Mark Gluzman
mg3627
HW 2
Due 2/26
ELEN 4903

Problem 1

For data $x \in \mathbb{R}^d$ and K classes (where class i has regression vector w_i) the class y of x is distributed as

$$P(y|x, w_1, \ldots, w_K) = \prod_{i=1}^{K} \left(\frac{e^{x^T w_i}}{\sum_{j=1}^{K} e^{x^T w_j}} \right)^{I(y=i)}$$

1) Write out the log likelihood $\mathcal{L}$ of data $(x_1, y_1), \ldots, (x_n, y_n)$ using an i.i.d. assumption.

Solution

$$P(y|x, w_1, \ldots, w_K) = \prod_{i=1}^{K} \left(\frac{e^{x^T w_i}}{\sum e^{x^T w_j}} \right)^{I(y=i)}$$ or

$$P(y=i|x, w_1, \ldots w_K) = \frac{e^{x^T w_i}}{\sum_j^K e^{x^T w_j}}$$

Let $W = [w_1 \; w_2 \; \ldots \; w_K]$, where K is a number of classes ~~of samples~~.

Using i.i.d assumption the likelihood function is given by $p(T | w_1, w_2, \ldots, w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(y=k|x_n)^{t_{nk}}$,

where N is a number of samples.

$T = \{t_{nk}\}$, $t_{nk} = \begin{cases} 0 & \text{if } y \neq k \text{ in the n-th sample} \\ 1 & \text{if } y = k \text{ in the n-th sample} \end{cases}$

Or $t_{nk} = I(y_n = k)$

Then taking the log we will have

$$\mathcal{L}(\Theta) = \ln p(T,\Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} I(y_n = k) \times$$

$$\times \log \frac{e^{x_n^T w_k}}{\sum_j^K e^{x_n^T w_j}} = \sum_{n=1}^{N}\sum_{k=1}^{K} I(y_n = k)\left(x_n^T w_k - \log \sum_j^K e^{x_n^T w_j}\right)$$

2) Calculate $\nabla_{w_i}\mathcal{L}$ and $\nabla^2_{w_i}\mathcal{L}$

Solution

$$\nabla_{w_i}\mathcal{L}(\Theta) = \sum_{n=1}^{N} I(y_n = i)x_n - \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{1}{\sum_{j=1}^{K} e^{x_n^T w_j}} \cdot x_i \cdot e^{x_n^T w_k} \times$$

$$\times I(y_n = k) =$$

$$= \sum_{n=1}^{N} I(y_n = i)x_n - \sum_{n=1}^{N} x_i \frac{e^{x_n^T w_i}}{\sum_{j=1}^{K} e^{x_n^T w_j}} =$$

$$= \sum_{n=1}^{N} x_n\left(I(y_n = i) - \frac{e^{x_n^T w_i}}{\sum_{j=1}^{K} e^{x_n^T w_j}}\right)$$

$$\nabla^2_{w_i}\mathcal{L}(\Theta) = -x_n \sum_{n=1}^{N}\frac{x_n e^{x_n^T w_i}\sum_{j=1}^{K} e^{x_n^T w_j} - x_n e^{2x_n^T w_i}}{\left(\sum_{j=1}^{K} e^{x_n^T w_j}\right)^T\left(\sum_{j=1}^{K} e^{x_n^T w_j}\right)}$$

$$= -\sum_{n=1}^{N}\frac{x_n x_n^T\left(e^{2x_n^T w_i} - e^{2x_n^T w_i} + \sum_{j=1, j\neq i}^{K} e^{x_n^T(w_j + w_i)}\right)}{\left(\sum_{j=1}^{K} e^{x_n^T w_j}\right)^T\left(\sum_{j=1}^{K} e^{x_n^T w_j}\right)} =$$

$$= -\sum_{n=1}^{N} x_n x_n^T\left(+\underbrace{\frac{e^{x_n^T w_i}}{\sum e^{x_n^T w_j}}}_{p(y=i|x_n)} - \left(\frac{e^{x_n^T w_i}}{\sum e^{x_n^T w_j}}\right)^2\right) =$$

$$= -\sum_{n=1}^{N} x_n x_n^T\left(p(y=i|x_n)\left(1 - p(y=i|x_n)\right)\right)$$

Problem 2

In the integral case,

$$k(u,v)=\int_{\mathbb{R}^d} \phi_t(u)\phi_t(v)\,dt,$$

Show that the mapping $\phi_t(u)=\frac{1}{(2\pi\nu)^{d/2}}e^{-\frac{\|u-t\|^2}{2\nu}}$

reproduces the Gaussian kernel $k(u,v)=$

$=\alpha\exp\left(-\frac{\|u-v\|^2}{\beta}\right)$ for an appropriate setting

of α and β.

Solution

First of all, ~~note~~

$$\|u-t\|^2+\|v-t\|^2=u^Tu-2u^Tt+t^Tt+$$

$$+v^Tv-2v^Tt+t^Tt=\frac{1}{2}u^Tu+\frac{1}{2}v^Tv+u^Tv$$

$$+\frac{1}{2}u^Tu+\frac{1}{2}v^Tv-u^Tv+2t^Tt-2(u^T+v^T)t=$$

$$=\frac{1}{2}\|u-v\|^2+2t^Tt-2(u^T+v^T)t$$

$$+\frac{1}{2}(u+v)^T(u+v)=\frac{1}{2}\|u-v\|^2+2\left\|t-\frac{u+v}{2}\right\|^2$$

So, $\|u-t\|^2+\|v-t\|^2=\frac{1}{2}\|u-v\|^2+2\left\|t-\frac{u+v}{2}\right\|^2$

And so

$$\int_{\mathbb{R}^d} \phi_t(u)\,\phi_t(v)\,dt = \frac{1}{(2\pi\nu)^d}\int_{\mathbb{R}^d} e^{-\frac{1}{2\nu}\left(\|u-t\|^2+\|v-t\|^2\right)}\,dt$$

$$= \frac{1}{(2\pi\nu)^d}\int_{\mathbb{R}^d} e^{-\frac{1}{2\nu}\left(\frac{1}{2}\|u-v\|^2+2\left\|t-\frac{u+v}{2}\right\|^2\right)}\,dt =$$

$$= \frac{1}{(2\pi\nu)^d}\, e^{-\frac{1}{2\nu}\left(\frac{1}{2}\|u-v\|^2\right)}\int_{\mathbb{R}^d} e^{-\frac{1}{\nu}\left\|t-\frac{u+v}{2}\right\|^2}\,dt =$$

From the theory of Gaussian distribution the Gaussian density has the property

$$\int_{\mathbb{R}^d} e^{-\frac{\|x-\mu\|^2}{\nu}}\,dx = (\pi\nu)^{d/2}, \quad \text{for } \forall \mu > 0,\ x\in\mathbb{R}^d$$

Thus

$$\int_{\mathbb{R}^d} e^{-\frac{1}{\nu}\left\|t-\frac{u+v}{2}\right\|^2}\,dt =$$

$$= (\pi\nu)^{d/2}$$

Hence

$$\int_{\mathbb{R}^d} \phi_t(u)\,\phi_t(v)\,dt = \frac{1}{(2\pi\nu)^d}\, e^{-\frac{1}{2\nu}\left(\frac{1}{2}\|u-v\|^2\right)} \times (\pi\nu)^{d/2} = \frac{1}{2^d(\pi\nu)^{d/2}}\, e^{-\frac{1}{4\nu}\left(\|u-v\|^2\right)}$$

So, $\int_{\mathbb{R}^d} \phi_t(u)\,\phi_t(v)\,dt = k(u,v) =$

$$= \alpha\, e^{\frac{-\|u-v\|^2}{\beta}}, \text{ where } \alpha = \frac{1}{2^d(\pi\gamma)^{d/2}}$$

and $\beta = 4\gamma$

Problem 3a

- For each k calculate the confusion matrix and show the trace of this matrix divided by 500. This is the prediction accuracy. You don't need to show the confusion matrix.

