1. By Bayes' rule and the law of total probability,

$$\begin{aligned}
p(y=k|\mathbf{x},\boldsymbol{\mu},\boldsymbol{\sigma}) &= \frac{p(\mathbf{x}|y=k,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=k|\boldsymbol{\mu},\boldsymbol{\sigma})}{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\sigma})} \\
&= \frac{p(\mathbf{x}|y=k,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=k|\boldsymbol{\mu},\boldsymbol{\sigma})}{\sum_{j=1}^{k} p(\mathbf{x}|y=k,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=k|\boldsymbol{\mu},\boldsymbol{\sigma})} \\
&= \frac{\left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-\frac{1}{2}} \exp\left(-\sum_{i=1}^{D} \frac{1}{\sigma_i^2}(x_i-\mu_{ki})^2\right)\alpha_k}{\sum_{j=1}^{k}\left[\left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-\frac{1}{2}} \exp\left(-\sum_{i=1}^{D} \frac{1}{\sigma_i^2}(x_i-\mu_{ji})^2\right)\alpha_j\right]}
\end{aligned}$$

b) Since the data points are i.i.d,

$$\begin{aligned}
\ell(\boldsymbol{\theta};D) &= -\log p(y^{(1)},\mathbf{x}^{(1)},\ldots,y^{(N)},\mathbf{x}^{(N)}|\boldsymbol{\theta}) \\
&= -\log\left(\prod_{j=1}^{N} p(y^{(j)},\mathbf{x}^{(j)}|\boldsymbol{\theta})\right) \\
&= -\sum_{j=1}^{N} \log p(y^{(j)},\mathbf{x}^{(j)}|\boldsymbol{\theta}) \\
&= -\sum_{j=1}^{N} \log(p(\mathbf{x}^{(j)}|y=y^{(j)},\boldsymbol{\theta})p(y=y^{(j)}|\boldsymbol{\theta})) \\
&= -\sum_{j=1}^{N} [\log p(\mathbf{x}^{(j)}|y=y^{(j)},\boldsymbol{\theta}) + \log p(y=y^{(j)}|\boldsymbol{\theta})] \\
&= -\sum_{j=1}^{N}\left[\log\left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-\frac{1}{2}} \exp\left(-\sum_{i=1}^{D}\frac{1}{\sigma_i^2}(x_i^{(j)}-\mu_{y^{(j)}i})^2\right) + \log\alpha_{y^{(j)}}\right] \\
&= -\sum_{j=1}^{N}\left[-\frac{1}{2}\left(\sum_{i=1}^{D}\log(2\pi\sigma_i^2)\right) - \left(\sum_{i=1}^{D}\frac{1}{\sigma_i^2}(x_i^{(j)}-\mu_{y^{(j)}i})^2\right) + \log\alpha_{y^{(j)}}\right] \\
&= \sum_{j=1}^{N}\left[\frac{1}{2}\left(\sum_{i=1}^{D}\log(2\pi\sigma_i^2)\right) + \left(\sum_{i=1}^{D}\frac{1}{\sigma_i^2}(x_i^{(j)}-\mu_{y^{(j)}i})^2\right) - \log\alpha_{y^{(j)}}\right]
\end{aligned}$$

c) For $1 \leq m \leq k$ and $1 \leq n \leq D$,

$$\begin{aligned}
\frac{d}{d\mu_{mn}}\ell(\boldsymbol{\theta}; D) &= \frac{d}{d\mu_{mn}} \sum_{j=1}^{N} \left[\frac{1}{2} \left(\sum_{i=1}^{D} \log(2\pi\sigma_i^2) \right) + \left(\sum_{i=1}^{D} \frac{1}{\sigma_i^2} (x_i^{(j)} - \mu_{y^{(j)}i})^2 \right) - \log \alpha_{y^{(j)}} \right] \\
&= \frac{1}{2} \left(\sum_{j=1}^{N} \sum_{i=1}^{D} \frac{d}{d\mu_{mn}} \log(2\pi\sigma_i^2) \right) + \left(\sum_{j=1}^{N} \frac{d}{d\mu_{mn}} \frac{1}{\sigma_n^2} (x_n^{(j)} - \mu_{y^{(j)}n})^2 \right) - \sum_{j=1}^{N} \frac{d}{d\mu_{mn}} \log \alpha_{y^{(j)}} \\
&= \frac{1}{2} \left(\sum_{j=1}^{N} \sum_{i=1}^{D} 0 \right) + \left(\sum_{j=1}^{N} \mathbb{I}\{y^{(j)} = m\} \frac{1}{\sigma_n^2} (2)(x_n^{(j)} - \mu_{y^{(j)}n})(-1) \right) - \sum_{j=1}^{N} 0 \\
&= 2 \frac{1}{\sigma_n^2} \sum_{j=1}^{N} \mathbb{I}\{y^{(j)} = m\} (\mu_{y^{(j)}n} - x_n^{(j)})
\end{aligned}$$

