HW1 — ISYE6416

Ashish Dhiman — ashish.dhiman9@gatech.edu

January 22, 2023

1 Question 1: Moment Generating Function (MGF)

We are given, MGF as

$$G(\theta) = E[e^{\theta X}]$$

Now by Taylor expansion we have

$$e^{\theta X} = \sum_{i=0}^{\infty} \frac{(\theta X)^r}{r!}$$

This can be expanded as:

$$\begin{aligned} e^{\theta X} = 1 + \frac{\theta X}{1} + \frac{(\theta X)^2}{2} + \frac{(\theta X)^3}{6} + ... \\ \text{Taking partial derivative wrt theta} \\ \frac{\partial e^{\theta X}}{\partial \theta} = 0 + X + (\theta X).X + \theta(...) \\ \text{similarly, } \frac{\partial^2 e^{\theta X}}{\partial \theta^2} = 0 + 0 + (X).X + \theta(...) \end{aligned} \tag{1}$$

Now

$$\frac{\partial G(\theta)}{\partial \theta} = E[\frac{\partial e^{\theta X}}{\partial \theta}] \quad \text{By definition of MGF}$$

$$\begin{aligned} \implies \frac{\partial G(\theta)}{\partial \theta} = E[0 + X + (\theta X).X + \theta(...)] \\ \implies \frac{\partial G(\theta)}{\partial \theta}|_{\theta=0} = E[X] \\ \text{Similarly} \\ \implies \frac{\partial^2 G(\theta)}{\partial \theta^2}|_{\theta=0} = E[X^2] \end{aligned} \tag{2}$$

2 Question 2: Maximum likelihood estimator

2.1 part a: MLE Estimate

We are given $\{X_i\}$ as IIDs RVs sampled from pdf:

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

For MLE of a and b we need to find the Likelihood of $X_1, X_2 ... X_n$ and maximise it.

$$a_{MLE}, b_{MLE} = arg \max_{(a,b)} = Lik(X_1, X_2 ... X_n | a, b)$$

$$\begin{aligned} Lik(X_1, X_2 ... X_n | a, b) &= p(X_1, X_2 ... X_n | a, b) \\ = p(X_1|a,b) * p(X_1|a,b) ... p(X_1|a,b) &= \prod_{i=1}^{n} p(X_i|a,b) \\ &= \prod_{i=1}^{n} f(x_i) \quad \text{(From pdf)} \end{aligned} \tag{3}$$

Now the above $Lik > 0$ only if all product terms are greater than 0. This is true if $\forall X_i \in [a, b]$, otherwise Lik becomes 0. In other words:

$$b >= \max(X_1, X_2 ... X_n) \quad \text{and } a <= \min(X_1, X_2 ... X_n)$$

Keeping the above constraints in mind Lik is maximised if $b - a$ is as small as possible. Now minimising $b - a \implies$ pick the smallest b and largest a.

$$\begin{aligned} a_{MLE} &= \min(X_1, X_2 ... X_n), \quad \text{and} \\ b_{MLE} &= \max(X_1, X_2 ... X_n) \end{aligned} \tag{4}$$

2.2 part b: Unbiased Estimate

A estimator θ^* is unbiased for parameter θ iff $E[\theta^*] = \theta$. Hence for a_{MLE}, b_{MLE} to be unbiased, $E[a_{MLE}] = a$ and $E[b_{MLE}] = b$.

Let's find $P(b_{MLE} <= b)$ or the CDF of b_{MLE}.

$$\begin{aligned} P(b_{MLE} <= b) = P(b_{MLE} < b) \quad &\text{(for continuous distribution point prob. = 0)} \\ &\text{From part a we also have} \\ b >= \max(X_1, X_2 ... X_n) \text{ and } b_{MLE} &= \max(X_1, X_2 ... X_n) \\ &\implies b >= b_{MLE} \\ \implies P(b_{MLE} <= b) &= P(b_{MLE} < b) = 1 \\ &\implies E[b_{MLE}] < b \end{aligned} \tag{5}$$

Hence b_{MLE} is biased. Similar proof for a_{MLE}

3 Question 3: Bayesian Inference

We are given: $x \sim \mathcal{N}(\mu, \sigma^2)$

3.1 part a

$\mu \sim \mathcal{N}(\theta, \tau^2)$ We need to find posterior for μ, or

$$p(\mu|x) = \frac{p(x|\mu) * p(\mu)}{p(x)} \propto p(x|\mu) * p(\mu)$$

Hence plugging in pdf function we have:

$$\begin{aligned} p(\mu|x) &\propto \big[\frac{1}{\sqrt{2\pi\sigma^2}} * exp(\frac{-(x-\mu)^2}{2\sigma^2})\big] * \big[\frac{1}{\sqrt{2\pi\tau^2}} * exp(\frac{-(\mu-\theta)^2}{2\tau^2})\big] \\ &\propto \exp\big(-0.5 * [\frac{-(x-\mu)^2}{2\sigma^2} + \frac{-(\mu-\theta)^2}{2\tau^2}]\big) \\ &\text{Looking on terms only inside exponenent we have} \\ [\frac{-(x-\mu)^2}{2\sigma^2} + \frac{-(\mu-\theta)^2}{2\tau^2}] &= \frac{\tau^2 * (x-\mu)^2) + \sigma^2 * (\mu-\theta)^2}{\sigma^2\tau^2} \\ &= \frac{\mu^2(\sigma^2+\tau^2) - 2\mu(\tau^2 x + \sigma^2\theta) + \tau^2 x^2 + \sigma^2\theta^2}{\sigma^2\tau^2} \\ &\text{Dividing num and deonom by } (\tau^2+\sigma^2) \end{aligned} \tag{6}$$

$$\begin{aligned} \frac{num}{(\tau^2+\sigma^2)} &= \mu^2 - 2\mu(\frac{\tau^2 x}{\tau^2+\sigma^2} + \frac{\sigma^2\theta}{\tau^2+\sigma^2}) + \frac{(\tau^2x^2+\sigma^2\theta^2)}{\tau^2+\sigma^2} \\ &= ... \pm (\frac{2\tau^2\sigma^2 x\theta}{(\tau^2+\sigma^2)^2}) \quad \text{completing squares} \\ &= (\frac{\tau^2 x}{\tau^2+\sigma^2} + \frac{\sigma^2\theta}{\tau^2+\sigma^2} - \mu)^2 + \text{constant} \end{aligned} \tag{7}$$

