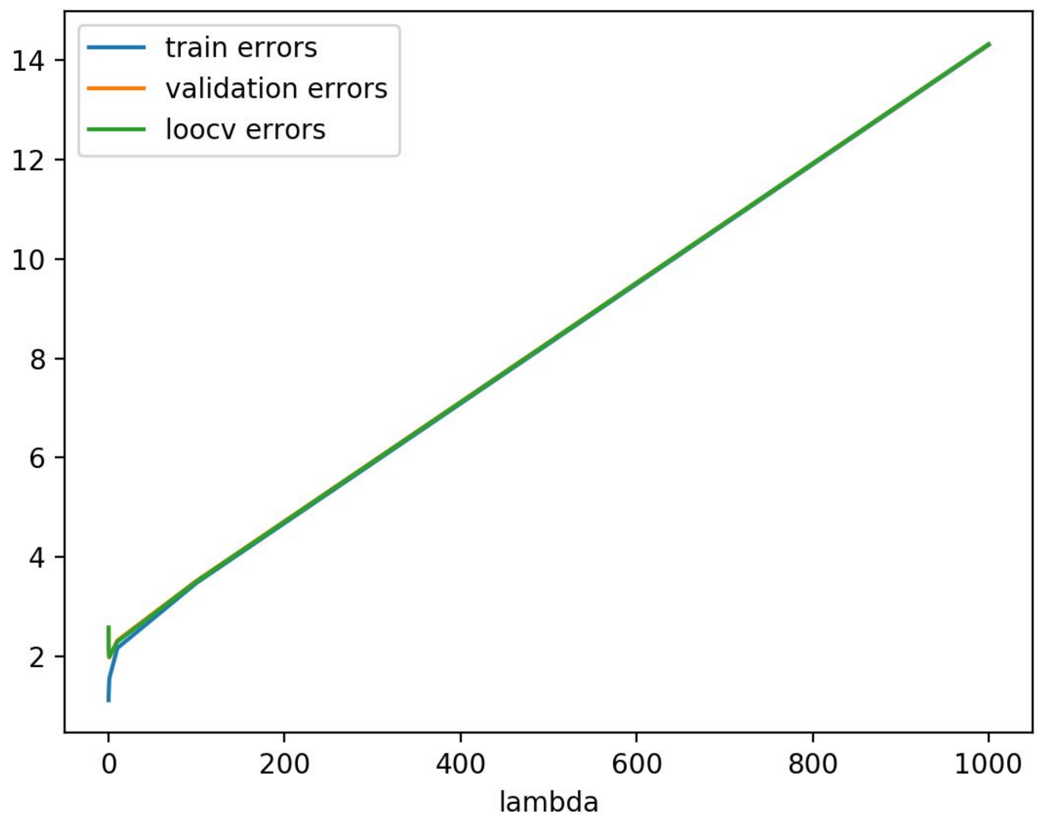
~~complexity = O~~

If $n > k \Rightarrow O(nk^2)$

If $n \leq k \Rightarrow O(k^3)$

Q3) 3.2.1

lambda	0.01	0.1	1	10	100	1000
Train Error	1.1204952813079863	1.222956840333819	1.568587189191946	2.16843745352464	3.4814321724863637	14.305635313079083
Val Error	2.5790449038104275	2.1557423907430207	1.9835860902518037	2.32204405212673	3.522936918785176	14.318915675050874
LOOCV	2.5800807745580427	2.1807668196014123	1.9980524704617526	2.2979058663641525	3.511911362808516	14.321265280702645



Loocv errors minimize at lambda = 0.793.

Q3) 3.2.2
The lambda which achieves best LOOCV performance on training data in 0.793.
Objective value = 16190.961703525318
Sum of squared errors = 11580.308410266747
Regularization term = 4610.65329325857

The lambda which achieves best LOOCV performance on training plus validation data is 0.839
The values on training data for the corresponding lambda are:
Objective value = 16448.21534249532
Sum of squared errors = 11750.560739969724
Regularization term = 4697.654602525596

Q3) 3.2.3
For lambda = 0.793, weight vector w is changed to take absolute value of the weight to calculate the important features.
Top 10 important features with their weights are:

Infused -	7.312501672952173
Pineapple orange -	6.034676760588582
Red -	5.876732475095196
Sweet black -	5.651101984591463
New French -	5.315713614971287
Future -	5.211911096535886
Little heavy -	5.162153523067218
Lifesaver -	4.973682122460218
Cocktail -	4.950345241857093
Cigar -	4.948205514948214

Top 10 least important features with their weights are:

Hazelnut -	-0.0003690725814067264
Softripe -	-0.0011641260498436168
Honey -	-0.0015629034897628458
Slight -	0.0018202491579017988
Acidity dry -	-0.0022761884634121543
Oakville -	-0.0029166312660322546
Florals -	-0.003906382083982862
Black cherries -	-0.004318998175222077
Fruit tart -	-0.004539502457639344
Lemons -	-0.004796385873220288

The type of wine seems to be more important than the flavour.

Q3) 3.2.4

For this question the model is trained on both training and validation sets combined. So the lambda which achieves minimum LOOCV is 0.839. Using this lambda weight vectors are calculated and Y values are predicted for test set.

RMSE for lambda 0.839 is 1.89244