1.

1. Solution:

1). $$p(y=k \mid x, \mu, \sigma) = \frac{p(x \mid y=k, \mu, \sigma)\, p(y=k \mid \mu, \sigma)}{p(x \mid \mu, \sigma)}$$

$$\text{numerator} = \left(\prod_{i=1}^{d} 2\pi\sigma_i^2\right)^{-1/2} \exp\left(-\sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right) \cdot a_k$$

$$\text{denom}: \; p(x \mid \mu, \sigma) = \sum_{j=1}^{k} p(x \mid y=j, \mu, \sigma)\, p(y=j \mid \mu, \sigma)$$

$$= \sum_{j=1}^{k} a_j \left(\prod_{i=1}^{d} 2\pi\sigma_i^2\right)^{-1/2} \exp\left(-\sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i - \mu_{ji})^2\right)$$

$$= \left(\prod_{i=1}^{d} 2\pi\sigma_i^2\right)^{-1/2} \sum_{j=1}^{k} \exp\left(-\sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i - \mu_{ji})^2\right) \cdot a_j$$

$$p(y=k \mid x, \mu, \sigma) = \frac{a_k \exp\left(-\sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right)}{\sum_{j=1}^{k} a_j \cdot \exp\left(-\sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i - \mu_{ji})^2\right)}$$

2). $$p(y^{(i)}, x^{(i)} \mid \theta) = p(x^{(i)} \mid y^{(i)}, \mu, \sigma) \cdot p(y^{(i)} \mid \mu, \sigma) = p(x^{(i)} \mid y^{(i)}, \mu, \sigma) \cdot a_{y^{(i)}}$$

$$L(\theta; D) = p(y^{(1)}, x^{(1)}, \ldots, y^{(N)}, x^{(N)} \mid \theta) = \prod_{j=1}^{N} p(y^{(j)}, x^{(j)} \mid \theta) = \prod_{j=1}^{N} p(x^{(j)} \mid y^{(j)}, \mu, \sigma)\, a_{y^{(j)}}$$

$$= \prod_{j=1}^{N} \left(\prod_{i=1}^{d} 2\pi\sigma_i^2\right)^{-1/2} \exp\left(-\sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i^{(j)} - \mu_{y^{(j)}i})^2\right) \cdot a_{y^{(j)}}$$

$$= \left(\prod_{i=1}^{d} 2\pi\sigma_i^2\right)^{-N/2} \prod_{j=1}^{N} a_{y^{(j)}} \exp\left(-\sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i^{(j)} - \mu_{y^{(j)}i})^2\right)$$

$$l(\theta; D) = -\left[-\frac{N}{2}\left(\sum_{i=1}^{d} \log 2 + \log\pi + 2\log\sigma_i\right) + \sum_{j=1}^{N}\left(\log a_{y^{(j)}} - \sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i^{(j)} - \mu_{y^{(j)}i})^2\right)\right]$$

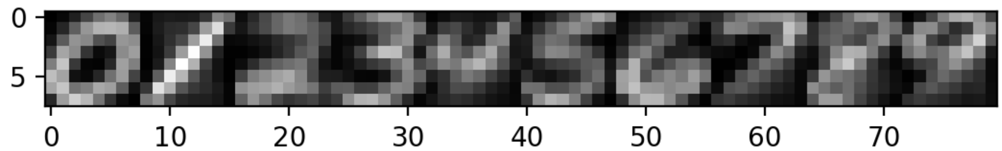
$$= \frac{N}{2}\sum_{i=1}^{d} \log(2\pi\sigma_i^2) - \sum_{j=1}^{N}\left(\log a_{y^{(j)}} - \sum_{i=1}^{d} \frac{1}{2\sigma_i^2}(x_i^{(j)} - \mu_{y^{(j)}i})^2\right)$$

3), 4). $$\frac{\partial l}{\partial \mu_{ki}} = \sum_{j=1}^{N} 1(y^{(j)}=k) \frac{\mu_{ki} - x_i^{(j)}}{\sigma_i^2} = 0 \Rightarrow \hat{\mu}_{ki} = \frac{\sum_{j=1}^{N} 1(y^{(j)}=k)\, x_i^{(j)}}{\sum_{i=1}^{N} 1(y^{(j)}=k)}$$

$$\frac{\partial l}{\partial \sigma_i^2} = \frac{N}{2\sigma_i^2} - \sum_{j=1}^{N} 1(y^{(j)}=k) \frac{(x_i^{(j)} - \mu_{ki})^2}{2(\sigma_i^2)^2} = 0 \Rightarrow N - \sum_{j=1}^{N} 1(y^{(j)}=k) \frac{(x_i^{(j)} - \hat{\mu}_{ki})^2}{\hat{\sigma}_i^2} = 0$$

$$\Rightarrow \hat{\sigma}_i^2 = \frac{\sum_{j=1}^{N} 1(y^{(j)}=k)(x_i^{(j)} - \hat{\mu}_{ki})^2}{N}$$