For $1 \leq n \leq D$,

$$\begin{aligned}
\frac{d}{d\sigma_n^2}\ell(\boldsymbol{\theta}; D) &= \frac{d}{d\sigma_n^2} \sum_{j=1}^{N} \left[\frac{1}{2} \left(\sum_{i=1}^{D} \log(2\pi\sigma_i^2) \right) + \left(\sum_{i=1}^{D} \frac{1}{\sigma_i^2} (x_i^{(j)} - \mu_{y^{(j)}i})^2 \right) - \log \alpha_{y^{(j)}} \right] \\
&= \frac{1}{2} \left(\sum_{j=1}^{N} \sum_{i=1}^{D} \frac{d}{d\sigma_n^2} \log(2\pi\sigma_i^2) \right) + \left(\sum_{j=1}^{N} \sum_{i=1}^{D} \frac{d}{d\sigma_n^2} \frac{1}{\sigma_i^2} (x_i^{(j)} - \mu_{y^{(j)}i})^2 \right) - \sum_{j=1}^{N} \frac{d}{d\sigma_n^2} \log \alpha_{y^{(j)}} \\
&= \frac{1}{2} \left(\sum_{j=1}^{N} 2\pi \left(\frac{1}{2\pi\sigma_n^2} \right) \right) + \left(\sum_{j=1}^{N} -\frac{1}{\sigma_n^4} (x_n^{(j)} - \mu_{y^{(j)}n})^2 \right) - \sum_{j=1}^{N} 0 \\
&= \frac{1}{2} \left(\sum_{j=1}^{N} \frac{1}{\sigma_n^2} \right) - \left(\sum_{j=1}^{N} \frac{1}{\sigma_n^4} (x_n^{(j)} - \mu_{y^{(j)}n})^2 \right) \\
&= \frac{N}{2\sigma_n^2} - \frac{1}{\sigma_n^4} \left(\sum_{j=1}^{N} (x_n^{(j)} - \mu_{y^{(j)}n})^2 \right)
\end{aligned}$$

d) For $1 \leq m \leq k$ and $1 \leq n \leq D$, setting $\dfrac{d}{d\mu_{mn}}\ell(\boldsymbol{\theta}; D)$ to 0 gives

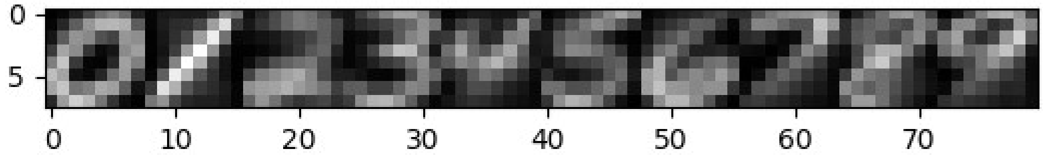
$$\begin{aligned}
2\frac{1}{\sigma_n^2}\sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}(\mu_{y^{(j)}n} - x_n^{(j)}) = 0 \\
\Rightarrow \sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}\mu_{y^{(j)}n} &= \sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}x_n^{(j)} \\
\Rightarrow \mu_{mn}\sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\} &= \sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}x_n^{(j)} \\
\Rightarrow \mu_{mn} &= \frac{\sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}x_n^{(j)}}{\sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}}
\end{aligned}$$

The mth row of $\boldsymbol{\mu}$ is $\dfrac{\sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}\mathbf{x}^{(j)\intercal}}{\sum_{j=1}^{N}\mathbb{I}\{y^{(j)} = m\}}$.

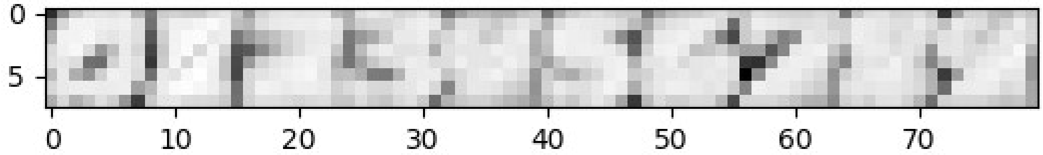
For $1 \leq n \leq D$, setting $\dfrac{d}{d\sigma_n^2}\ell(\boldsymbol{\theta}; D)$ to 0 gives

$$\begin{aligned}
\frac{N}{2\sigma_n^2} - \frac{1}{\sigma_n^4}\left(\sum_{j=1}^{N}(x_n^{(j)} - \mu_{y^{(j)}n})^2\right) = 0 \\
\Rightarrow \frac{N}{2\sigma_n^2} &= \frac{1}{\sigma_n^4}\left(\sum_{j=1}^{N}(x_n^{(j)} - \mu_{y^{(j)}n})^2\right) \\
\Rightarrow \sigma_n^2 &= \frac{2\left(\sum_{j=1}^{N}(x_n^{(j)} - \mu_{y^{(j)}n})^2\right)}{N}
\end{aligned}$$

