CSE512 Fall 2018 Machine Learning - Homework 2

Your Name: Caitao Zhan

Solar ID: 111634527

NetID email address: caitao.zhan@stonybrook.edu

Names of people whom you discussed the homework with: None

1 Question 1 - Parameter Estimation

1.1 MLE

1. $P(\mathbf{X}|\lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \times \cdots \times \frac{\lambda^{x_n}}{x_n!}e^{-\lambda} = e^{-n\lambda} \times \frac{\lambda^{x_1+\cdots+x_n}}{x_1!\times\cdots\times x_n!}$
   $log(P(\mathbf{X}|\lambda)) = -n\lambda + (x_1 + \cdots + x_n)log\lambda - (logx_1! + \cdots + logx_n!)$
2. $\frac{\partial log(P(\mathbf{X}|\lambda))}{\partial\lambda} = -n + \frac{x_1+\cdots+x_n}{\lambda} = 0$
   $\Rightarrow \lambda = \frac{x_1+\cdots+x_n}{n}$
3. $\lambda = \frac{4+5+3+5+6+9+10}{7} = 6$

1.2 MAP

1.

$$\begin{aligned} P(\lambda|\mathbf{X}) &= \frac{P(\mathbf{X}|\lambda)P(\lambda)}{P(\mathbf{X})} \\ &= \frac{1}{P(\mathbf{X})} \times e^{-n\lambda} \cdot \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \times \cdots \times x_n!} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot e^{-\beta\lambda} \\ &= \frac{\beta^\alpha}{P(\mathbf{X}) \cdot (x_1! \times \cdots \times x_n!) \cdot \Gamma(\alpha)} \cdot \lambda^{x_1+\cdots+x_n+\alpha-1} \cdot e^{-(n+\beta)\lambda} \\ &\sim Gamma(\sum_{i=1}^{n} x_i + \alpha, n + \beta) \end{aligned}$$

2. $log(P(\lambda|\mathbf{X})) = log(\frac{\beta^\alpha}{P(\mathbf{X})\cdot(x_1!\times\cdots\times x_n!)\cdot\Gamma(\alpha)}) + (x_1 + \cdots + x_n + \alpha - 1)log\lambda - (n + \beta)\lambda$

   $\frac{\partial log(P(\lambda|\mathbf{X}))}{\partial\lambda} = \frac{x_1+\cdots+x_n+\alpha-1}{\lambda} - (n + \beta) = 0$
   $\Rightarrow \lambda = \frac{\sum_{i=1}^{n} x_i+\alpha-1}{n+\beta}$

1.3 Estimator Bias

1. $\eta = e^{-2\lambda} \Rightarrow \lambda = -\frac{1}{2}log\eta$
   $P(X|\eta) = \frac{1}{X!} \times (-\frac{1}{2}log\eta)^X \times e^{\frac{1}{2}log\eta}$
   $log(P(X|\eta)) = -log(X!) + Xlog(-\frac{1}{2}log\eta) + \frac{1}{2}log\eta$
   $\frac{\partial log(P(X|\eta))}{\partial\eta} = X(\frac{1}{-0.5log\eta} \cdot \frac{1}{-2\eta}) + \frac{1}{2\eta} = 0$
   $\Rightarrow \eta = e^{-2X}$

2.

$$\begin{aligned}
bias &= E[\hat{\eta}] - \eta \\
&= \sum_{x=0}^{\infty} e^{-2x} \cdot \frac{\lambda^x e^{-\lambda}}{x!} - e^{-2\lambda} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^{-2}\lambda)^x}{x!} - e^{-2\lambda} \\
&= e^{-\lambda} e^{e^{-2}\lambda} - e^{-2\lambda} \\
&= e^{-(1-e^{-2})\lambda} - e^{-2\lambda}
\end{aligned}$$

3. Let the unbiased estimator be $U(X)$.
The expectation of an unbiased estimator should equal to $e^{-2\lambda}$
$E(U(X)) = \sum_{x=0}^{\infty} U(x) \frac{\lambda^x}{x!} e^{-\lambda} = e^{-2\lambda}$
$\Rightarrow \sum_{x=0}^{\infty} U(x) \frac{\lambda^x}{x!} = e^{-\lambda}$
The only $U(X)$ that satisfy this is $U(X) = (-1)^X$, according to Taylor series expanding $e^{-\lambda}$. This is a bad estimator because it becomes 1 when X is even, and becomes -1 when X is odd, which is bad.