Solution

K=1

The prediction accuracy: 0.948

The confusion matrix:

48	0	0	1	0	1	0	0	0	0
0	50	0	0	0	0	0	0	0	0
0	0	49	0	0	0	1	0	0	0
0	1	0	44	0	2	0	1	2	0
0	0	0	0	49	0	0	0	0	1
0	0	0	0	2	46	1	0	0	1
0	0	1	0	1	0	48	0	0	0
0	1	1	0	0	0	0	47	0	1
0	0	1	0	0	0	1	0	47	1
0	1	0	0	2	0	1	0	0	46

K=2

The prediction accuracy: 0.93

The confusion matrix:

48	0	0	1	0	1	0	0	0	0
0	50	0	0	0	0	0	0	0	0
0	0	49	0	0	0	1	0	0	0
0	1	0	47	0	0	0	1	1	0
0	0	0	0	49	0	0	0	0	1
0	0	0	2	2	44	1	0	0	1
0	0	1	0	2	0	47	0	0	0
0	1	1	0	1	0	0	46	0	1

0	0	2	3	0	1	1	0	42	1
0	1	0	1	3	0	2	0	0	43

K=3

The prediction accuracy: 0.938

The confusion matrix:

48	0	0	1	0	1	0	0	0	0
0	50	0	0	0	0	0	0	0	0
0	0	46	1	0	0	1	0	2	0
0	1	0	44	0	0	0	1	4	0
0	0	0	0	48	0	0	1	0	1
0	0	0	1	1	46	1	0	0	1
0	0	0	0	2	0	48	0	0	0
0	1	1	0	1	0	0	46	0	1
0	0	1	1	0	1	0	0	46	1
0	1	0	0	1	0	1	0	0	47

K=4

The prediction accuracy: 0.946

The confusion matrix:

48	0	0	1	0	1	0	0	0	0
0	50	0	0	0	0	0	0	0	0
0	0	47	0	0	0	1	0	2	0
0	1	0	46	0	0	0	1	2	0
0	0	0	0	48	0	0	1	0	1
0	0	0	0	1	47	1	0	0	1
0	0	0	0	2	0	48	0	0	0
0	1	1	0	1	0	0	46	0	1
0	0	1	0	0	2	0	0	46	1
0	1	0	0	1	0	1	0	0	47

K=5

The prediction accuracy: 0.946

The confusion matrix:

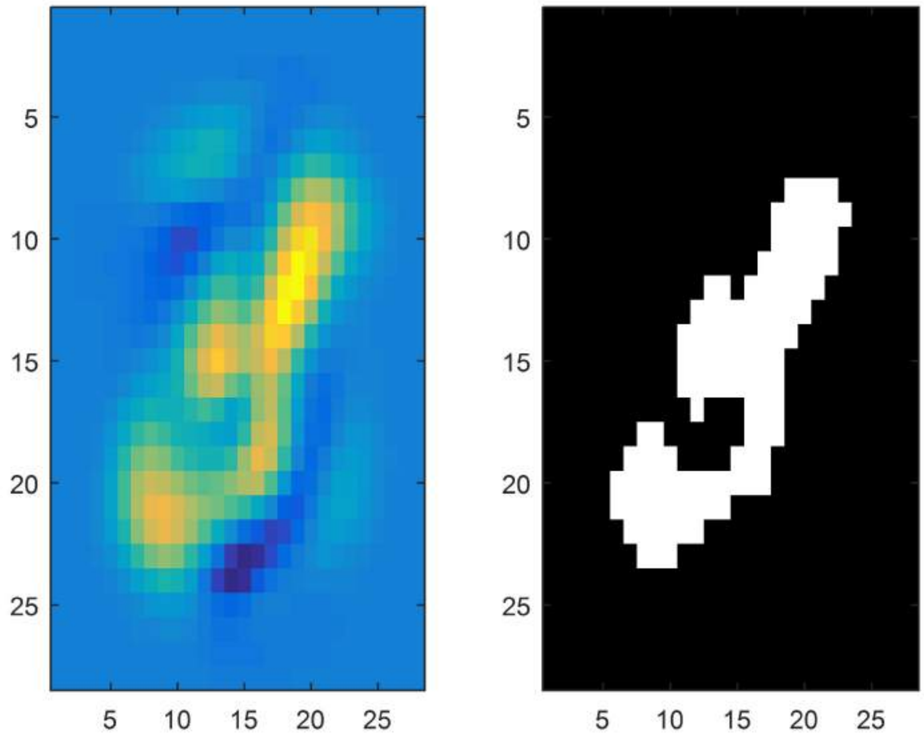
48	0	0	1	0	1	0	0	0	0
0	50	0	0	0	0	0	0	0	0
0	0	47	0	0	0	1	0	2	0
0	1	0	46	0	0	0	1	2	0
0	0	0	0	47	0	0	1	0	2
0	0	0	0	1	47	1	0	0	1
0	0	0	0	2	0	48	0	0	0
0	1	1	0	1	0	0	46	0	1
0	0	1	0	0	2	0	0	46	1
0	1	0	0	0	0	1	0	0	48

- For $k = 1, 3, 5$, show three misclassified examples as images and indicate the true class and the predicted class for each one.

