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: Ting Jin

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.

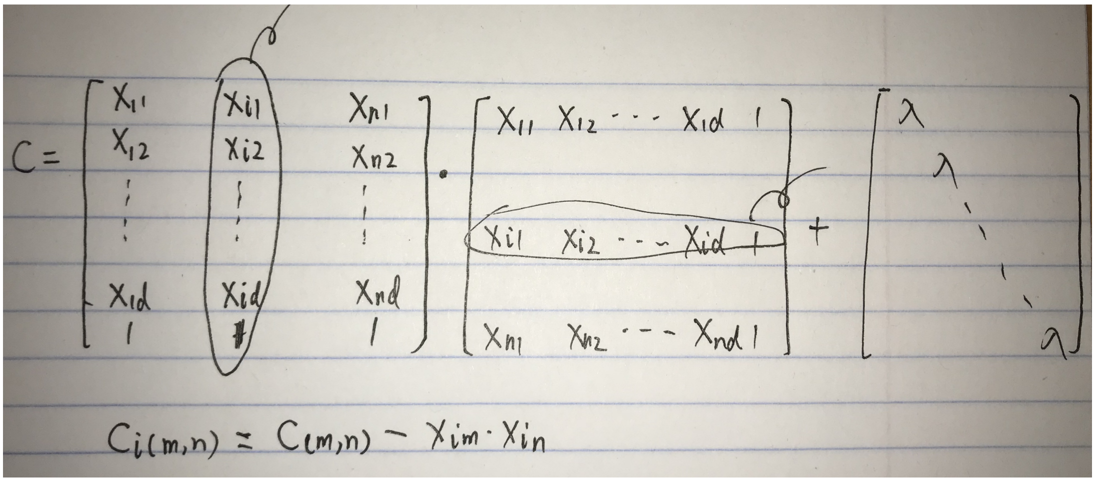

Figure 1: Visualize $\mathbf{C}$

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

3.1

3.2

1. Two plots. The first one is $\lambda = [0.01, 0.1, 1, 10, 100, 1000]$. Since 1000 is like an outlier, the second plot removes $\lambda = 1000$ to have a better observation on the curve.

