CSC311 Homework 3

Xuanqi Wei
1009353209

November 10, 2024

1 Question 1

(a)

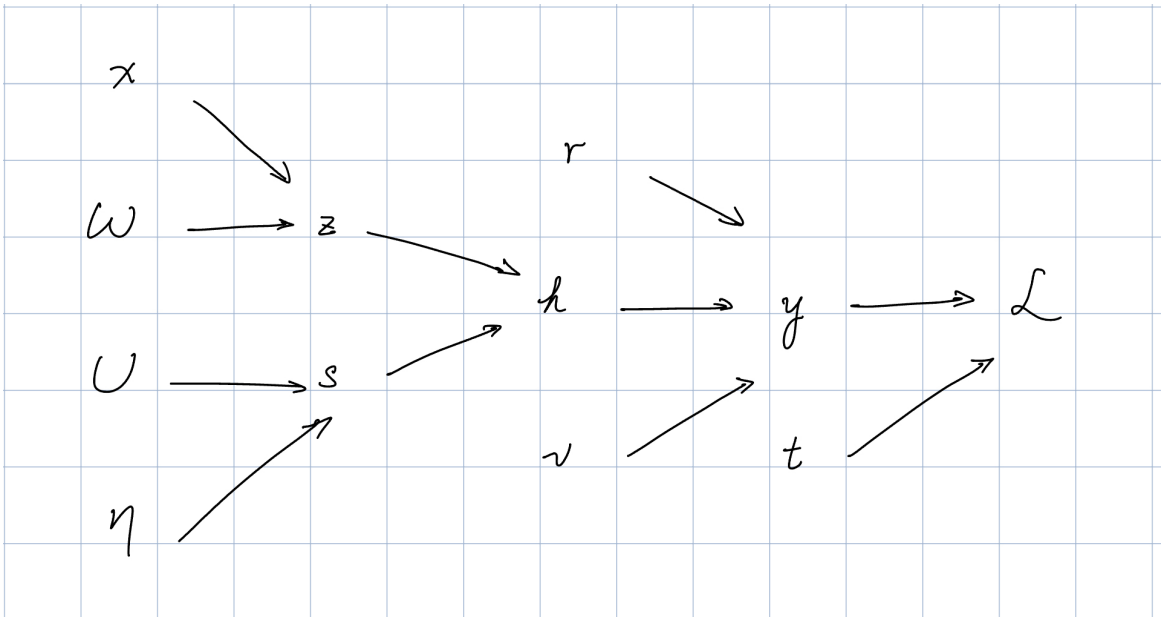

Figure 1: The computation graph

(b)

$$\frac{d\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) \tag{1}$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} \tag{2}$$

$$= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) \tag{3}$$

$$= \sigma(x)(1-\sigma(x)) \tag{4}$$

(c)

$$\bar{\mathcal{L}} = 1 \tag{5}$$

$$\bar{y} = \bar{\mathcal{L}}\frac{d\mathcal{L}}{dy} = \bar{\mathcal{L}}\frac{d}{dy}\left(t\log y + (1-t)\log(1-y)\right) \tag{6}$$

$$= \bar{\mathcal{L}}\left(\frac{t}{y} - \frac{1-t}{1-y}\right) \tag{7}$$

$$\bar{h}_i = \bar{y}\frac{dy}{dh_i} = \bar{y}\frac{dy}{d(v^T h + r^T x)}\frac{d(v^T h + r^T x)}{dh_i} \tag{8}$$

$$= \bar{y}\sigma(v^T h + r^T x)(1-\sigma(v^T h + r^T x))v_i \tag{9}$$

$$\bar{v}_i = \bar{y}\frac{dy}{dv_i} = \bar{y}\frac{dy}{d(v^T h + r^T x)}\frac{d(v^T h + r^T x)}{dv_i} \tag{10}$$
$$= \bar{y}\sigma(v^T h + r^T x)(1 - \sigma(v^T h + r^T x))h_i \tag{11}$$