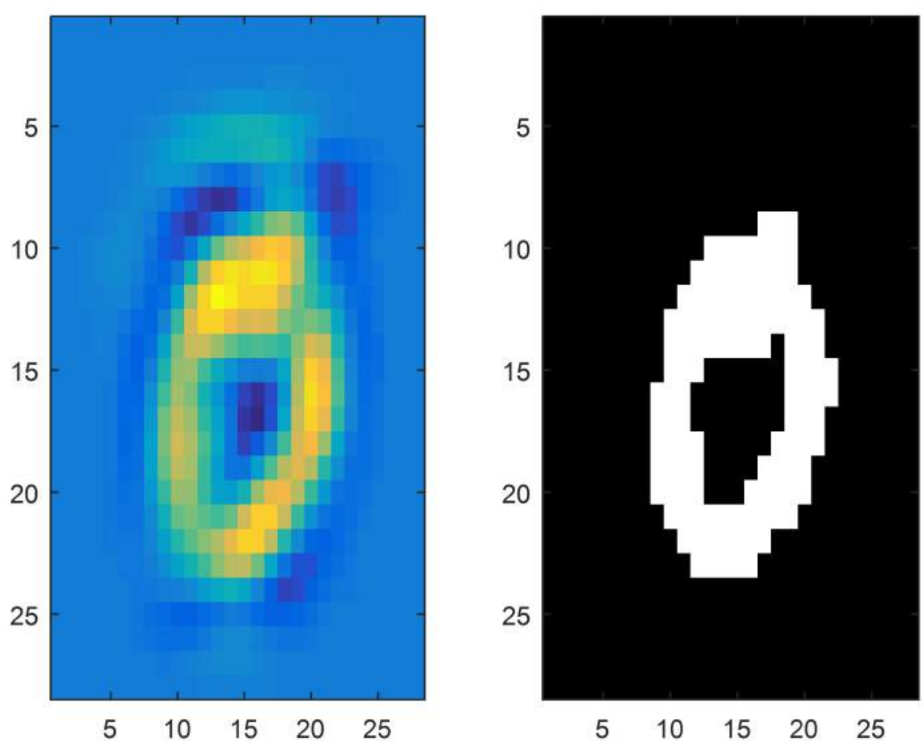
Solution

K=1

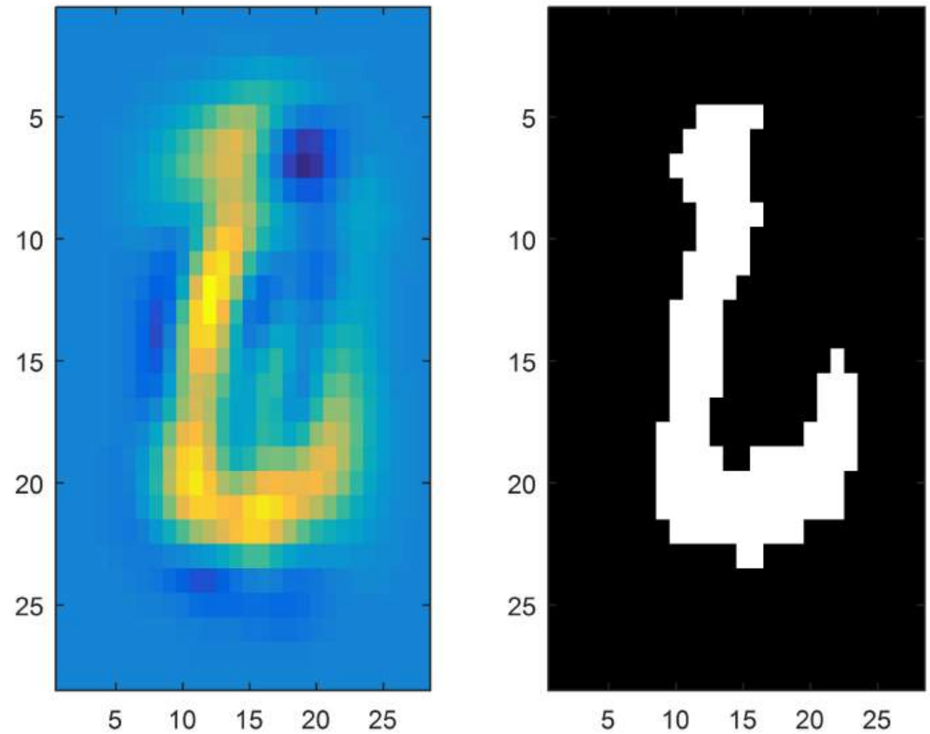
N = 11 ; Actual digit = 0 ; Predicted digit = 5



