CS-534 Assignment-1

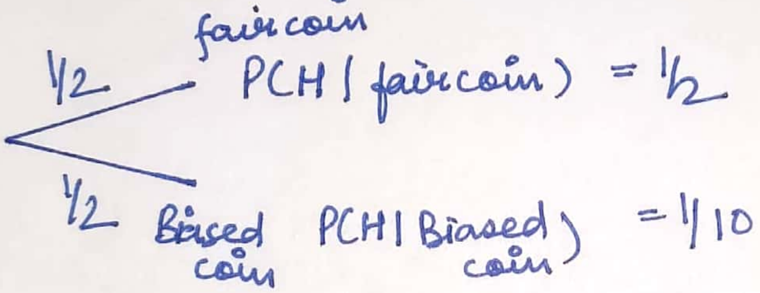
Submitted By: AKASH AGARWAL
OSU ID NUMBER: 933-471-097

Ques1. Consider two coins, one is fair and the other has a 1/10 probability for head. Now you randomly pick one of the coins, & toss it twice. Answer the following questions:

(A) What is the probability that you picked the fair coin? What is the probability of the first toss being head?

(B) If both tosses are heads, what is the probability that you have chosen the fair coin?

Ans. Given: 2 coins $\begin{cases} 1 \text{ fair coin} \\ 1 \text{ biased coin}, \quad P(H) = \frac{1}{10} \end{cases}$



(A) Let P(fair) be the probability of choosing fair coin.

Thus, $\boxed{P(\text{fair}) = 1/2}$

Let P(H) be the probability of first toss being head.

$$P(H) = P(\text{fair}) . P(H|\text{fair}) + P(\text{biased}) . P(H|\text{biased})$$

$$= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{10} = \frac{1}{2}\left[\frac{1}{2} + \frac{1}{10}\right] = \frac{1}{2}\left(\frac{6}{10}\right)$$

$$\boxed{P(H) = 0.3}$$

1 (B) We need to find, P(fair coin | two heads).

We know that, $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$ {Bayes Rule}

∴ using Bayes Rule,

$$P(\text{fair coin} \mid \text{two heads}) = \frac{P(\text{two heads} \mid \text{fair coin}) \cdot P(\text{fair coin})}{P(\text{two heads})} \quad -(1)$$

$P(\text{two heads} \mid \text{fair coin}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ ∵ Once the fair coin is choosen, $P(H) = 1/2$

$P(\text{fair coin}) = \frac{1}{2}$

$$P(\text{two heads}) = P(\text{fair coin}) \cdot P(\text{two heads} \mid \text{fair coin}) + P(\text{biased coin}) \cdot P(\text{two heads} \mid \text{biased coin})$$

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{10} \times \frac{1}{10}$$

$$= \frac{1}{2}\left[\frac{1}{4} + \frac{1}{100}\right] = \frac{1}{2}\left[\frac{26}{100}\right] = \frac{13}{100}$$

Putting values in eqⁿ (1)

$$P(\text{fair coin} \mid \text{two heads}) = \frac{\frac{1}{4} \times \frac{1}{2}}{\frac{13}{100}} = \frac{25}{26} \simeq 0.96 \text{ Ans.}$$

Ques2. Given a set of i.i.d samples $x_1, x_2, x_3, \ldots x_n \sim$ uniform $(0, \theta)$

(A) Write down the likelihood function of θ.

(B) Find the maximum likelihood estimator for θ.

Ans. The uniform distribution for any set of samples x in $[a, b]$ is given as follows:

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

A.T.Q. $a = 0$, $b = \theta$

$$\Rightarrow p(x \mid \theta) = \begin{cases} \frac{1}{\theta}, & \text{for } 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$$

Such that, $0 \le x_1, x_2, x_3 \ldots x_n \le \theta$

(A).

Now, $\underset{\text{Likelihood of } \theta}{L(\theta)} = p(x_1, x_2, x_3 \ldots x_n \mid \theta)$

$$= p(x_1 \mid \theta)\, p(x_2 \mid \theta)\, p(x_3 \mid \theta) \ldots p(x_n \mid \theta)$$

$\because$ the sample points are independently & identically distributed.

$$= \prod_{i=1}^{n} p(x_i \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta}$$

$$\boxed{L(\theta) = \frac{1}{\theta^n}}$$

let $l(\theta)$ be the log likelihood of θ.

Thus, $l(\theta) = \ln(L(\theta))$

$$l(\theta) = \ln(\theta^{-n})$$

$$\Rightarrow \boxed{l(\theta) = -n\ln(\theta)}$$

2(B) To find the value of θ that maximizes $l(\theta)$, we first need to take gradient of $l(\theta)$ w.r.t θ.

$$\Rightarrow \frac{\partial l(\theta)}{\partial\theta} = \frac{\partial(-n\ln(\theta))}{\partial\theta} = \frac{-n}{\theta}$$

We can see that as the value of θ increases, $\frac{\partial l(\theta)}{\partial(\theta)}$ decreases, and vice versa.

Thus, to maximize $\frac{\partial l(\theta)}{\partial\theta}$, we must choose smallest possible value of θ.

