CSE512 Fall 2018 - Machine Learning - Homework 2

Your Name: Amit Dharmadhikari

Solar ID: 112044244

NetID email address: amit.dharmadhikari@stonybrook.com
adharmadhika@cs.stonybrook.edu

Names of people whom you discussed the homework with:

1 Question 1 - Parameter Estimation

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

1.1 Question 1.1 - MLE

1. Log-likelihood function of X given λ.

$$\cancel{L(x_1, x_2, \dots x_n | \lambda)}$$

$$L(x_1, x_2, \dots x_n|\lambda) = P(x_1|\lambda) \cdot P(x_2|\lambda) \dots P(x_n|\lambda)$$

$$= \prod_{i=1}^{n} P(x_i|\lambda)$$

We can get the log-likelihood function by taking the logarithm of the likelihood function.

$$l(x_1, x_2 \dots x_n|\lambda) = \log\left(\prod_{i=1}^{n} P(x_i|\lambda)\right)$$

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

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

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

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

2. MLE for λ in the general case.

$$\hat{\lambda}_n = \arg\max_{\lambda} (\text{likelihood function})$$

$$= \arg\max_{\lambda} (\text{log-likelihood function})$$

We can get the MLE by differentiating the log-likelihood function w.r.t. λ and setting it to 0.

$$\frac{\partial}{\partial \lambda} l(\lambda) = 0$$

$$\therefore \frac{\partial}{\partial \lambda}\left(\log \lambda . \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log (x_i!)\right) = 0$$

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

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

$\therefore$ MLE is :-

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

3. MLE for λ using the observed X.

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{7}(4+5+3+5+6+9+10)$$

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

$$= 6$$

1.2 Question 1.2 - MAP

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

Posterior distribution over λ is calculated as follows.

$$\cancel{P(\lambda|x_1, \ldots x_n) = P(\lambda \; P(x_1, \ldots x_n)}$$

$$\cancel{P(\lambda|x_1 \ldots x_n) = P(x_1 \ldots x_n)}$$

$$P(\lambda|x_1, x_2 \ldots x_n) = P(x_1, x_2, \ldots x_n|\lambda) \; P(\lambda) \quad - (1)$$

$$P(x_1, x_2 \ldots x_n|\lambda) = \prod_{i=1}^{n} P(x_i|\lambda)$$

We already have the following equation from the previous part of the question.

$$P(x_1, x_2 \ldots x_n|\lambda) = \log \lambda . \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

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

Put in equation 1

$$P(\lambda|x_1, x_2 \ldots x_n) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} P(\lambda)$$

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

$$\therefore P(\lambda \mid x_1, x_2 \ldots x_n) = e^{-\lambda(n+\beta)} \cdot \lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} \frac{\beta^{\alpha}}{\prod_{i=1}^{n} x_i! \, \Gamma(\alpha)}$$

2. Analytic expression for the MAP estimate of λ.

$$\lambda_{MAP} = \arg\max_{\lambda} \log\left(e^{-\lambda(n+\beta)} \cdot \lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} \frac{\beta^{\alpha}}{\prod_{i=1}^{n} x_i! \, \Gamma(\alpha)}\right)$$

$$= \arg\max_{\lambda} \left[-\lambda(n+\beta) + \left(\sum_{i=1}^{n} x_i + \alpha - 1\right)\log \lambda + \log\left(\frac{\beta^{\alpha}}{\prod_{i=1}^{n} x_i! \, \Gamma(\alpha)}\right)\right]$$

Take derivative with respect to λ and set it to 0.

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

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

1.3 Question 1.3 - Estimator Bias

$X \sim \text{Poisson}(\lambda)$

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

1. $\hat{\eta} = e^{-2X}$. We have to show that $\hat{\eta}$ is the maximum likelihood estimate of η

We have $P(X=k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

– (Poisson's distribution)

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

$\therefore \lambda = \frac{-\log \eta}{2}$

We will substitute this value of λ in the equation for Poisson's distribution.

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

In order to find the maximum likelihood estimate of η, we need to take the derivative w.r.t η and set it to 0.