N = 20 ; Actual digit = 0 ; Predicted digit = 3

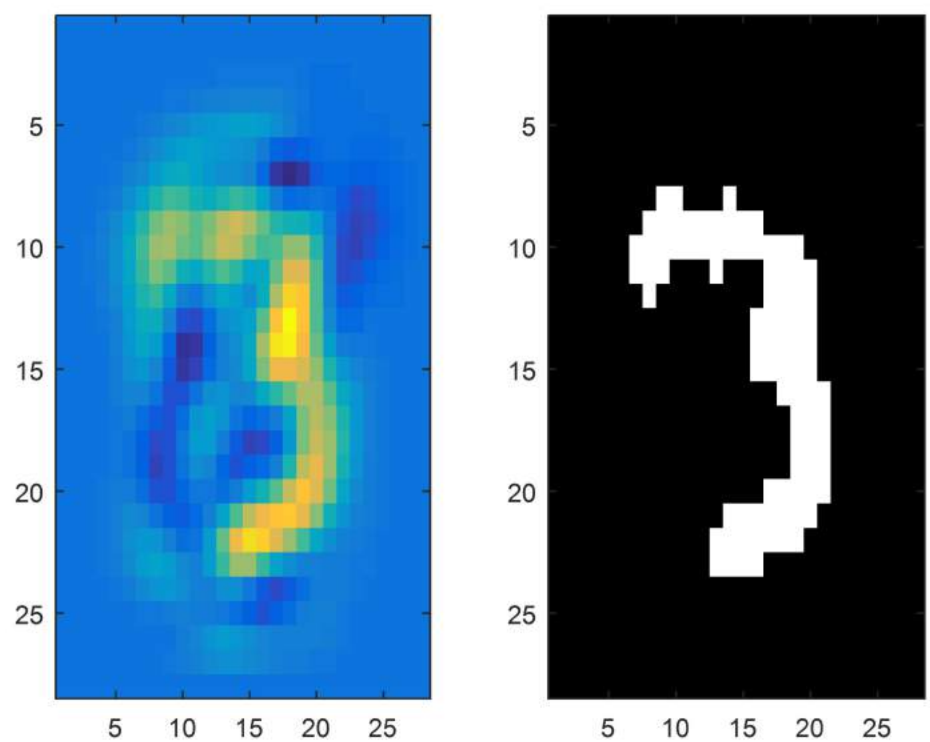


N = 141 ; Actual digit = 2 ; Predicted digit = 6

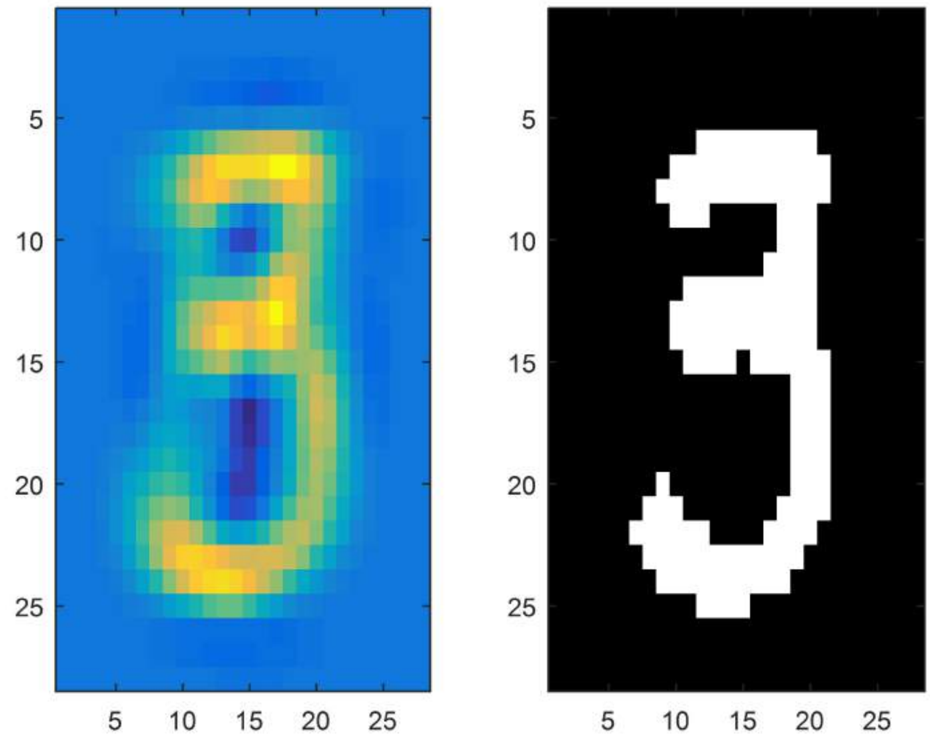


K=3

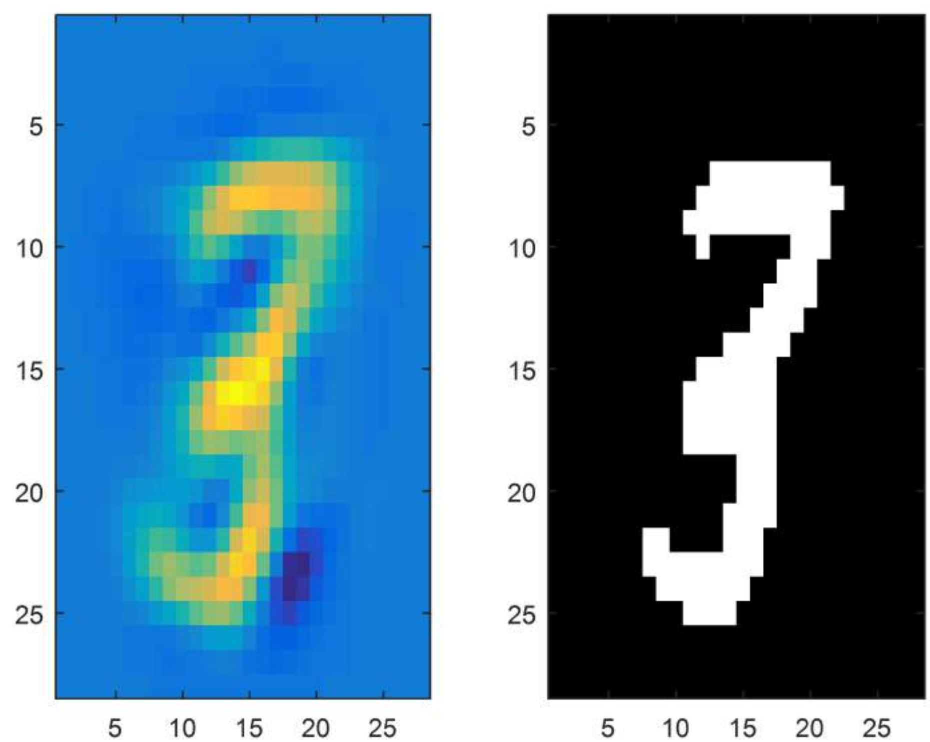
N = 155 ; Actual digit = 3   ; Predicted digit = 7