$$\bar{r}_i = \bar{y}\frac{dy}{dr_i} = \bar{y}\frac{dy}{d(v^T h + r^T x)}\frac{d(v^T h + r^T x)}{dr_i} \tag{12}$$
$$= \bar{y}\sigma(v^T h + r^T x)(1 - \sigma(v^T h + r^T x))x_i \tag{13}$$

$$\bar{s}_i = \sum_j \bar{h}_j \frac{dh_j}{ds_i} = \bar{h}_j z_i \tag{14}$$

$$\bar{z}_i = \sum_j \bar{s}_j \frac{ds_j}{dz_i} = \bar{h}_j s_i \tag{15}$$

$$\bar{U}_{ij} = \sum_k \bar{s}_k \frac{ds_k}{dU_{ij}} = s_i \eta_j \tag{16}$$

$$\bar{w}_{ij} = \bar{z}_i \frac{dz_i}{dw_{ij}} = \bar{z}_i x_j \tag{17}$$

$$\bar{\eta}_j = \sum_i \bar{s}_i \frac{ds_i}{d\eta_j} = \sum_i \bar{s}_i U_{ij} \tag{18}$$

$$\bar{x}_j = \sum_i \bar{z}_i \frac{dz_i}{dx_j} + \bar{y}\frac{dy}{dx_j} \tag{19}$$
$$= \sum_i \bar{z}_i w_{ij} + \bar{y}\sigma(v^T h + r^T x)(1 - \sigma(v^T h + r^T x))r_j \tag{20}$$

2 Question 2

(a) Given the joint probability specified in the problem, we can compute the log-likelihood:

$$l(\theta, \pi) = \sum_{i=1}^{N} \log p(x^{(i)}, c^{(i)}|\theta, \pi) \tag{21}$$

$$= \sum_{i=1}^{N} \log p(c^{(i)}|\theta, \pi) p(x^{(i)}|c^{(i)}, \theta, \pi) \tag{22}$$

$$= \sum_{i=1}^{N} \log p(c^{(i)}|\pi) \prod_{j=1}^{784} p(x_j^{(i)}|c^{(i)}, \theta_{jc}) \tag{23}$$

$$= \sum_{i=1}^{N} \left[\log p(c^{(i)}|\pi) + \sum_{j=1}^{784} \log p(x_j^{(i)}|c^{(i)}, \theta_{jc}) \right] \tag{24}$$

$$= \sum_{i=1}^{N} \log p(c^{(i)}|\pi) + \sum_{i=1}^{N} \sum_{j=1}^{784} \log p(x_j^{(i)}|c^{(i)}, \theta_{jc}). \tag{25}$$

I'll maximize the log-likelihood by maximizing each part. To Maximize $\sum_{i=1}^{N} \log p(c^{(i)}|\pi)$, by the categorical distribution:

$$\sum_{i=1}^{N} \log p(c^{(i)}|\pi) = \sum_{i=1}^{N} \log \prod_{j=0}^{9} \pi_j^{t_j^{(i)}} \tag{26}$$

$$= \sum_{i=1}^{N} \sum_{j=0}^{9} \log \pi_j^{t_j^{(i)}} = \sum_{i=1}^{N} \sum_{j=0}^{9} t_j^{(i)} \log \pi_j \tag{27}$$

$$= \sum_{i=1}^{N} \left(\sum_{j=0}^{8} t_j^{(i)} \log \pi_j + t_9^{(i)} \log \left(1 - \sum_{j=0}^{8} \pi_j \right) \right). \tag{28}$$

To maximize (28), set the derivative with respect to π_j to zero:

$$\frac{\partial}{\partial \pi_j} \sum_{i=1}^{N} \log p(c^{(i)}|\pi) = 0 \tag{29}$$

$$\implies \sum_{i=1}^{N} \frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{1 - \sum_{j=0}^{8} \pi_j} = \sum_{i=1}^{N} \frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{\pi_9} = 0 \tag{30}$$

$$\implies \sum_{i=1}^{N} \frac{t_j^{(i)}}{\pi_j} = \sum_{i=1}^{N} \frac{t_9^{(i)}}{\pi_9} \tag{31}$$

$$\implies \frac{\sum_{i=1}^{N} t_j^{(i)}}{\pi_j} = \frac{\sum_{i=1}^{N} t_9^{(i)}}{\pi_9} \tag{32}$$

$$\implies \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} = \frac{\hat{\pi}_j}{\hat{\pi}_9} \tag{33}$$

$$\implies \hat{\pi}_j = \hat{\pi}_9 \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}}. \tag{34}$$

