1 Class-Conditional Gaussians:

1.)

$\because P(y=k|x,\mu,\sigma) = \frac{P(x|y=k,\mu,\sigma)P(y=k)}{P(x)}$

$\because P(y=k) = \alpha_k$

$\because \sigma^2$ is the shared variance for all features in all classes.

$\therefore P(x|y=k,\mu,\sigma) = (2\pi\sigma^2)^{-D/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}$

$\because P(x) = \sum_k (2\pi\sigma^2)^{-D/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}\alpha_k$

$= (2\pi\sigma^2)^{-D/2}\sum_k exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}\alpha_k$

$\therefore P(y=k|x,\mu,\sigma) = \frac{(2\pi\sigma^2)^{-D/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}\alpha_k}{(2\pi\sigma^2)^{-D/2}\sum_k exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}\alpha_k}$

$= \frac{exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}\alpha_k}{\sum_k exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}\alpha_k}$

2.)

$\because P(y^1,x^1,y^2,x^2,...,y^N,x^N|\theta) = \Pi_{i=1}^{N}P(x^i,y^i) = \Pi_{i=1}^{N}P(x^i|y^i)P(y^i)$

$\therefore P(y^1,x^1,y^2,x^2,...,y^N,x^N|\theta) = \Pi_{i=1}^{N}(2\pi\sigma^2)^{-D/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2\}\alpha_k$

$\therefore -log(P(y^1,x^1,y^2,x^2,...,y^N,x^N|\theta)) = \sum_N(-\frac{D}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{D}(x_i-\mu_{ki})^2 - log(\alpha_k))$

3.)

$\frac{\partial l(\theta;D)}{\partial \mu_{ki}} = -\frac{1}{\sigma^2}\sum_N\sum_{i=1}^{D}(x_i-\mu_{ki})$

$\frac{\partial l(\theta;D)}{\partial \theta^2} = \frac{-ND}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_N\sum_{i=1}^{D}(x_i-\mu_{ki})^2$

4.)

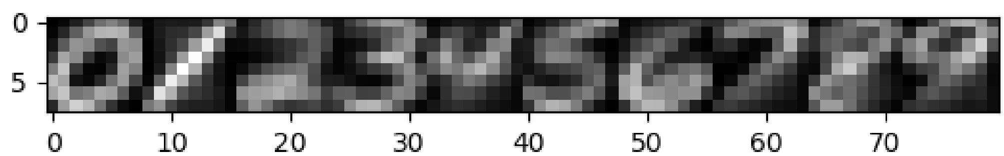
$MLE_{\mu_{ki}} = -\frac{1}{\sigma^2}\sum_N\sum_{i=1}^{D}(x_i-\mu_{ki}) = 0$

$MLE_{\theta} = \frac{-ND}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_N\sum_{i=1}^{D}(x_i-\mu_{ki})^2 = 0$

2 Handwritten Digit Classification:

2.1 $K - NN$ $Classifier$

0.)



1.)

Training classification accuracy for K = 1: 1.0

Test classification accuracy for K = 1: 0.96875

Training classification accuracy for K = 15: 0.9634285714285714

Test classification accuracy for K = 15: 0.961

For detail, please visit q2_1.py

2.)

For K > 1 K-NN, when the algorithm encounter ties, the algorithm will look the closest neighbour whether belongs one of them, if yes, then make the closest neighbour's class as decision. If not, the find the second closest neighbour until we find a neighbour is belonged one of them.

For detail, please visit q2_1.py's function query_knn

3.)

average accuracy across 10 folds	
K	average accuracy
1	0.964428571429
2	0.964428571429
3	0.964428571429
4	**0.965428571429**
5	0.963571428571
6	0.964285714286
7	0.960571428571
8	0.961428571429
9	0.958
10	0.956714285714
11	0.955571428571
12	0.954857142857
13	0.953142857143
14	0.954285714286
15	0.949571428571

value of K: 4

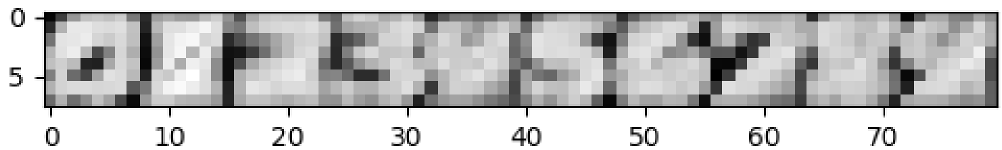
Average accuracy across folds for K = 4 : 0.965428571429

Training classification accuracy for K = 4 : 0.9862857142857143

Test classification accuracy for K = 4 : 0.9725

2.2 *Conditional Gaussian Classifier Training*

1.)



2.)

average conditional log-likelihood for Training data: -0.124624436669

average conditional log-likelihood for Test data: -0.196673203255

For detail, please visit q2_2.py

3.)

accuracy on the train data: 0.9814285714285714

accuracy on the test data: 0.97275

For detail, please visit q2_2.py

2.3 *Naive Bayes Classifier Training*

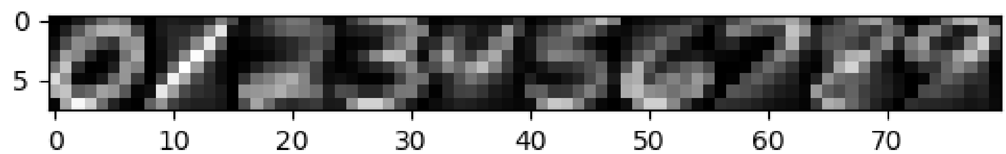
1.)

For detail, please visit q2_3.py

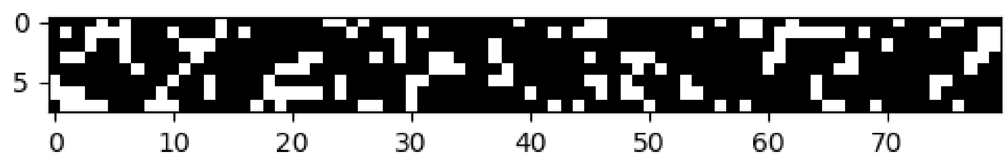
2.)

For detail, please visit q2_3.py

3.)



4.)



5.)

average conditional log-likelihood for Training data: -0.9437538618

average conditional log-likelihood for Test data: -0.987270433725

For detail, please visit q2_3.py

6.)

accuracy on the train data: 0.7741428571428571

accuracy on the test data: 0.76425

For detail, please visit q2_3.py

2.4 *Model Comparison*

Conditional Gaussian Classifier is best among these three. Base on the data, Conditional Gaussian Classifier has the highest accuracy rate on test data. Naive Bayes Classifier is worst base on the accuracy rate. The reason that Naive Bayes Classifier is worst is features on digits are actually not independent, but we have to assume they are independent which makes the prediction inaccuracy. For K-NN, it performs not bad, but we have to run k-fold to find the best k, which takes a long time. Therefore, we consider Conditional Gaussian Classifier has better performance than K-NN.