2.0

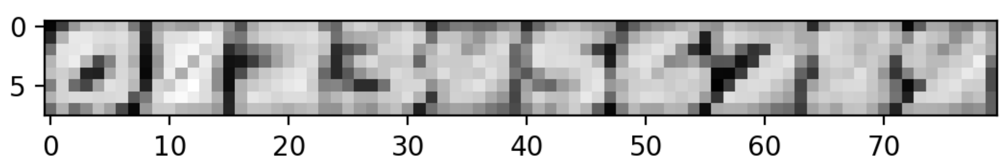


2.1.1 1a. The train classification accuracy for k=1 is 100%.
The test classification accuracy for k = 1 is 96.875%.
1b. The train classification accuracy for k=15 is 96.10%.
The test classification accuracy for k = 15 is 95.925%.

2.1.2 Find the classification by (k-1)-NN recursively until the tie is broken or the classification is given by 1-NN. Since it is impossible to encounter ties in 1-NN, then the recursion will terminate and give out a reasonable result.

2.1.3 Average accuracy for k=1 is 0.964428571429.
Average accuracy for k=2 is 0.964428571429.
Average accuracy for k=3 is 0.965142857143.
Average accuracy for k=4 is 0.964.
Average accuracy for k=5 is 0.962285714286.
Average accuracy for k=6 is 0.961571428571.
Average accuracy for k=7 is 0.959142857143.
Average accuracy for k=8 is 0.958714285714.
Average accuracy for k=9 is 0.957285714286.
Average accuracy for k=10 is 0.955285714286.
Average accuracy for k=11 is 0.953285714286.
Average accuracy for k=12 is 0.952571428571.
Average accuracy for k=13 is 0.951714285714.
Average accuracy for k=14 is 0.951714285714.
Average accuracy for k=15 is 0.949285714286.
The optimal value of k is 3.
The training classification accuracy for k=3 is 0.986571428571.
The test classification accuracy for k=3 is 0.96975.

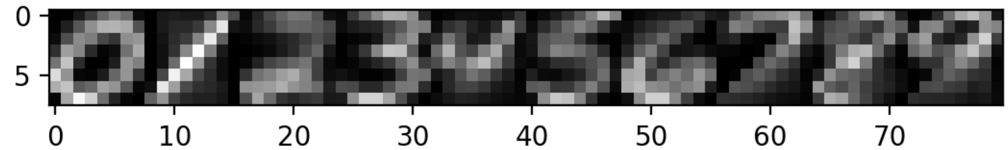
2.2.1



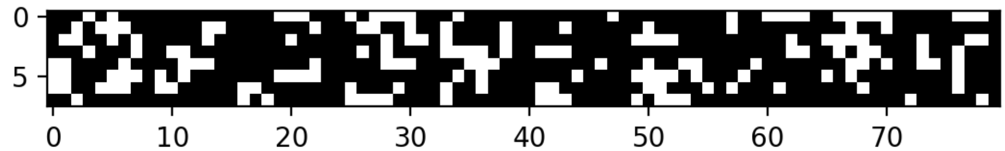
2.2.2 The average conditional likelihood of train data is -0.124587972087, and the exponential of it is 0.882860590861.
The average conditional likelihood of test data is -0.196609089674, and the exponential of it is 0.821511707963.

2.2.3 The accuracy of most likely posterior class on train data is 0.981428571429.
The accuracy of most likely posterior class on test data is 0.97275.

2.3.3



2.3.4



2.3.5 The average conditional likelihood of train data is -0.9437538618, and the exponential of it is 0.38916422127.
The average conditional likelihood of test data is -0.987270433725, and the exponential of it is 0.372592319709.

2.3.6 The accuracy of most likely posterior class on train data is 0.774142857143.
The accuracy of most likely posterior class on test data is 0.76425.

2.4 The k-NN model performs almost as well as the Gaussian Classifier model, while the Bernoulli Naïve Bayes model performs the worst. K-NN performs slightly better on train data and Gaussian Classifier performs slightly better on test data. The difference of performances between Naïve Bayes and the other two can be expected, because Naïve Bayes takes the least time and memory during the training procedure. On the other hand, although we find the optimal k to avoid overfitting, the model is relatively slow. Also, Gaussian Classifier takes much memory space.