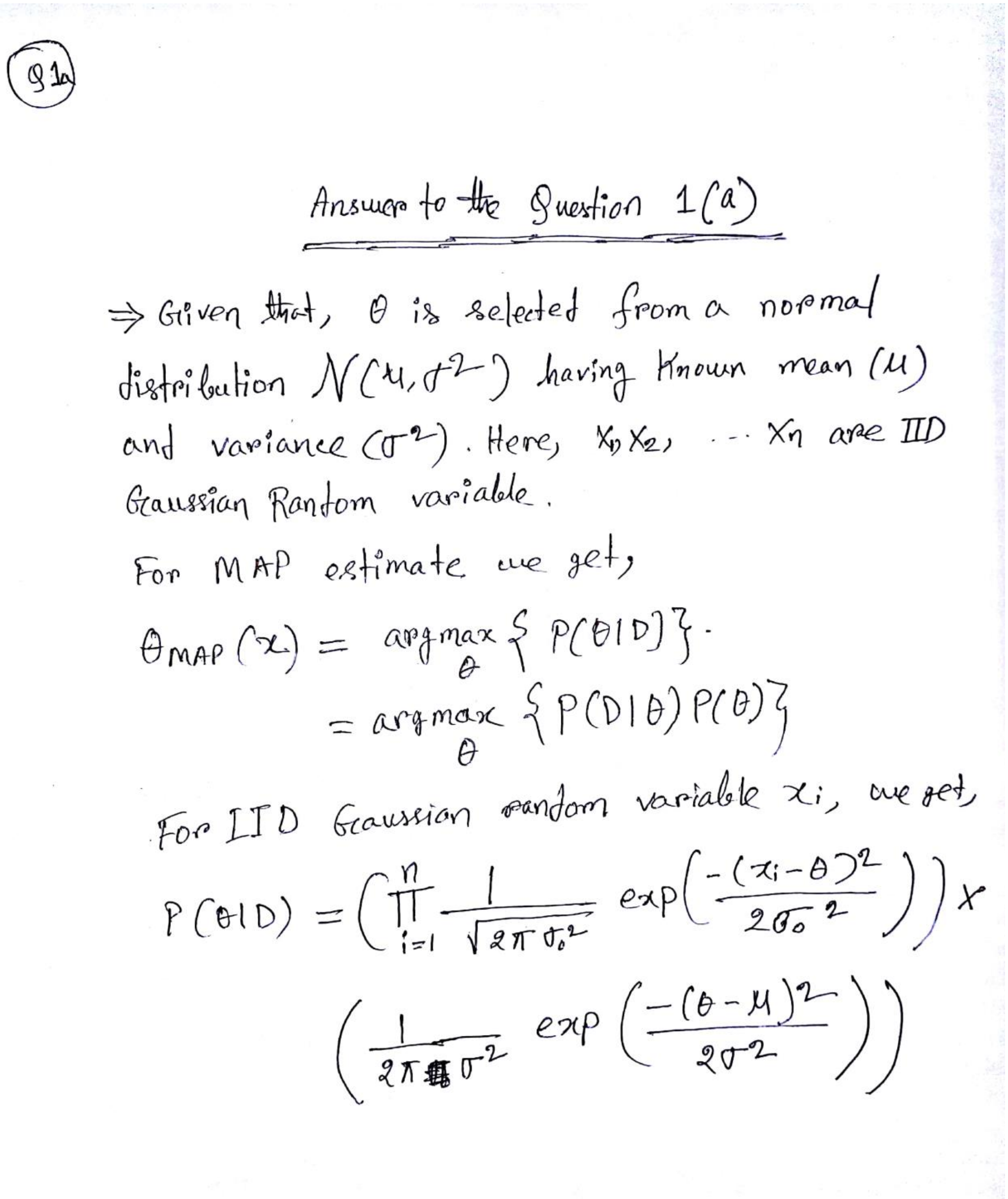
CMPUT - 566

Introduction to Machine Learning

Assignment – 2

Jakaria Rabbi



Q1a

Answer to the Question 1(a)

⇒ Given that, θ is selected from a normal distribution $N(\mu, \sigma^2)$ having known mean (μ) and variance (σ^2). Here, $X_1, X_2, \dots X_n$ are IID Gaussian Random variable.