N =165 ; Actual digit = 3 ; Predicted digit = 8

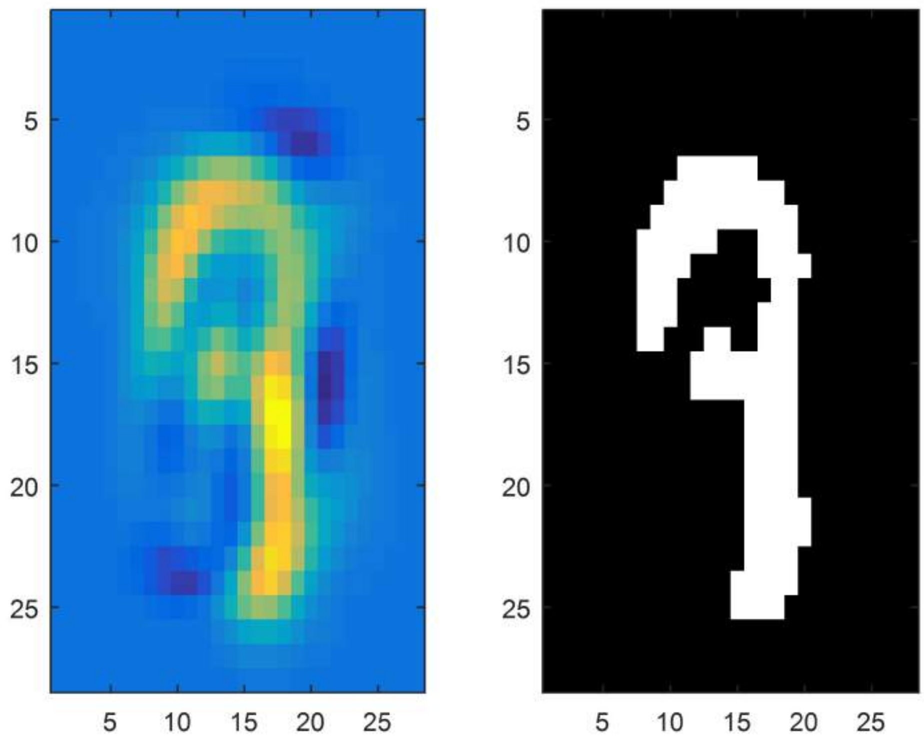


N = 166 ; Actual digit = 3 ; Predicted digit = 8

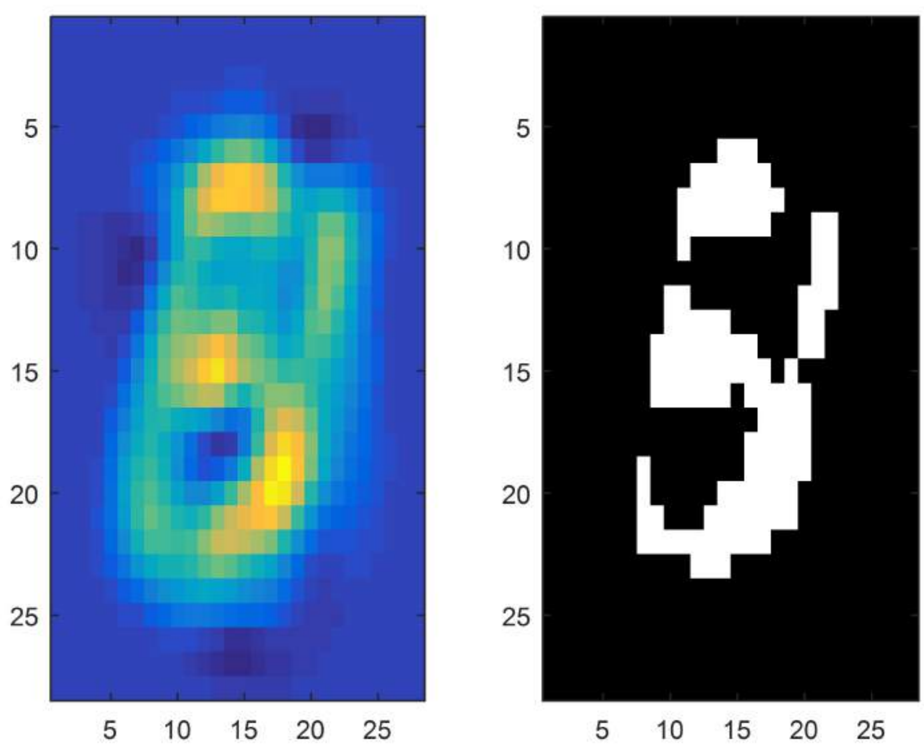


K=5

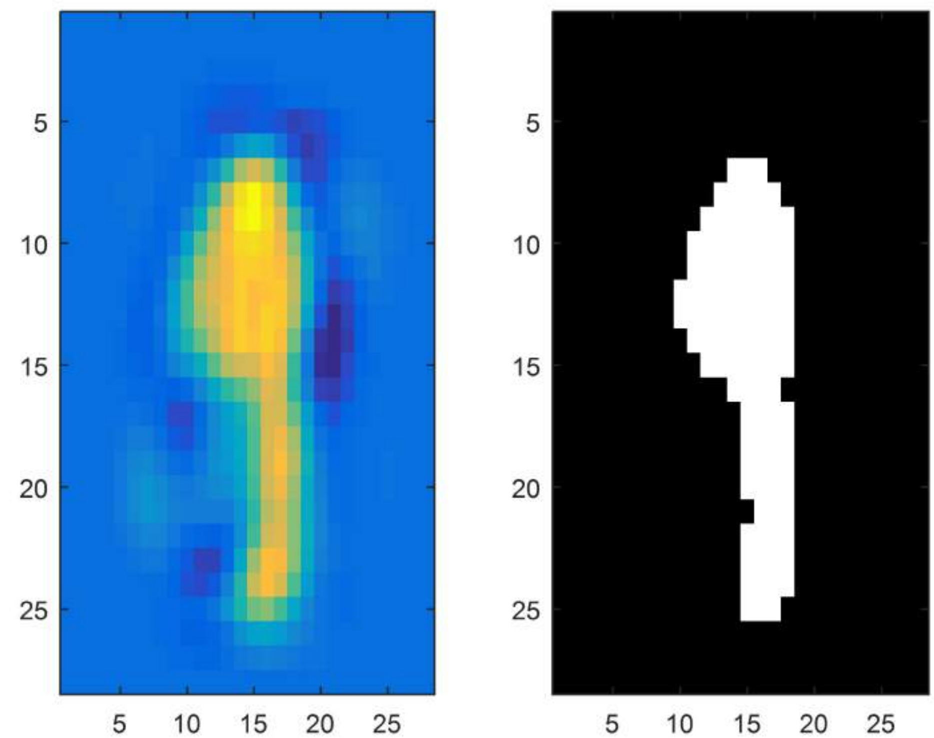
N = 436 ; Actual digit = 8 ; Predicted digit = 9



N = 448 ; Actual digit =8 ; Predicted digit = 2



N = 456 ; Actual digit =9 ; Predicted digit = 1



5
10
15
20
25
5
10
15
20
25
5
10
15
20
25
5
10
15
20
25

Problem 3 b 1)

~~Implement the Bayes classifier using a~~

Derive the maximum likelihood estimate for the mean and covariance for a particular class j. Show the answer you obtain for the mean and covariance, as well as the estimate for the class prior.

Solution

We consider a particular class j.

$$p(D_j \mid \mu^*, \Sigma^*) = \prod_{i=1}^{n} p(x_i \mid \mu, \Sigma),$$

where D_j - a part of training data, such that the target class $y=j$. n is total amount of samples such that $y=j$. We also used i.i.d assumption.

Since we have multivariate Gaussian case the log-likelihood function is

$$\mathcal{L}(\mu, \Sigma) = \sum_i \ln p(x_i \mid \mu, \Sigma) =$$

$$= \sum_{i=1}^{n} -\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) - \frac{1}{2}\ln(2\pi)^d|\Sigma|$$

$$\nabla_\mu \mathcal{L}(\mu, \Sigma) = \sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu) = 0$$

So, $\mu^* = \frac{1}{n}\sum_{j=1}^{n} x_j$.

$$\nabla_\Sigma \mathcal{L}(\mu, \Sigma) = \sum_{i=1}^{n} \left(+\frac{1}{2}(x_i - \mu^*)^T \Sigma^{-2} (x_i - \mu^*)\right)$$

$$= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-2}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

And so $\Sigma^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^*)(x_i - \mu^*)^T$

As a result gen. class j we have

$$\mu_j^* = \frac{1}{n}\sum_{i=1}^{n} x_i^j \; ; \; \Sigma_j^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_j^*)(x_i - \mu_j^*)^T$$

where x_i^j - samples with outputs $y = j$ (or with targets in class j.)

Next, denote the prior probability of class j by π_j. Then we will assume that

$\pi_j = \frac{N_j}{N}$, where N is total number of samples in the treing data, N_j is number of samples in class j.

Problem 3b

- Show the confusion matrix in a table. As in Problem 3a, indicate the prediction accuracy by summing along the diagonal and dividing by 500.

Solution

The confusion matrix is

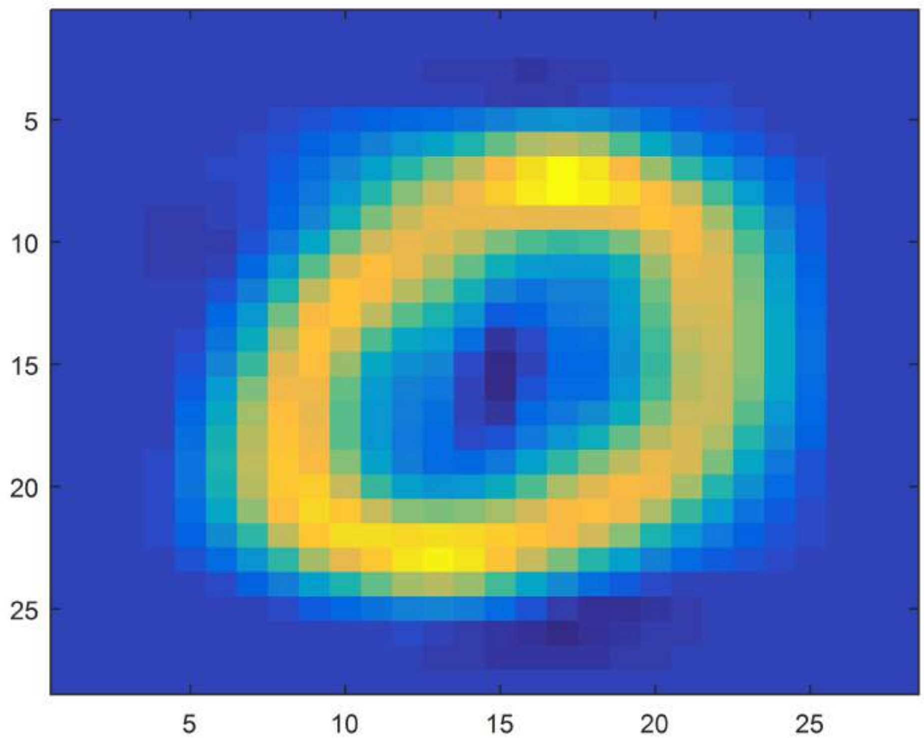
48	0	0	1	0	1	0	0	0	0
0	49	0	0	0	0	0	0	1	0
0	0	48	0	1	0	1	0	0	0
0	0	1	47	0	0	0	0	2	0
0	0	0	0	48	0	0	0	1	1
0	0	0	1	0	45	2	0	1	1
0	0	0	0	1	5	43	0	0	1
0	0	2	0	2	0	0	46	0	0
0	0	1	0	0	1	0	0	47	1
1	0	0	0	2	0	0	0	0	47