2 Question 2

2.1

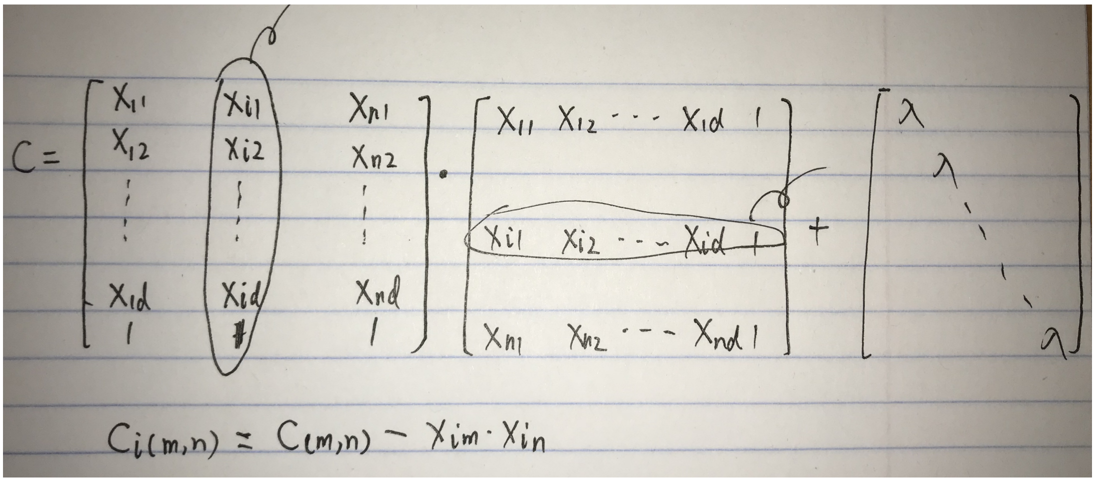
First derive the loss function, then let the differentiation of the loss function equal to zero.

$$
\begin{aligned}
L(\bar{\mathbf{w}}) &= ||\mathbf{X}^T\bar{\mathbf{w}} - \mathbf{y}||^2 + \lambda||\bar{\mathbf{w}}||^2 \\
&= (\mathbf{X}^T\bar{\mathbf{w}} - \mathbf{y})^T(\mathbf{X}^T\bar{\mathbf{w}} - \mathbf{y}) + \lambda\bar{\mathbf{w}}^T\bar{\mathbf{w}} \\
&= (\bar{\mathbf{w}}^T\mathbf{X} - \mathbf{y}^T)(\mathbf{X}^T\bar{\mathbf{w}} - \mathbf{y}) + \lambda\bar{\mathbf{w}}^T\bar{\mathbf{w}} \\
&= \bar{\mathbf{w}}^T\mathbf{X}\mathbf{X}^T\bar{\mathbf{w}} - 2\mathbf{y}^T\mathbf{X}^T\bar{\mathbf{w}} + \mathbf{y}^T\mathbf{y} + \lambda\bar{\mathbf{w}}^T\bar{\mathbf{w}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L(\bar{\mathbf{w}})}{\partial \bar{\mathbf{w}}} &= 2\mathbf{X}\mathbf{X}^T\bar{\mathbf{w}} - 2\mathbf{X}\mathbf{y} + 2\lambda\bar{\mathbf{w}} = 0 \\
\Rightarrow (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})\bar{\mathbf{w}} &= \mathbf{X}\mathbf{y} \\
\Rightarrow \bar{\mathbf{w}} &= (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y} \\
&= \mathbf{C}^{-1}\mathbf{d}
\end{aligned}
$$

2.2

$\mathbf{C}$ is a (d+1)×(d+1) matrix. $\mathbf{C}_{(i)}$ is also a (d+1)×(d+1) matrix.