Since $\sum_{j=0}^{9} \pi_j = 1$, we have:

$$1 = \sum_{j=0}^{9} \hat{\pi}_j = \sum_{j=0}^{8} \hat{\pi}_j + \hat{\pi}_9 \tag{35}$$

$$= \sum_{j=0}^{8} \left(\hat{\pi}_9 \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} \right) + \hat{\pi}_9 \frac{\sum_{i=1}^{N} t_9^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} \tag{36}$$

$$= \frac{\hat{\pi}_9}{\sum_{i=1}^{N} t_9^{(i)}} \left(\sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)} + \sum_{i=1}^{N} t_9^{(i)} \right) \tag{37}$$

$$= \frac{\hat{\pi}_9}{\sum_{i=1}^{N} t_9^{(i)}} \sum_{j=0}^{9} \sum_{i=1}^{N} t_j^{(i)} = \frac{\hat{\pi}_9 \cdot N}{\sum_{i=1}^{N} t_9^{(i)}} \tag{38}$$

$$\implies \hat{\pi}_9 = \frac{\sum_{i=1}^{N} t_9^{(i)}}{N} \tag{39}$$

Thus, from (34), we know that

$$\hat{\pi}_j = \frac{\sum_{i=1}^{N} t_j^{(i)}}{N}, \quad \text{for } j = 0, 1, \ldots, 9, \tag{40}$$

where the numerator $\sum_{i=1}^{N} t_j^{(i)}$ is the number of samples in class j and the denominator N represents the total number of samples.

For the second part, to maximize $\sum_{i=1}^{N} \sum_{j=1}^{784} \log p(x_j^{(i)} | c^{(i)}, \theta_{jc})$, by the Bernoulli distribution:

$$\sum_{i=1}^{N} \sum_{j=1}^{784} \log p(x_j^{(i)} | c^{(i)}, \theta_{jc}) = \sum_{i=1}^{N} \sum_{j=1}^{784} \mathbb{I}(c^{(i)} = c) \left[x_j^{(i)} \log \theta_{jc} + (1 - x_j^{(i)}) \log(1 - \theta_{jc}) \right]. \tag{41}$$

To maximize (41), setting the derivative with respect to θ_{jc} to zero:

$$\frac{\partial}{\partial \theta_{jc}} \sum_{i=1}^{N} \sum_{j=1}^{784} \mathbb{I}(c^{(i)} = c) \left[x_j^{(i)} \log \theta_{jc} + (1 - x_j^{(i)}) \log(1 - \theta_{jc}) \right] = 0 \tag{42}$$

$$\implies \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c) \left[\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1 - x_j^{(i)}}{1 - \theta_{jc}} \right] = 0 \tag{43}$$

$$\implies \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c) \frac{x_j^{(i)}}{\theta_{jc}} = \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c) \frac{1 - x_j^{(i)}}{1 - \theta_{jc}} \tag{44}$$

$$\implies \frac{\sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c) x_j^{(i)}}{\theta_{jc}} = \frac{\sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c)(1 - x_j^{(i)})}{1 - \theta_{jc}} \tag{45}$$

$$\implies \theta_{jc} = \frac{\sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c) x_j^{(i)}}{\sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c)} \tag{46}$$

$$\implies \hat{\theta}_{jc} = \frac{\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\sum_{i=1}^{N} t_c^{(i)}}. \tag{47}$$

Therefore, for $\hat{\theta}_{jc}$, the numerator $\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}$ represents the sum of feature values for class c and the denominator $\sum_{i=1}^{N} t_c^{(i)}$ is the total number of samples in class c.