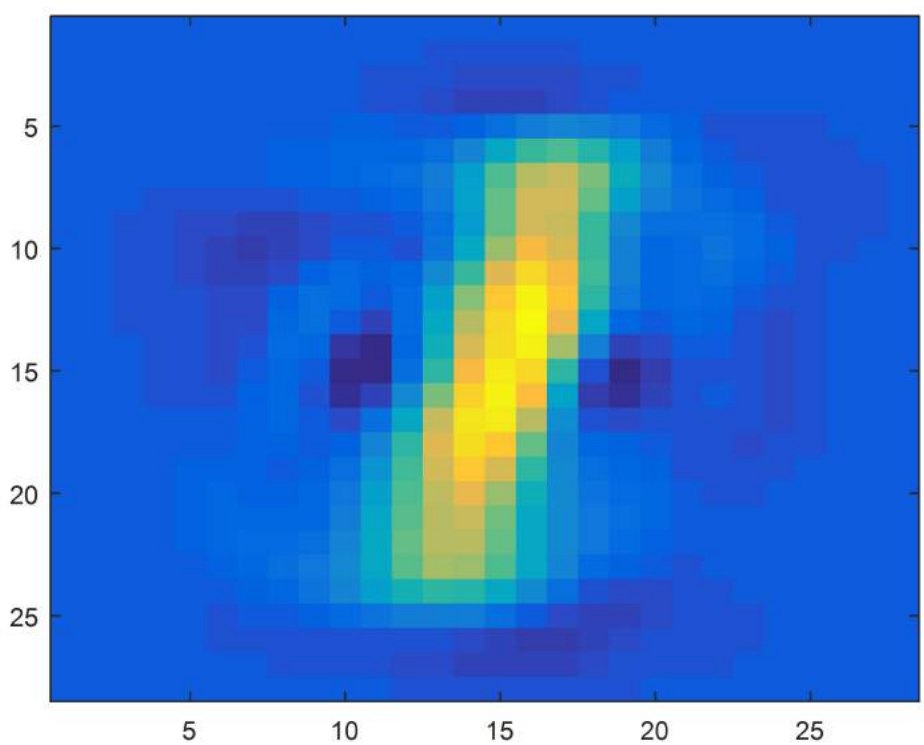
The prediction accuracy is

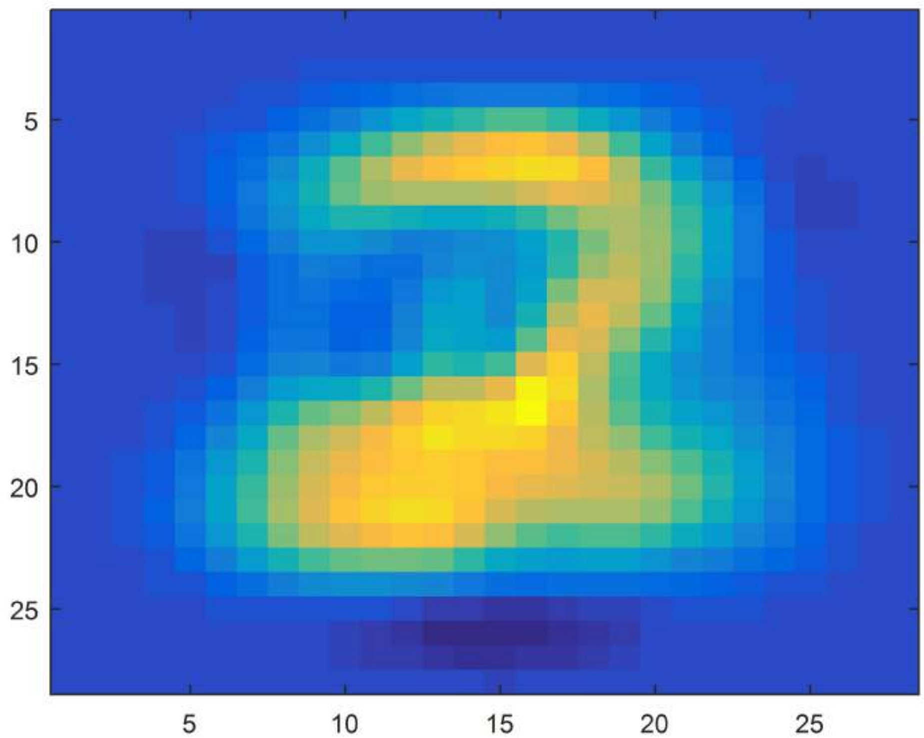
0.9360

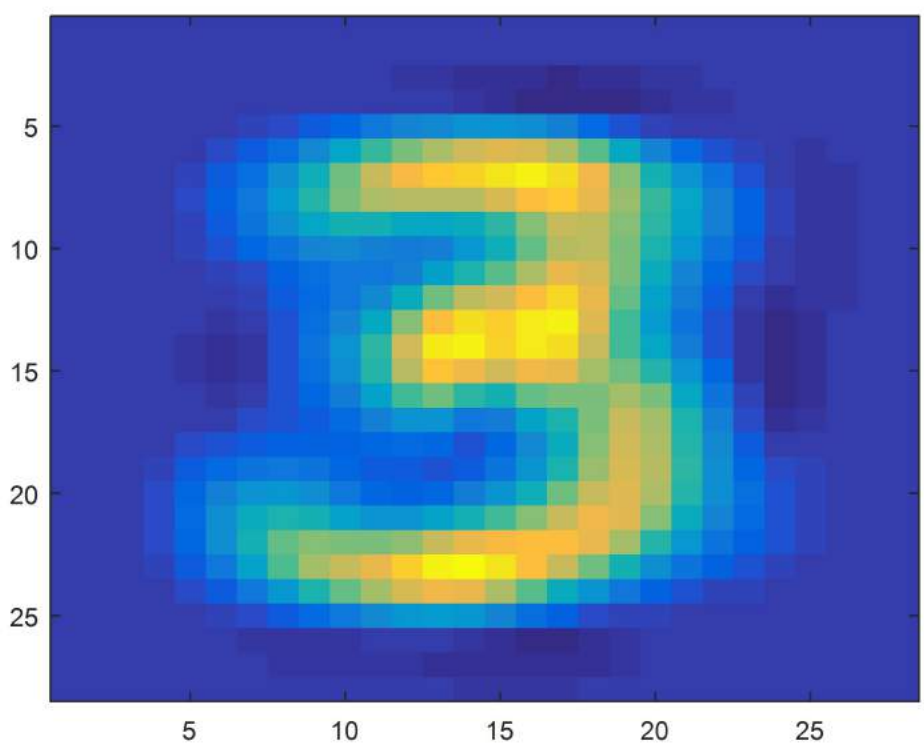
- Show the mean of each Gaussian as an image using the provided Q matrix.