Hence,

$$\begin{aligned} \implies p(\mu|x) \propto \exp -0.5 * \Big(\frac{\frac{(\frac{\tau^2 x}{\tau^2+\sigma^2} + \frac{\sigma^2\theta}{\tau^2+\sigma^2} - \mu)^2}{\tau^2+\sigma^2}}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\Big) \\ \text{Hence} \\ \mu|x \sim \mathcal{N}(\frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\theta, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}) \end{aligned} \tag{8}$$

3.2 part b

Here $\mu \sim \mathcal{U}(0,1)$

$$\begin{aligned} \implies p(\mu|x) \propto \big[\frac{1}{\sqrt{2\pi\sigma^2}} * exp(\frac{-(x-\mu)^2}{2\sigma^2})\big] * \big[\frac{1}{1-0}\big] \\ \text{Hence} \\ \mu|x \sim \mathcal{N}(\mu, \sigma^2) \end{aligned} \tag{9}$$

4 Question 4: Basic Optimisation

4.1 part a

4.2 part b

4.3 part c

5 Question 5: Weighted Regression

Given loss of Regression function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n} w_i(\theta^T x_i - y_i)^2.$$

5.1 part a

Let X is matrix of observations with dimensions $(n{*}p)$ i.e. n obs. and p features, and theta be the weights vector of dimensions $(p*1)$.

In other words

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{1p} \\ x_{21} & x_{22} & x_{2p} \\ x_{n1} & x_{n2} & x_{np} \end{bmatrix}_{n*p} = \begin{bmatrix} {x_1}^T \\ {x_2}^T \\ {x_n}^T \end{bmatrix}_{n*p}$$

$$\implies \hat{y} = X\theta \quad \text{then,}$$

$$X\theta - y = \hat{y} - y = \begin{bmatrix} x_1\theta - y1 \\ x_2\theta - y2 \\ x_n\theta - yn \end{bmatrix}_{n*1}$$

$$\text{Also let diagonal weights matrix} W = \begin{bmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_n \end{bmatrix}_{n*n}$$

Now we can see that

$$J(\theta) = (X\theta - y)^T W(X\theta - y) =$$

$$\begin{bmatrix} x_1\theta - y1 & x_2\theta - y2 & x_n\theta - yn \end{bmatrix}_{1*n} * \begin{bmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_n \end{bmatrix}_{n*n} * \begin{bmatrix} x_1\theta - y1 \\ x_2\theta - y2 \\ x_n\theta - yn \end{bmatrix}_{n*1}$$

$$= \begin{bmatrix} w_1(x_1\theta - y1) & w_2(x_2\theta - y2) & w_n(x_n\theta - yn) \end{bmatrix}_{1*n} * \begin{bmatrix} x_1\theta - y1 \\ x_2\theta - y2 \\ x_n\theta - yn \end{bmatrix}_{n*1}$$

$$\implies J(\theta) = w1(x_1\theta - y1)^2 + ... + w_n(x_n\theta - yn)^2 \tag{10}$$

5.2 part b

We are given (x_i, y_i), $i = 1, \ldots, n$ of n independent examples, and $y_i = \theta^T x_i + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, {\sigma_i}^2)$

This means,

$$p(y_i|x_i,\theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp(-\frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2})$$

Now we want to find θ_{MLE} s.t.:

$$\theta_{MLE} = arg\max_{\theta} Lik(y_1, y_2, ..., y_n|x_1, x_2, ..., x_n, \theta)$$

$$
\begin{aligned}
Lik &= p(y_1, y_2, ..., y_n | x_1, x_2, ..., x_n, \theta) \\
&= \prod_{i=1}^{n} p(y_i | x_i, \theta) \quad \text{Independent examples} \\
&= \prod_{i=1}^{n} [\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2})] \\
\text{ignoring } \pi, \sigma_i &\text{ terms as they won't impact optimisation wrt } \theta \qquad (11) \\
&\propto \exp\big(-0.5 * \sum_{i=1}^{n} \frac{(y_i - \theta^T x_i)^2}{\sigma_i^2})\big) \\
&\text{Let } w_i = 1/\sigma_i^2 \text{, then:} \\
Lik &\propto \exp\big(-0.5 * \sum_{i=1}^{n} w_i (y_i - \theta^T x_i)^2\big)
\end{aligned}
$$

Now maximising above likelihood is equivalent to minimising the terms in exponent since if x goes to -inf exp(-x) goes to inf

$$
\implies \theta_{MLE} = arg \max_{\theta} Lik = arg \min_{\theta} \sum_{i=1}^{n} w_i (y_i - \theta^T x_i)^2 \tag{12}
$$

The above problem is convex b'cos from part 1 above

$$
J(\theta) = \frac{1}{2} \sum_{i=1}^{n} w_i (\theta^T x_i - y_i)^2 = (X\theta - y)^T W (X\theta - y)
$$

In such form Heissian of J is W, and since W is diagonal with positive entries, its semi definite hence problem is convex.

5.3 part c

5.4 part d

6 Question 6: Neural Networks and Back Propogation