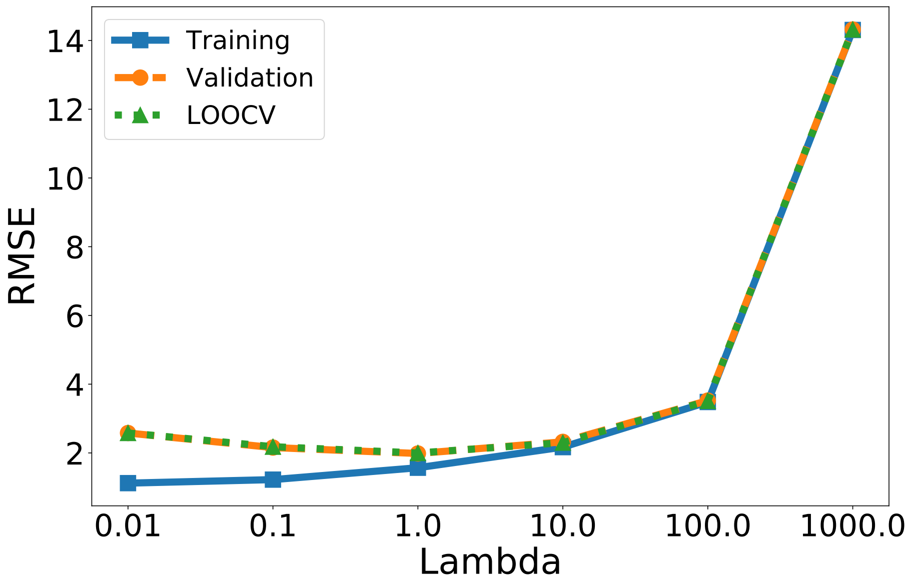

Figure 2: RMSE as a function of λ

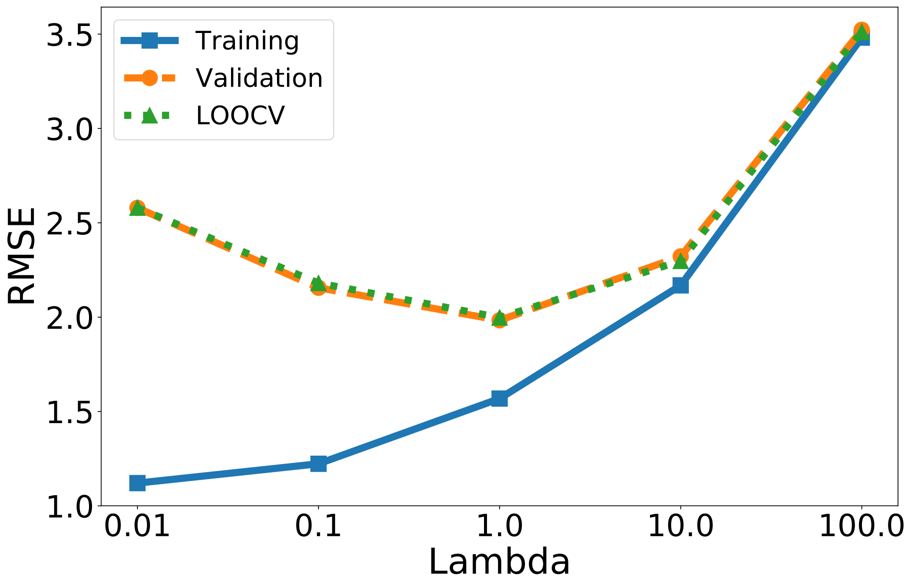

Figure 3: RMSE as a function of λ

2. From LOOCV curve or validation curve in figure 3, we observe that the best λ among [0.01, 0.1, 1, 10, 100, 1000] is **1**.
   The objective value = 17279.591821618673
   squared errors = 12302.328850486883
   regularization term = 4977.2629711317904.
   Objective = squared errors + regularization term

3. We do a sorting on the absolute values of the weights computed by the ridge regression with $\lambda = 1$. The important features are in figure 5 and the least important figures are in figure 4.
   It make sense intuitively that important features have high weights, unimportant features have low weights. Also, from observation, the least important features usually has a many zero values, usually over 4990 (out of 5000). This makes the distribution very skewed. The important features usually have more non-zero values, so the distribution is less skewed.

```
In [395]: index = 0
          weights = []
          for index in range(len(w)):
              tmp = Weight(index, w[index], features[index])
              weights.append(tmp)
          weights.sort()           # sort based on the absolute value of weights
          for weight in weights:
              print(weight)
```

```
index=2695, weight=0.000326, name=sweet wine
index=95, weight=0.000756, name=highlights
index=2510, weight=0.000947, name=tough tannins
index=1085, weight=0.001594, name=flavors black cherry
index=480, weight=0.002029, name=cases produced
index=2299, weight=0.002225, name=fare
index=2298, weight=0.002225, name=mouth
index=2551, weight=0.003226, name=assertive
index=1098, weight=0.004347, name=seamless
index=2673, weight=0.004427, name=lower
index=1169, weight=0.004536, name=lemons
```

least 10

Figure 4: Least 10 important features

```
index=2068, weight=4.806781, name=cocktail
index=781, weight=4.810647, name=price dry
index=2368, weight=4.811832, name=flavors nice
index=1924, weight=4.975534, name=future
index=2835, weight=5.055671, name=currant cola
index=186, weight=5.168260, name=new french
index=1272, weight=5.202842, name=sweet black
index=642, weight=5.211456, name=little heavy
index=754, weight=5.663754, name=pineapple orange
index=773, weight=5.835733, name=red
index=184, weight=7.056742, name=infused
```

top 10

Figure 5: Top 10 important features

4. Two Kaggle summition shown in figure 6

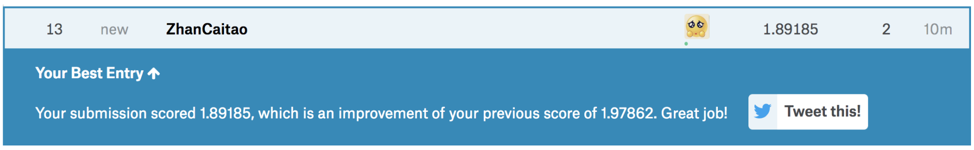

Figure 6: Kaggle submit