First, we will take the log-likelihood.

$$l(\eta) = k \log\left(\frac{-\log \eta}{2}\right) + \frac{\log \eta}{2} - \log k!$$

Take derivative and equate it to zero.

$$\frac{d}{d\eta}\left[k \log\left(\frac{-\log \eta}{2}\right) + \frac{\log \eta}{2} - \log k!\right] = 0$$

$$\frac{k}{\eta \log \eta} + \frac{1}{2\eta} = 0$$

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

2. We need to prove that the bias of $\hat{n}$ is $e^{-(1-1/e^2)\lambda} - e^{-2\lambda}$

$\text{bias}(\hat{n}) = E[\hat{n}] - n$

$E[\hat{n}] = E[e^{-2X}] = \sum_{k=0}^{\infty} e^{-2k} P(X=k)$

According to Poisson's distribution,

$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Substituting this in the above equation, we get,

$E[e^{-2X}] = \sum_{k=0}^{\infty} e^{-2k} \frac{\lambda^k e^{-\lambda}}{k!}$

$= e^{-\lambda} + \frac{e^{-2}\lambda e^{-\lambda}}{1!} + \frac{e^{-4}\lambda^2 e^{-\lambda}}{2!} + \dots$

$= e^{-\lambda}\left[1 + \frac{e^{-2}\lambda}{1!} + \frac{e^{-4}\lambda^2}{2!} + \dots\right]$

~~Let's substitute e^{-2λ} = m for the sake of simplicity.~~

Let's substitute $e^{-2}\lambda = m$ for the sake of simplicity

$$E\left[e^{-2X}\right] = e^{-\lambda}\left[1 + \frac{m}{1!} + \frac{m^2}{2!} + \dots\right]$$

$1 + \frac{m}{1!} + \frac{m^2}{2!} + \dots$ is a Taylor expansion of e^m

$$\therefore E\left[e^{-2X}\right] = e^{-\lambda} \cdot e^{m}$$

Substituting the original value of m, we get,

$$E\left[e^{-2X}\right] = e^{-\lambda} e^{e^{-2}\lambda}$$
$$= e^{\lambda(-1+e^{-2})}$$
$$= e^{\lambda\left(-1+\frac{1}{e^2}\right)}$$
$$= e^{-\left(1-\frac{1}{e^2}\right)\lambda}$$

Substituting this in the equation for bias$(\hat{\eta})$, we get,

~~bias$(\hat{\eta})$ = E~~

$$\cancel{\text{bias}(\hat{\eta}) = e^{\lambda\left(\frac{1}{e^2}-1\right)}} - e^{-2\lambda}$$

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

Thus, we have proved that

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

Scanned by CamScanner

3. $\text{bias}(\hat{n}) = E[\hat{n}] - n$

Here, $\hat{n} = (-1)^x$. We need to prove that this is unbiased.

$$E[(-1)^x] = \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k e^{-\lambda}}{k!}$$

Expanding the summation, ~~we~~ we get,

$$E[(-1)^x] = e^{-\lambda} - \frac{\lambda e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} - \frac{\lambda^3 e^{-\lambda}}{3!} + \dots$$

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

According to the Taylor expansion,

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

Substituting this, we get,

$$E[(-1)^x] = e^{-\lambda} \cdot e^{-\lambda}$$
$$= e^{-2\lambda}$$

Substituting the values for $E[(-1)^x]$ and n in the equation for ~~bi~~ bias.

~~$\text{bias}(\hat{n}) = E[(-1)^x]$~~

$$\text{bias}(\hat{n}) = E[\hat{n}] - n = E[(-1)^x] - e^{-2\lambda}$$
$$= e^{-2\lambda} - e^{-2\lambda}$$
$$= 0$$

Since, the bias is 0, the estimator is unbiased. The reason why this is a bad estimator to use is that it doesn't generalize when the sample space increases.

2 Question 2 - Ridge Regression and LOOCV

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

2.1 $\bar{w} = [w; b]$

$\bar{X} = [X; 1_n^T]$

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

$C = \bar{X}\bar{X}^T + \lambda\bar{I}$

$d = \bar{X}y$

We need to show that the solution of Ridge regression is

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

The solution to ~~a~~ ridge regression can be obtained by minimizing the given equation.

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

First of all, we will ~~sum~~ remove the summation sign by vectorising.

$$\underset{w,b}{\text{minimize}}\ \lambda\|\bar{w}\|^2 + \|\bar{X}^T\bar{w} - y\|_2^2$$

Take gradient w.r.t. $\overline{w}$ and ~~set~~ set to 0.

$$2\lambda\overline{w} + 2\overline{x}(\overline{x}^T\overline{w} - y) = 0$$
$$\therefore \overline{x}(\overline{x}^T\overline{w} - y) + \lambda\overline{w} = 0$$
$$\therefore \overline{x}\,\overline{x}^T\overline{w} - \overline{x}y + \lambda\overline{w} = 0$$
$$\therefore (\overline{x}\,\overline{x}^T + \lambda I)\overline{w} = \overline{x}y$$
$$\therefore \overline{w} = (\overline{x}\,\overline{x}^T + \lambda I)^{-1}\overline{x}\,y$$
$$\therefore \overline{w} = c^{-1}d$$

2.2 $c = \overline{x}\,\overline{x}^T + \lambda I$

Removing x_i, we get C_i as follows.