For MAP estimate we get,

$$\theta_{MAP}(x) = \underset{\theta}{\operatorname{argmax}} \{ P(\theta|D) \}.$$

$$= \underset{\theta}{\operatorname{argmax}} \{ P(D|\theta) P(\theta) \}$$

For IID Gaussian random variable x_i, we get,

$$P(\theta|D) = \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-(x_i-\theta)^2}{2\sigma_0^2} \right) \right) \times$$

$$\left(\frac{1}{2\pi\sigma^2} \exp\left(\frac{-(\theta-\mu)^2}{2\sigma^2} \right) \right)$$

Q1a

$$\Rightarrow \log P(\theta|D) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-(x_i-\theta)^2}{2\sigma_0^2}\right) +$$

$$\log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\theta-\mu)^2}{(2\sigma^2}\right)$$

$$= \sum_{i=1}^{n} \frac{-(x_i-\theta)^2}{2\sigma_0^2} - \log\sqrt{2\pi\sigma_0^2} -$$

$$\frac{(\theta-\mu)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$$

$$= -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i-\theta)^2 - \log\sqrt{2\pi\sigma_0^2} -$$

$$\frac{(\theta-\mu)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$$

Differentiating this with respect to θ and setting it to 0, we get,

$$\frac{dLL}{d\theta} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (x_i-\theta) - \frac{1}{\sigma^2}(\theta-\mu) = 0$$

$$\Rightarrow \frac{\theta-\mu}{\sigma^2} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (x_i-\theta)$$

$$\Rightarrow \frac{\theta-\mu}{\sigma^2} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} x_i - \frac{n\theta}{\sigma_0^2}$$

Q1a Q1b

$$\Rightarrow \frac{\theta-\mu}{\sigma^2}+\frac{n\theta}{\sigma_0^2}=\frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \frac{\theta}{\sigma^2}+\frac{n\theta}{\sigma_0^2}=\frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_i+\frac{\mu}{\sigma^2}$$

$$\Rightarrow \theta\left(\frac{1}{\sigma^2}+\frac{n}{\sigma_0^2}\right)=\frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_i+\frac{\mu}{\sigma^2}$$

$$\Rightarrow \theta=\frac{\frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_i+\frac{\mu}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{n}{\sigma_0^2}}$$

⇒ Ans:/

Ans. to the Question 1 (b)

⇒ Here, θ is selected from a Laplace distribution $\mathcal{L}(\mu, b)$ and we have

$$p(x)=\frac{1}{2b}\exp\left(\frac{-|x-\mu|}{b}\right)$$

For simplicity $\mu=0$

Hence, $$P(\theta)=\frac{1}{2b}\exp\left(\frac{-|\theta|}{b}\right)$$

Q1b

For map estimate, we get,

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \{ P(\theta|D) \}$$

$$= \underset{\theta}{\operatorname{argmax}} \{ P(D|\theta) P(\theta) \}$$

Here,

$$P(\theta|D) = \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-(x_i-\theta)^2}{2\sigma_0^2} \right) \right) \times$$

$$\left(\frac{1}{2b} \exp \left(\frac{-|\theta|}{b} \right) \right.$$

$$\Rightarrow \log P(\theta|D) = \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \right) - \sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2\sigma_0^2} +$$

$$\log\left(\frac{1}{2b}\right) - \frac{|\theta|}{b}$$

Differentiating the above equation with respect to θ and setting it to 0, we get,

$$\frac{dLL}{d\theta} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (x_i - \theta) - \frac{1}{b} \frac{d|\theta|}{d\theta} = 0$$

Hence ~~θ_{MAP}~~ $\Rightarrow \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (x_i - \theta) - \frac{1}{b} \frac{d|\theta|}{d\theta} = 0$

Q1b

For finding θ_{MAP}, here we don't get any closed form solution. We can not solve θ for laplace distribution as it provides a stationary point.