(b) For $p(t|x, \theta, \pi)$, we know,

$$p(t|x,\theta,\pi) = \frac{p(t,x|\theta,\pi)}{p(x|\theta,\pi)} = \frac{p(t,x|\theta,\pi)}{\sum_{j=0}^{9} p(t_j = 1, x|\theta,\pi)} \tag{48}$$

$$= \frac{p(t,x|\theta,\pi)}{\sum_{i=0}^{9} p(t_i = 1|\pi) \prod_{j=1}^{784} p(x_j|t_i = 1, \theta_{jc})} \tag{49}$$

$$= \frac{p(t,x|\theta,\pi)}{\sum_{i=0}^{9} (\pi_i \prod_{j=1}^{784} \theta_{ij}^{x_j} (1-\theta_{ij})^{1-x_j})}. \tag{50}$$

Thus, from (50), taking the log gives,

$$log p(t|x,\theta,\pi) = \log p(t,x|\theta,\pi) - \log \sum_{i=0}^{9} \left(\pi_i \prod_{j=1}^{784} \theta_{ij}^{x_j} (1-\theta_{ij})^{1-x_j} \right) \tag{51}$$

$$= \sum_{i=0}^{9} t_i \log \pi_i + \sum_{j=1}^{784} (x_j \log \theta_{jc} + (1-x_j) \log(1-\theta_{jc})) \tag{52}$$

$$- \log \sum_{i=0}^{9} \left(\pi_i \prod_{j=1}^{784} \theta_{ij}^{x_j} (1-\theta_{ij})^{1-x_j} \right). \tag{53}$$

(c)

```
Average log-likelihood for MLE is  nan
```

Figure 2: MLE Average Log-likelihood

By fitting the parameters and using equation 1, the average log-likelihood is undefined. The reason for that is some classes have a likelihood ($\hat{\theta_{jc}}$) of 0 but taking the log of 0 will result in nan.

(d)

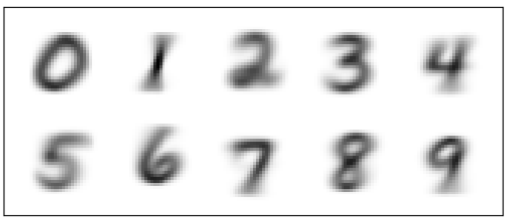

Figure 3: The MLE estimator $\hat{\theta}$

(e) By the Beta prior we learnt, we know,

$$p(\theta_{jc}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta_{jc}^{a-1}(1-\theta_{jc})^{b-1} \propto \theta_{jc}^{a-1}(1-\theta_{jc})^{b-1}. \tag{54}$$

To maximize the prior, we will maximize the log-prior, which the log-prior from (54) is given by,

$$\log p(\theta_{jc}) = \log \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} + (a-1)\log\theta_{jc} + (b-1)\log(1-\theta_{jc}). \tag{55}$$

Thus, we know that,

$$\log p(\theta_{jc}) \propto (a-1)\log\theta_{jc} + (b-1)\log(1-\theta_{jc}) \tag{56}$$

$$\implies \log p(\theta_{jc}|D) \propto (a-1)\log\theta_{jc} + (b-1)\log(1-\theta_{jc}) + \sum_{i=1}^{N} t_c^{(i)}\left[x_j^{(i)}\log\theta_{jc} + (1-x_j^{(i)})\log(1-\theta_{jc})\right]. \tag{57}$$

Thus, taking the log with repsect to θ_{jc} for (55) will give,

$$\frac{d\log p(\theta_{jc})}{d\theta_{jc}} = 0 \tag{58}$$

$$\implies \frac{a-1}{\theta_{jc}} - \frac{b-1}{1-\theta_{jc}} + \sum_{i=1}^{N} t_c^{(i)}\frac{x_j^{(i)}}{\theta_{jc}} - \sum_{i=1}^{N} t_c^{(i)}\frac{1-x_j^{(i)}}{1-\theta_{jc}} = 0 \tag{59}$$