Now, we know that $0 \leq x_1, x_2, x_3 \dots x_n \leq \theta$.

Thus, for the given sample $x: \{x_1, x_2 \dots x_n\}$ the minimum possible value of $\theta = x_n$

$\therefore$ Maximum likelihood estimator for $\theta = x_n$

Ans

Ques 3. In class when discussing linear regression, we assume that the Gaussian noise is independently identically distributed. Now assume that the Gaussian noise is independently but each $\epsilon_m \sim N(0, \sigma_m^2)$, i.e. it has its own distinct variance.

A) Write down the log likelihood of w.

B) Show that maximizing the log likelihood is equivalent to minimizing a weighted least square loss function $J(W) = \frac{1}{2}\sum_{m=1}^{n} a_m (W^T x_m - y_m)^2$, express each a_m in terms of σ_m.

(C) Derive a batch gradient descent algorithm for optimizing this objective.

(D) Derive a closed form solution to ~~this~~ optimizing this objective.

Ans. We know that the data is distributed in Normal distribution and the noise ϵ_i is not i.i.d.

i.e.: $$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\left(\frac{-(\epsilon_i)^2}{2\sigma_i^2}\right)}$$

$\therefore$ for $y = w^T x + \epsilon$

$$L(w) = p(y_1, y_2, \ldots y_N \mid x_1, x_2 \ldots x_n ; w)$$

$$= \prod_{i=1}^{N} p(y_i \mid x_i ; w)$$

(A) Let $l(w)$ be the log-likelihood of w

$\therefore$ $$l(w) = \log(L(w)) = \log\left[\prod_{i=1}^{n} p(y_i \mid x_i ; w)\right]$$

$$= \log\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(\frac{-(y_i - w^T x_i)^2}{2\sigma_i^2}\right)\right]$$

$$= -\sum_{i=1}^{n} \log(\sqrt{2\pi}\,\sigma_i) + \sum_{i=1}^{n} \log\left[\exp\left(\frac{-(y_i - w^T x_i)^2}{2\sigma_i^2}\right)\right]$$

$$= -\sum_{i=1}^{n} \log\sqrt{2\pi} - \sum_{i=1}^{n} \log(\sigma_i) - \sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2\sigma_i^2}$$

$$l(w) = -n\log\sqrt{2\pi} - \sum_{i=1}^{n} \log(\sigma_i) - \sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2\sigma_i^2}$$

Ans.

3.(B) To maximize $l(w)$, we take derivative of $l(w)$ w.r.t. w.

$$\therefore \frac{\partial(l(w))}{\partial w} = 0 - 0 + \operatorname{argmin}\left[\sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2\sigma_i^2}\right]$$

$$= \operatorname{argmin}\left[\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - w^T x_i)^2\right] \quad \text{— (i)}$$

If $J(w) = \frac{1}{2}\sum_{m=1}^{n} a_m (w^T x_m - y_m)^2$ — (ii)

then on comparing (i) & (ii)

maximizing the log likelihood is equivalent to minimizing the weighted least square loss function $J(w)$.

Also, $\boxed{a_m = \frac{1}{\sigma_m^2}}$ Ans.

Q3(C) $J(\omega) = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [\omega^T x_i - y_i]^2$

$$\frac{\partial J(\omega)}{\partial \omega_m} = \frac{1}{2} \sum_{i=1}^{n} \sigma_i^{-2} . 2 [\omega^T x_i - y_i] x_i$$

$$= \sum_{i=1}^{n} \sigma_i^{-2} [\omega^T x_i - y_i] x_i$$

Batch Gradient Descent Algo :-

Repeat {

$$\frac{\partial E(\omega)}{\partial (\omega)} = \sum_{i=1}^{n} \sigma_i^{-2} (\omega^T x_i - y_i) x_i$$

$$\omega = \omega - \lambda \frac{\partial E(\omega)}{\partial (\omega)}$$

} while $\frac{\partial E(\omega)}{\partial \omega} \leq \underbrace{\phi}_{\text{Threshold}}$

Q3(D) $J(\omega) = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [\omega^T x_i - y_i]^2$

let $S = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & \cdots & 0 \\ & \frac{1}{\sigma_2^2} & \cdots & & \vdots \\ \vdots & & \ddots & & \\ & & & & \frac{1}{\sigma_n^2} \\ \vdots & 0 & \cdots & & \end{bmatrix}_{n \times n}$

~~$X = [x_0 \; x_i \; n]$~~ $X = \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{bmatrix}_{n \times m}$

$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$

let W be the weight matrix.

$\therefore \frac{\partial J(w)}{\partial w} = 0$ to minimize the cost function

Now, in matrix form, $J(w) = \frac{1}{2}\left((XW-Y)^T S(XW-Y)\right)$

$$\frac{\partial J(W)}{\partial W} = \frac{1}{2}\frac{\partial}{\partial W}\left[(XW-Y)^T S(XW-Y)\right]$$

$$= X^T SXW - X^T SY$$

$$\frac{\partial J(w)}{\partial (W)} = 0$$

$$\Rightarrow X^T SXW = X^T SY$$

$$\Rightarrow \boxed{W = (X^T SX)^{-1} X^T SY}$$

This is the closed form solution for the optimization problem.