From observation, we see $C_i(m,n) = C(m,n) - x_{im}x_{in}$. In matrix expression, it is

$$
\begin{aligned}
\bar{\mathbf{x}}_i &= [\mathbf{x_i}; 1] \\
\mathbf{C}_{(i)} &= \mathbf{C} - \bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T
\end{aligned}
$$

Similarly,

$$\mathbf{d}_{(i)} = \mathbf{d} - \bar{\mathbf{x}}_i y_i$$

where y_i is the ith element of $\mathbf{y}$

2.3

The Sherman-Morrison formula:

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

Replace $\mathbf{A}$ with $\mathbf{C}$, replace $\mathbf{u}$ with $-\bar{\mathbf{x}}_i$, replace $\mathbf{v}$ with $\bar{\mathbf{x}}_i$, then we get:

$$\begin{aligned}\mathbf{C}^{-1} &= (\mathbf{C} - \bar{\mathbf{x}}_\mathbf{i}\bar{\mathbf{x}}_\mathbf{i}^T)^{-1} \\ &= \mathbf{C}^{-1} + \frac{\mathbf{C}^{-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i}\end{aligned}$$

2.4

Use the result of subsection 2.2 and 2.3 to solve the problem in 2.4

$$\begin{aligned}\bar{\mathbf{w}}_{(i)} &= \mathbf{C}_{(i)}^{-1}\mathbf{d}_{(i)} \\ &= (\mathbf{C}^{-1} + \frac{\mathbf{C}^{-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i})(\mathbf{d} - \bar{\mathbf{x}}_i y_i) \\ &= \mathbf{C}^{-1}\mathbf{d} + \mathbf{C}^{-1}\bar{\mathbf{x}}_i(\frac{-y_i + y_i\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\mathbf{d} - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i y_i}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i}) \\ &= \bar{\mathbf{w}} + \mathbf{C}^{-1}\bar{\mathbf{x}}_i(\frac{-y_i + \bar{\mathbf{x}}_i\bar{\mathbf{w}}}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i})\end{aligned}$$

2.5

Use the result of subsection 2.4 to solve this problem. $\mathbf{C} = \mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}$. Note that $\mathbf{C} = \mathbf{C}^T$. So $(\mathbf{C}^{-1})^T = (\mathbf{C}^T)^{-1} = \mathbf{C}^{-1}$

$$\begin{aligned}\bar{\mathbf{w}}_{(i)}\bar{\mathbf{x}}_i - y_i &= [\bar{\mathbf{w}} + \mathbf{C}^{-1}\bar{\mathbf{x}}_i(\frac{-y_i + \bar{\mathbf{x}}_i^T\bar{\mathbf{w}}}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i})]^T\bar{\mathbf{x}}_i - y_i \\ &= [\bar{\mathbf{w}}^T + (\frac{-y_i + \bar{\mathbf{x}}_i^T\bar{\mathbf{w}}}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i})\bar{\mathbf{x}}_i^T(\mathbf{C}^{-1})^T]\bar{\mathbf{x}}_i - y_i \\ &= \frac{\bar{\mathbf{w}}^T\bar{\mathbf{x}}_i - \bar{\mathbf{w}}^T\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i - y_i\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_i^T\bar{\mathbf{w}}\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i} - y_i \\ &= \frac{\bar{\mathbf{w}}^T\bar{\mathbf{x}}_i - y_i\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i - y_i + y_i\bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i} \\ &= \frac{\bar{\mathbf{w}}^T\bar{\mathbf{x}}_i - y_i}{1 - \bar{\mathbf{x}}_i^T\mathbf{C}^{-1}\bar{\mathbf{x}}_i}\end{aligned}$$

2.6

1. $O(k^2n^2 + k^3n)$. By using formula 2.5

term	complexity
$\mathbf{X}\mathbf{X}^T$	$O(k^2n)$
$(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}$	$O(k^3)$
$(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}$	$O(k^2n)$
$\bar{\mathbf{w}}$	$O(k^2n + k^3)$
$\mathbf{C}^{-1}$	$O(k^3)$
$error \times 1$	$O(k^2n + k^3)$
$error \times n$	$O(k^2n^2 + k^3n)$

2. $O(k^2n + k^3)$. For the usual way of computing LOOCV, we compute $\bar{\mathbf{w}}$ and $\mathbf{C}^{-1}$ only once before entering the leave-one-out loop. So when entering the leave-one-out loop, $\bar{\mathbf{w}}$ and $\mathbf{C}^{-1}$ are treated as known vector and matrix:

term	complexity
$\bar{\mathbf{w}}^T\bar{\mathbf{x}}_i$	$O(k)$
$\bar{\mathbf{x}}_i\mathbf{C}^{-1}\bar{\mathbf{x}}_i$	$O(k^2)$
$error \times 1$	$O(k^2)$
$error \times n$	$O(k^2n)$
pre-compute $\bar{\mathbf{w}}$ and $\mathbf{C}^{-1}$ + $error \times n$	$O(k^2n + k^3)$

3 Question 3