$$\implies \frac{a-1}{\theta_{jc}} + \sum_{i=1}^{N} t_c^{(i)}\frac{x_j^{(i)}}{\theta_{jc}} = \frac{b-1}{1-\theta_{jc}} + \sum_{i=1}^{N} t_c^{(i)}\frac{1-x_j^{(i)}}{1-\theta_{jc}} \tag{60}$$

$$\implies \frac{a-1+\sum_{i=1}^{N} t_c^{(i)}x_j^{(i)}}{\theta_{jc}} = \frac{b-1+\sum_{i=1}^{N} t_c^{(i)}(1-x_j^{(i)})}{1-\theta_{jc}} \tag{61}$$

$$\implies \hat{\theta}_{jc_{\text{MAP}}} = \frac{a-1+\sum_{i=1}^{N} t_c^{(i)}x_j^{(i)}}{a+b-2+\sum_{i=1}^{N} t_c^{(i)}}. \tag{62}$$

Evaluate when $\alpha = 3$, $\beta = 3$,

$$\hat{\theta}_{jc_{\text{MAP}}} = \frac{3 - 1 + \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{3 + 3 - 2 + \sum_{i=1}^{N} t_c^{(i)}} = \frac{2 + \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{4 + \sum_{i=1}^{N} t_c^{(i)}}.$$

Comparing to the MLE estimator, we observer two constants based on the α and β values added to the numerator and denominator in the MAP estimator, 2 and 4 respectively, where the rest keeps the same.