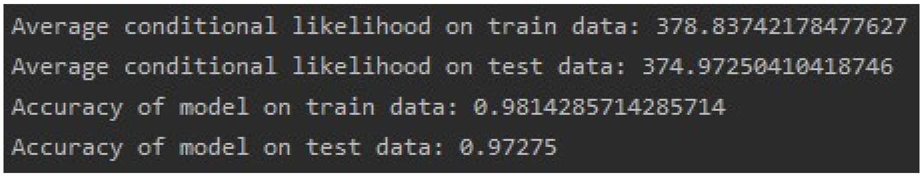
2.0



2.1.1

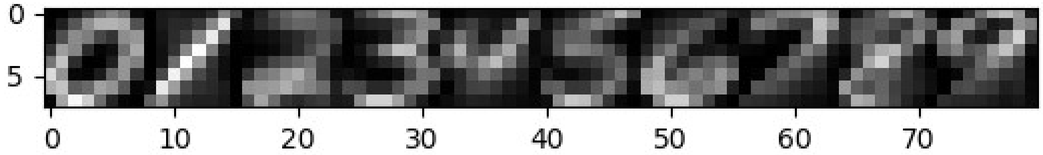


2.1.2/2.1.3

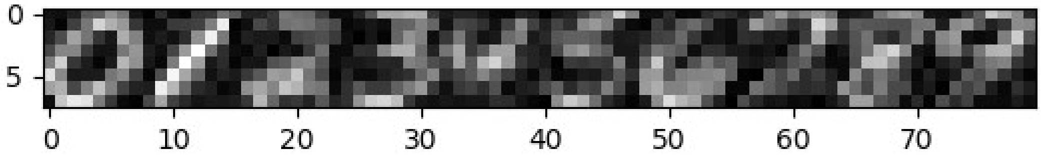


```
Average conditional likelihood on train data: 378.83742178477627
Average conditional likelihood on test data: 374.97250410418746
Accuracy of model on train data: 0.9814285714285714
Accuracy of model on test data: 0.97275
```

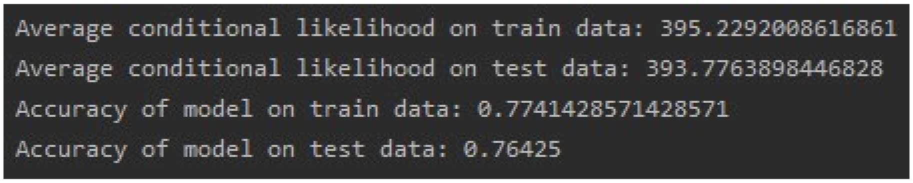
2.2.3



2.2.4



2.2.5/2.2.6



```
Average conditional likelihood on train data: 395.2292008616861
Average conditional likelihood on test data: 393.7763898446828
Accuracy of model on train data: 0.7741428571428571
Accuracy of model on test data: 0.76425
```

2.3 Both models performed much better than chance (10% accuracy). For both models, the accuracy on the test data was slightly less than the accuracy on the training data. The Gaussian classifier is better than the Bayes classifier as while the Gaussian classifier performed with a very low test error rate of 2.725%, the Bayes classifier has a test error rate of 23.575%, approximately 8.7 greater than that of the Gaussian classifier.

These results matched my expectations because the Bayes classifier assumes that the value of a pixel is independent from the value of any other pixel, which often does not hold in the digits data because there are pixel patches where many digits pass through or many digits do not pass through so pixels in these patches are not independent, and this information was not learned by the Bayes classifier.

The Gaussian model performed well because the covariance matrices were computed for each digit, which stores the degree of dependence between any two pixels for each set of images corresponding to a digit, which is beneficial to be learned since the pixels are dependent as mentioned before. The data was also ”natural” in that the structure of the pixel data for each digit class was not artificially made to conform to a distribution (e.g. Bernoulli, uniform) so the inherent normal distributions in such data allow the Gaussian classifier to model it well.