Ques. 4 Consider a binary classification task with the following loss matrix.

$$\begin{array}{cc|cc} & & \multicolumn{2}{c}{y} \\ & & 0 & 1 \\ \hline \hat{y} & 0 & 0 & 10 \\ & 1 & 5 & 0 \end{array}$$

We have build a probabilistic model that for each example x gives us an estimated $P(y=1|x)$. It can be shown that, to minimize the expected loss for our decision, we should set a probability threshold θ & predict $\hat{y}=1$ if $P(y=1|x)>\theta$ & $\hat{y}=0$ otherwise.

(A) Compute θ for the above given loss matrix.

(B) Show a loss matrix where the threshold is 0.1.

Ans. 4(A) Our classifier should predict $\hat{y}=1$ only if the expected cost of predicting 1 is less than expected cost of predicting 0.

Let the cost incurred for predicting $\hat{y}=1$ when $y=0$ be a.

Let the cost incurred for predicting $\hat{y}=0$ when $y=1$ be b.

$\therefore$ $$P(y=0|x).a < P(y=1|x).b$$

$$\Rightarrow [1-P(y=1|x)].a < P(y=1|x).b$$

$$a - P(y=1|x).a < P(y=1|x).b$$

$$-(a+b).P(y=1|x) < -a$$

$$\Rightarrow \boxed{P(y=1|x) > \frac{a}{a+b}}$$

A.T.Q $a = 5$
$b = 10$

$\therefore \quad P(y=1|x) > \frac{5}{5+10}$

$\Rightarrow \quad P(y=1|x) > \frac{1}{3}$

$\therefore \quad \boxed{\text{Threshold } (\theta) = \frac{1}{3}}$

4.(B) For threshold = 0.1

$$\frac{a}{a+b} = \frac{1}{10}$$

$\therefore \quad a = 1$
and $\quad b = 9$

$\therefore$

$\hat{y}$ \ y	0	1
0	0	9
1	1	0

Ans.

Ques 5. Consider the maximum likelihood estimation problem for multi-class logistic regression using the soft-max function defined below:

$$p(y=k|x) = \frac{\exp(w_k^T x)}{\sum_{j=1}^{K} \exp(w_j^T x)}$$

We can write the likelihood function as:

$$L(\omega) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(y=k|x_i)^{I(y_i=k)}$$

where $I(y_i=k)$ is the indicator function, taking value 1 if y_i is k.

(A) What are i & k in this likelihood function?

(B) Compute the log-likelihood function.

(C) What is the gradient of the log-likelihood function w.r.t the weight vector w_c of class c?

Ans. (A) 'i' denotes the i^{th} datapoint in the training set. and 'k' denotes the k^{th} class that the output y belongs to.

(B) Let $l(\omega)$ be the log-likelihood function of ω.

$$\therefore \; l(\omega) = \log L(\omega) = \log\left[\prod_{i=1}^{N}\prod_{k=1}^{K} p(y=k|x_i)^{I(y_i=k)}\right]$$

$$= \log\left[\prod_{i=1}^{N}\prod_{k=1}^{K}\left(\frac{e^{w_k^T x_i}}{\sum_{j=1}^{K} e^{w_j^T x_i}}\right)^{I(y_i=k)}\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} I(y_i=k)\log\left[\frac{e^{w_k^T x_i}}{\sum_{j=1}^{K} e^{w_j^T x_i}}\right]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \left[I(y_i = k) \left[\log\left[e^{w_k^T x_i}\right] - \log\left[\sum_{j=1}^{K} e^{w_j^T x_i}\right] \right] \right]$$

$$\boxed{l(w) = \sum_{i=1}^{N} \sum_{k=1}^{K} I(y_i = k)\, w_k^T x_i - \sum_{i=1}^{N} \sum_{k=1}^{K} I(y_i = k) \log\left[\sum_{j=1}^{K} e^{w_j^T x_i}\right]}$$

Ans.

5(c) Gradient of log-likelihood function w.r.t the weight vector $w_c = \frac{\partial\, l(w)}{\partial w_c}$

$$\Rightarrow \frac{\partial(l(w))}{\partial w_c} = \sum_{i=1}^{N} I(y_i = c) x_i - \sum_{i=1}^{N} \left(\frac{1}{\sum_{j=1}^{K} e^{w_j^T x_i}} \right) \cdot e^{w_c^T x_i} \cdot x_i$$

$$= \sum_{i=1}^{N} \left[I(y_i = c) - \frac{e^{w_c^T x_i}}{\sum_{j=1}^{K} e^{w_j^T x_i}} \right] x_i$$

$$\boxed{\frac{\partial(l(w))}{\partial w_c} = \sum_{i=1}^{N} \underbrace{\left[I(y_i = c) - p(y = c \mid x) \right]}_{\text{difference b/w predicted \& true value}} x_i}$$

Ans.