(f)

```
Average log-likelihood for MAP is  -3.357063137860285
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

Figure 4: MAP Average log-likelihood, MAP Training Accuracy, and MAP Test Accuracy

(g)

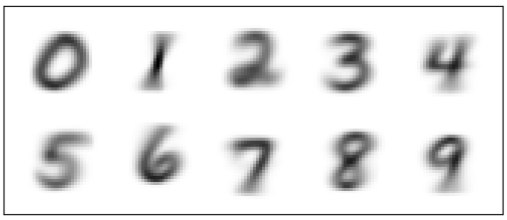

Figure 5: The MAP estimator $\hat{\theta}$

(h) For the Naive Bayes Algorithm,

- Advantage: It's computationally efficient for both training and testing, which can be simply implemented.
- Why might note be reasonable: The Naive Bayes Algorithm make assumptions on the feature being conditionally independent given the class, which is not guaranteed in reality.

3 Question 3

(a) Given $p(y=1|\mathbf{x},\theta) = \frac{1}{1+\exp(-\mathbf{x}^T\theta)} = \sigma(x^T\theta)$, the likelihood function:

$$L(\theta) = \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)},\theta) \tag{63}$$

$$= \prod_{i=1}^{N} \sigma(\mathbf{x}^{(i)T}\theta)^{y^{(i)}} \left(1-\sigma(\mathbf{x}^{(i)T}\theta)\right)^{1-y^{(i)}}. \tag{64}$$

Taking the log, we get the log-likelihood function:

$$\mathcal{L}(\theta) = \log L(\theta) \tag{65}$$

$$= \sum_{i=1}^{N} \left[y^{(i)} \log \sigma(\mathbf{x}^{(i)T}\theta) + (1-y^{(i)}) \log(1-\sigma(\mathbf{x}^{(i)T}\theta))\right]. \tag{66}$$

Thus, since $\sigma(x) = \frac{1}{1+e^{-x}}$, deriving with respect to θ, gives,

$$\frac{d\mathcal{L}(\theta)}{d\theta} \tag{67}$$

$$= \sum_{i=1}^{N} \left[y^{(i)} \frac{1}{\sigma(\mathbf{x}^{(i)T}\theta)} \sigma(\mathbf{x}^{(i)T}\theta)(1-\sigma(\mathbf{x}^{(i)T}\theta))\mathbf{x}^{(i)} - (1-y^{(i)}) \frac{1}{1-\sigma(\mathbf{x}^{(i)T}\theta)} \sigma(\mathbf{x}^{(i)T}\theta)(1-\sigma(\mathbf{x}^{(i)T}\theta))\mathbf{x}^{(i)} \right] \tag{68}$$

$$= \sum_{i=1}^{N} \left[y^{(i)}(1-\sigma(\mathbf{x}^{(i)T}\theta))\mathbf{x}^{(i)} - (1-y^{(i)})\sigma(\mathbf{x}^{(i)T}\theta)\mathbf{x}^{(i)}\right] \tag{69}$$

$$= \sum_{i=1}^{N} \left[y^{(i)}\mathbf{x}^{(i)} - \sigma(\mathbf{x}^{(i)T}\theta)\mathbf{x}^{(i)}\right]. \tag{70}$$

From the derivation above, the log-likelihood would be optimized using,

$$\theta \leftarrow \theta + \alpha \frac{d\mathcal{L}(\theta)}{d\theta} = \theta + \alpha \sum_{i=1}^{N} \left[y^{(i)}\mathbf{x}^{(i)} - \sigma(\mathbf{x}^{(i)T}\theta)\mathbf{x}^{(i)}\right].$$

(b) Given that $p(\theta) \sim \mathcal{N}(0, \sigma^2 I)$, from the property of normal distribution, we know that, the prior, $p(\theta) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\theta^T\theta\right)$.

Thus, the posterior is given by $p(\theta|D) \propto p(D|\theta)p(\theta)$. I'll maximize the posterior by maximizing the log-posterior, which,

$$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta)$$

.

Given the log-posterior, we have,

$$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) = \log \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)}, \theta) + \log p(\theta) \tag{71}$$

$$= \sum_{i=1}^{N} \left[y^{(i)} \log \sigma(\mathbf{x}^{(i)T}\theta) + (1 - y^{(i)}) \log(1 - \sigma(\mathbf{x}^{(i)T}\theta)) \right] + \log \left(\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \theta^T \theta \right) \right) \tag{72}$$

$$= \sum_{i=1}^{N} \left[y^{(i)} \log \sigma(\mathbf{x}^{(i)T}\theta) + (1 - y^{(i)}) \log(1 - \sigma(\mathbf{x}^{(i)T}\theta)) \right] - \frac{d}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \theta^T \theta. \tag{73}$$

Taking the derivative with respect to θ, gives

$$\frac{d \log p(\theta|D)}{d\theta} = \sum_{i=1}^{N} \left[y^{(i)} \mathbf{x}^{(i)} - \sigma(\mathbf{x}^{(i)T}\theta) \mathbf{x}^{(i)} \right] - \frac{1}{\sigma^2} \theta. \tag{74}$$

Therefore, from the derivation above, the log-posterior would be optimized using

$$\theta \leftarrow \theta + \alpha \frac{d \log p(\theta|D)}{d\theta}) = \theta + \alpha \left[\sum_{i=1}^{N} \left(y^{(i)} \mathbf{x}^{(i)} - \sigma(\mathbf{x}^{(i)T}\theta) \mathbf{x}^{(i)} \right) - \frac{1}{\sigma^2} \theta \right].$$

4 Question 4

(a)

```
The average conditional log-likelihood for the training dataset: -0.125.
The average conditional log-likelihood for the test dataset: -0.197.
```

Figure 6: The average conditional likelihood

(b)

```
The accuracy for the train dataset: 0.981.
The accuracy for the test dataset: 0.973.
```

Figure 7: The accuracies for the train and test dataset

(c)

```
The average conditional log-likelihood for the training dataset by assuming Diagonal-Covariance Matrix: -1.23.
The average conditional log-likelihood for the test datasetby assuming Diagonal-Covariance Matrix: -1.29.
The accuracy for the training dataset by assuming Diagonal-Covariance Matrix: 0.850.
The accuracy for the test dataset by assuming Diagonal-Covariance Matrix: 0.840.
```

Figure 8: Result under diagonal covariance matrix

Comparing (c) and (a)(b), I find that the average conditional log-likelihoods for a diagonal covariance matrix are lower. Since the diagonal covariance matrix can't model dependence between pixel, comparing to a full covariance matrix, the ability to capture the data is lower; and, when assuming a diagonal covariance matrix, the model is also less flexible. The average accuracies are also lower due to the two reasons stated.