In this case, we can use iterative approach to find the solution.

the gradient $\frac{d|\theta|}{d\theta} = \frac{b}{\sigma_0^2} \sum_{i=1}^{n} x_i - n\theta$

can be used to find out new θ value by subtracting this gradient from current θ at each step of the iteration.

We also can use proximal method along with gradient descent.

Q1c

Ans. to que. no. 1(c)

we know that,

$$\theta_{MAP} = \underset{\theta}{\arg\max}\{P(\theta|D)\}$$

$$= \underset{\theta}{\arg\max}\{P(D|\theta)\,P(\theta)\}$$

For multivariate Gaussian we get,

$$\theta_{MAP} = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d|\Sigma_0|}} \exp\left(-\frac{1}{2}(x_i-\theta)^T\Sigma_0^{-1}(x_i-\theta)\right)$$

$$\times \cancel{\prod_{i=1}^{n}} \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\theta)^T\Sigma^{-1}(\theta)\right)$$

$$\log(\theta_{MAD}) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{(2\pi)^d|\Sigma_0|}} \exp\left(-\frac{1}{2}(x_i-\theta)^T\Sigma_0^{-1}(x_i-\theta)\right)$$

$$+ \log \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\theta)^T\Sigma^{-1}(\theta)\right)$$

Q1c

Here, $\Sigma_0 = I$ and $\Sigma = \sigma^2 I$,

Hence, we get,

$$\theta_{MAP} = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2} (x_i-\theta)^T I (x_i-\theta)\right)$$

$$\times \frac{1}{\sqrt{(2\pi)^d \sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (\theta)^T I (\theta)\right)$$

Here, $$\log(\theta_{MAP}) = \log\left(\frac{1}{\sqrt{(2\pi)^d}}\right) - \sum_{i=1}^{n} \frac{1}{2} (x_i-\theta)^T I (x_i-\theta)$$

$$+ \log\left(\frac{1}{\sqrt{(2\pi)^d \sigma^2}}\right) - \frac{1}{2\sigma^2} (\theta)^T I (\theta)$$

Differentiating with respect to θ and setting it to 0, we get,

$$\frac{dLL}{d\theta} = \frac{1}{2} \left[(I + I^T) \right] \sum_{i=1}^{n} (x_i - \theta)$$

$$- \frac{1}{2\sigma^2} \left[I + I^T \right] (\theta) = 0$$

Q1e

$$\Rightarrow \sum_{i=1}^{n} (x_i - \theta) = \frac{1}{\sigma^2}(\theta)$$

$$\Rightarrow \frac{1}{\sigma^2}(\theta) + n\theta = \sum_{i=1}^{n} x_i$$

$$\Rightarrow \theta = \frac{\sigma^{-2} \sum_{i=1}^{n} x_i}{1 + n\sigma^{-2}}$$

Ans./

Answer to the Question: 2(a)

If we run the linear regression with full feature set, we get singular matrix error due to the singularity of (X^TX). This error generally occurs when there are redundant features in the feature set and there is lack of linear independence between columns of the feature set. Also, if there is near linear dependence between the columns, this error generally occurs. The determinant of the matrix will become zero in that case. Here, we are trying to get a closed form solution and in that case, we need the calculate inverse of the feature matrix which is not possible for singular matrix.

We can easily solve this problem by using pseudo-inverse of the feature matrix. We can also add regularization term to the regression algorithm such as Laplace or Ridge regularizer, which will shift or truncate the small singular values which cause numerical stabilities. Another way is to use iterative approaches such as Batch or stochastic gradient descent.

Answer to the Question: 2(b)

Standard Error is implemented in the main method.