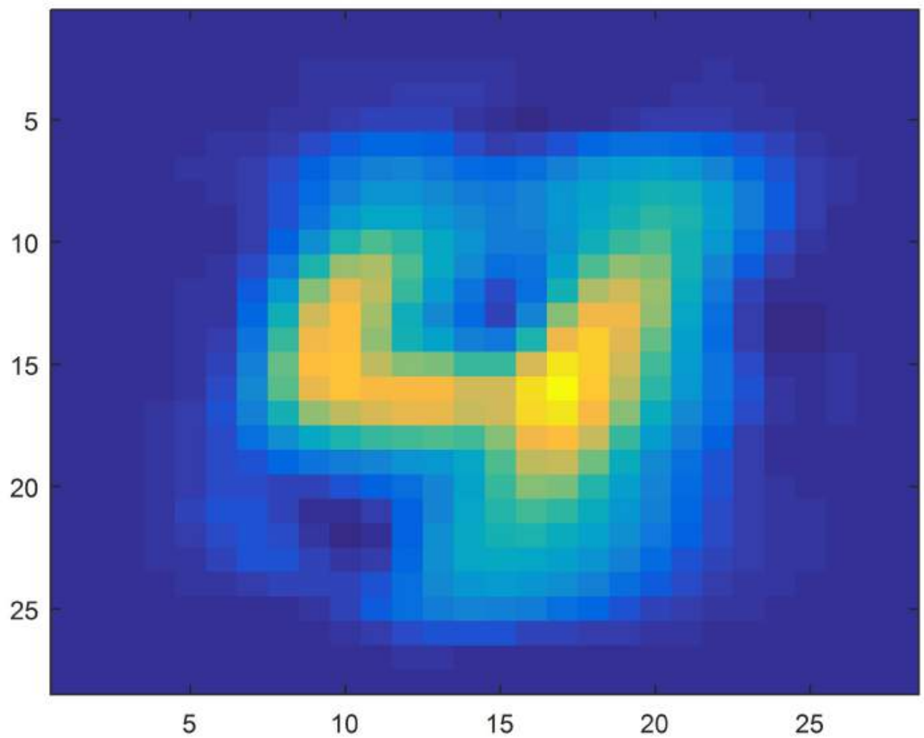
Solution

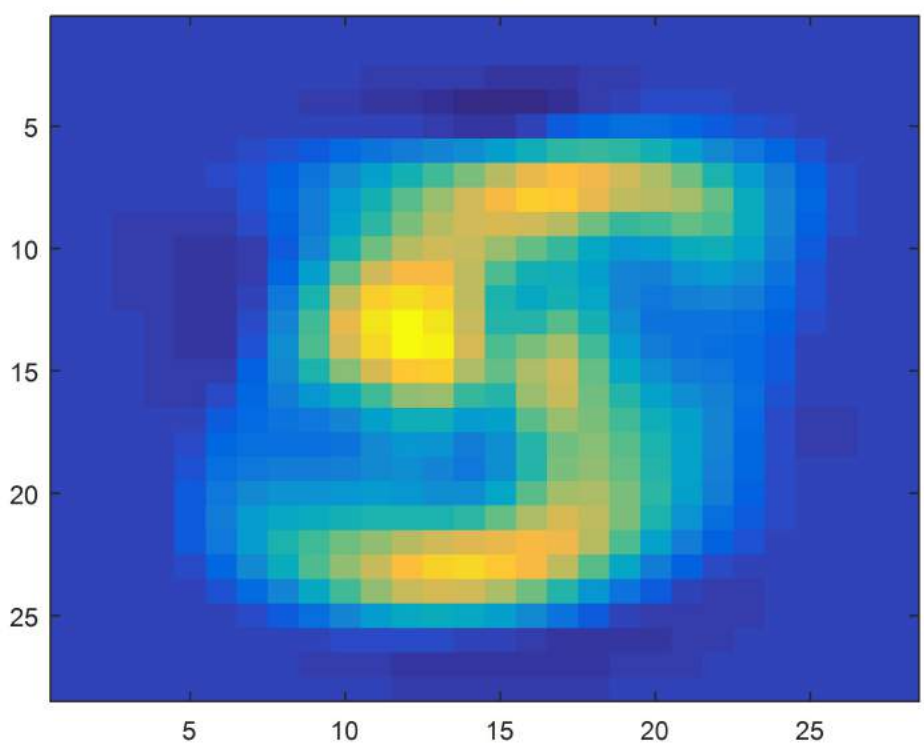

5
10
15
20
25
5
10
15
20
25

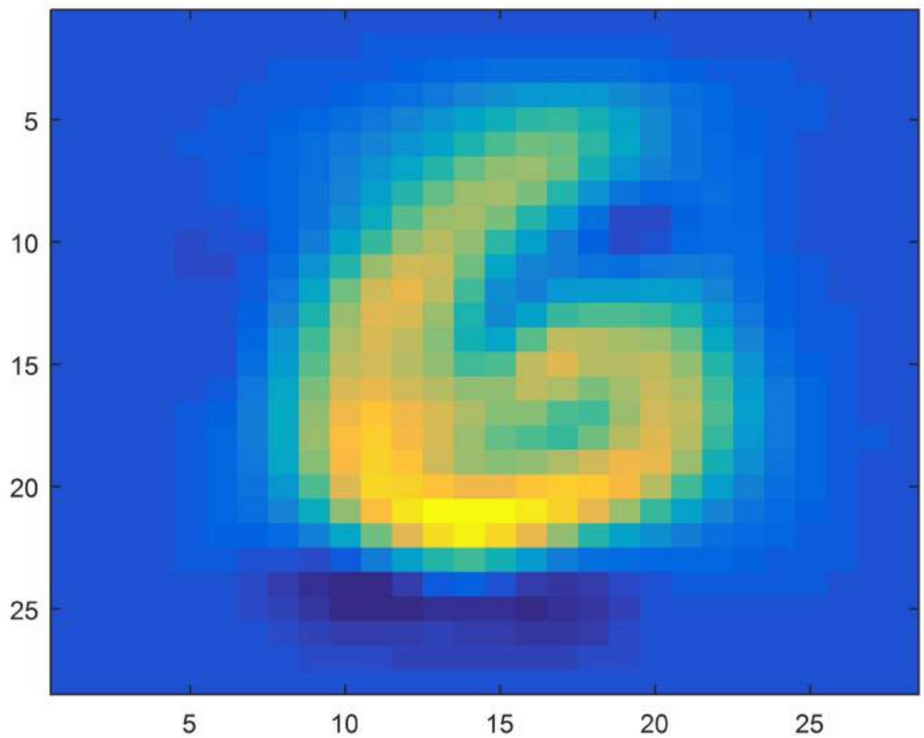

5
10
15
20
25
5
10
15
20
25

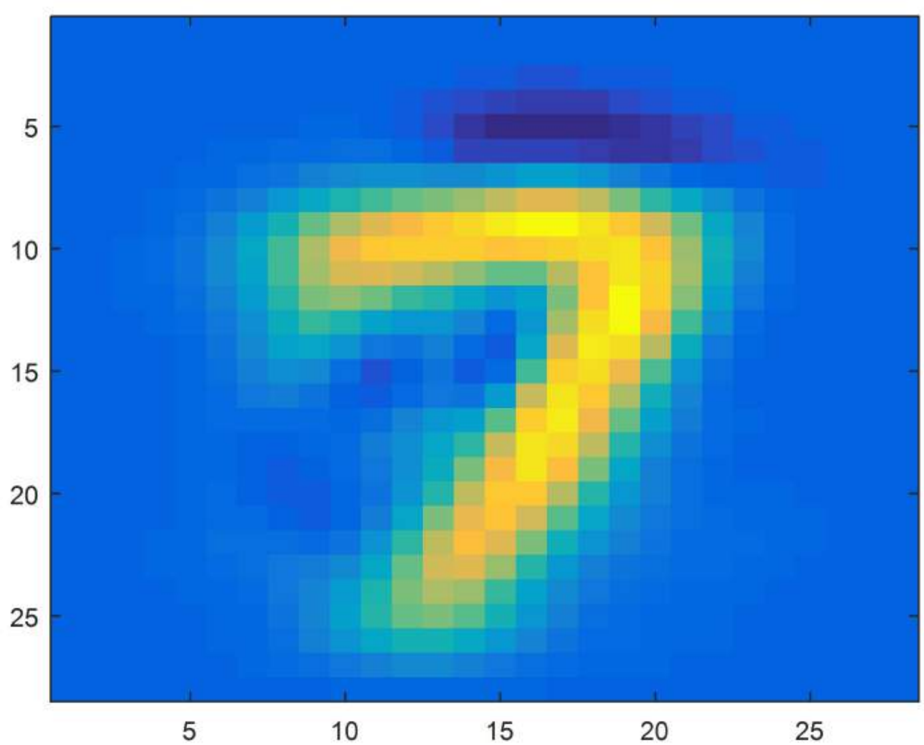

5
10
15
20
25
5 10 15 20 25