$$C_i = \overline{x}\,\overline{x}^T - \overline{x_i}\,\overline{x_i}^T + \lambda I \;\cancel{= C - \overline{x_i}\,\overline{x_i}^T}$$
$$= C - \overline{x_i}\,\overline{x_i}^T$$

~~2.3~~

$$d = \overline{x}y$$
$$d_i = (\overline{x} - \overline{x_i})y_i = \overline{x}y - \overline{x_i}\,y_i = d - \overline{x_i}\,y_i$$

2.3 We had calculated C_i in the previous question.

$$C_i = C - \overline{x_i}\,\overline{x_i}^T$$
$$C_i^{-1} = (C - \overline{x_i}\,\overline{x_i}^T)^{-1}$$

Now we will use the Sherman-Morrison formula according to the hint provided.

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

Here, A corresponds to C, u to $-\overline{x_i}$ and v^T to $\overline{x_i}^T$.

$$\therefore C_i^{-1} = C^{-1} + \frac{C^{-1}\bar{x}_i \bar{x}_i^T C^{-1}}{1 - \bar{x}_i^T C^{-1} \bar{x}_i}$$

2.4 ~~We kn~~

We need to show that:

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

We have $\bar{w} = C^{-1} d$ from question 2.1

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

$$\begin{aligned}
\bar{w}_i &= C_i^{-1} d_i \\
&= C_i^{-1} \bar{X} y - \bar{x}_i y_i \\
&= \left(C^{-1} + \frac{C^{-1}\bar{x}_i \bar{x}_i^T C^{-1}}{1 - \bar{x}_i^T C^{-1} \bar{x}_i}\right)(\bar{X} y - \bar{x}_i y_i) \\
&= C^{-1}\bar{X} y - C^{-1}\bar{x}_i y_i + \frac{C^{-1}\bar{x}_i \bar{x}_i^T C^{-1} \bar{X} y}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} \\
&\quad - \frac{C^{-1}\bar{x}_i \bar{x}_i^T C^{-1} \bar{x}_i y_i}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} \\
&= \bar{w} - C^{-1}\bar{x}_i y_i + \frac{C^{-1}\bar{x}_i \bar{x}_i^T C^{-1} \bar{X} y}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} \\
&\quad - \frac{C^{-1}\bar{x}_i \bar{x}_i^T C^{-1} \bar{x}_i y_i}{1 - \bar{x}_i^T C^{-1} \bar{x}_i}
\end{aligned}$$

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

$$+ \frac{C^{-1}\bar{x}_i\bar{x}_i^T C^{-1}\bar{X}y}{1-\bar{x}_i^T C^{-1}\bar{x}_i} - \frac{C^{-1}\bar{x}_i\bar{x}_i^T C^{-1}\bar{x}_i y_i}{1-\bar{x}_i^T C^{-1}\bar{x}_i}$$

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

$$+ \frac{C^{-1}\bar{x}_i\bar{x}_i^T C^{-1}\bar{X}y}{1-\bar{x}_i^T C^{-1}\bar{x}_i} - \frac{C^{-1}\bar{x}_i\bar{x}_i^T C^{-1}\bar{x}_i y_i}{1-\bar{x}_i^T C^{-1}\bar{x}_i}$$

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

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

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

$$\cancel{= \bar{w} - C^{-1}\bar{x}_i(y_i(1-x_i}$$

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

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

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

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

Thus, we have proved that

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

2.5 We have to show that

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

We will ~~pu~~ substitute the value of $\overline{w_i}$ obtained from the previous question into the LHS of the above equation.

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

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

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

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

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

$$\cancel{= (\overline{w}^T \overline{x_i} - y_i) - (\overline{w}^T \overline{x_i} - y}$$

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

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

Thus, we have proved that,

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

2.6 Usual way of computing LOOCV is

$$\sum_{i=1}^{n} \left(\overline{w_i}^T \overline{x_i} - y_i\right)^2$$

The formula given in question 2.5 is:

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

In the usual way, we need to perform the matrix inverse operation for each training sample. Hence, the time complexity is $O(nk^3)$, where n is the number of training samples. (The complexity of matrix inverse operation is $O(k^3)$ and it is the most expensive operation in our computation).

For the formula mentioned in question 2.5, we need to perform ~~only once~~ the matrix inverse operation only once, for C^{-1}.

The matrix inversion operation is the most expensive operation here as well, as the time to compute $\bar{w}^T x_i - y_i$ is ~~ok~~ ~~and~~ $O(k)$ and that for $1 - \bar{x}_i^T C^{-1} \bar{x}_i$ is $O(k^2)$

Thus,

Time complexity for the usual way $= O(nk^3)$

Time complexity for the formula mentioned in question 2.5 $= O(k^3)$

Thus, we see that the formula mentioned in question 2.5 is more efficient than the usual way.