Answer to the Question: 2(c)

The Ridge regression is implemented as *RidgeLinearRegression* class in*regressionalgorithms.py* file.
The results for regularization parameter **λ** = 0.01 are following:

1. Regularizer prevent the singular matrix error.
2. **Average Test Error: 41.77 Average Standard Error: 0.55**

Answer to the Question: 2(d)

The Lasso regression is implemented as *LassoRegression* class in*regressionalgorithms.py* file.

Average Test Error: 42.46 Average Standard Error: 0.078

Answer to the Question: 2(e)

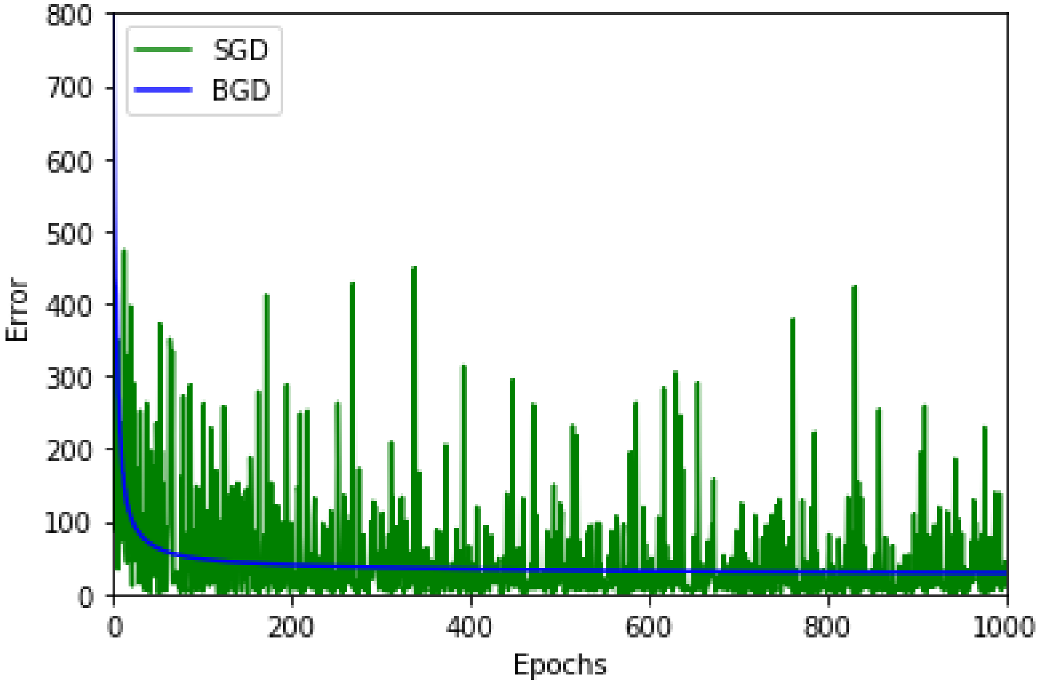
The Stochastic Gradient Descent is implemented as *StochasticGradientDescent* class in*regressionalgorithms.py* file. Process the whole data on 1000 times. After 1000 Epoch:

Average Test Error: 42.82 Average Standard Error: 0.04

Answer to the Question: 2(f)

The Batch Gradient Descent is implemented as *BatchGradientDescent* class in*regressionalgorithms.py* file. Process the whole data on average 2600 times.

Average Test Error: 41.35 Average Standard Error: 0.31



Answer to the Question: Bonus(a)

The Stochastic Gradient with RMSPROP is implemented as *StochasticGradientWithRMSPROP* class in*regressionalgorithms.py* file. After 1000 Epoch:

Average Test Error: 47.94 **Average Standard Error: 1.52**

Answer to the Question: Bonus(b)

The Stochastic Gradient with AMSGRAD is implemented as *StochasticGradientWithAMSGRAD* class in*regressionalgorithms.py* file. After 1000 Epoch:

Average Test Error: 46.39 **Average Standard Error: 1.4**