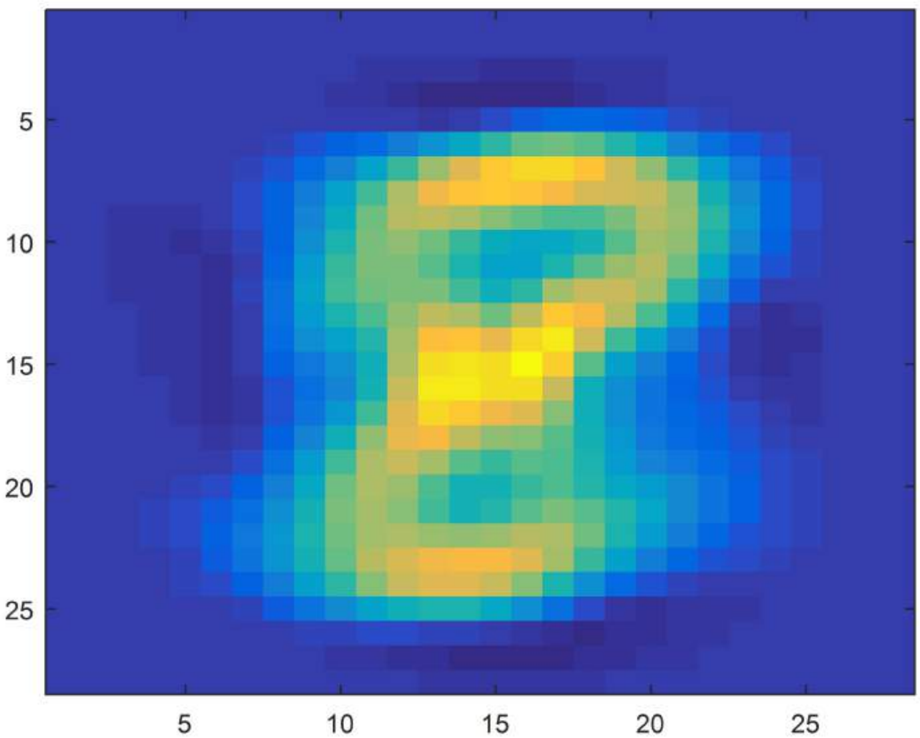

5
10
15
20
25
5 10 15 20 25

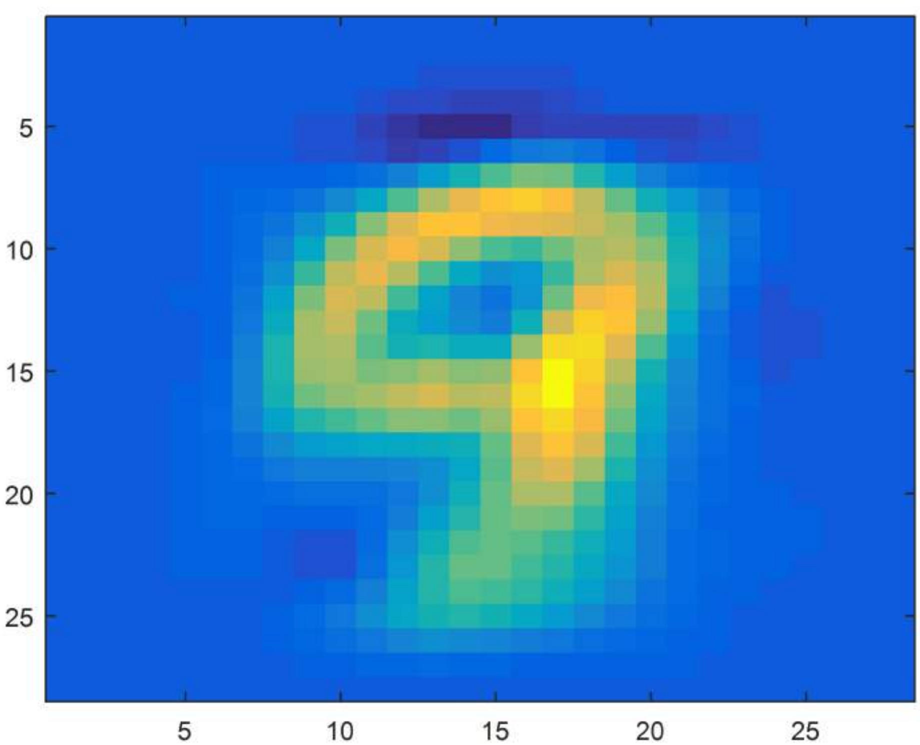

5
10
15
20
25
5
10
15
20
25


5
10
15
20
25
5
10
15
20
25


5
10
15
20
25
5
10
15
20
25


5
10
15
20
25
5
10
15
20
25

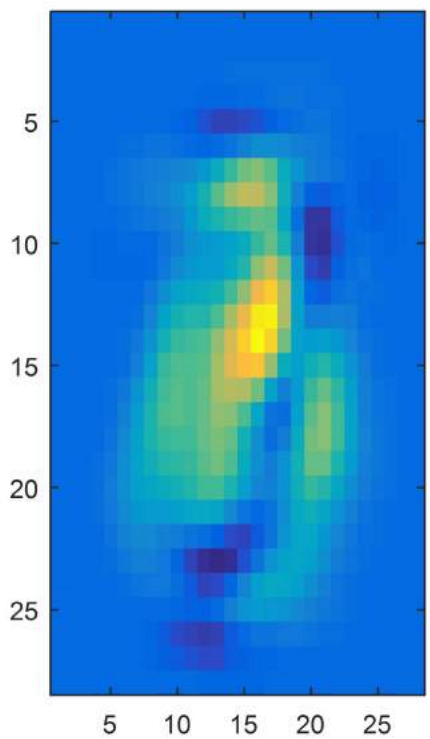

5
10
15
20
25
5
10
15
20
25

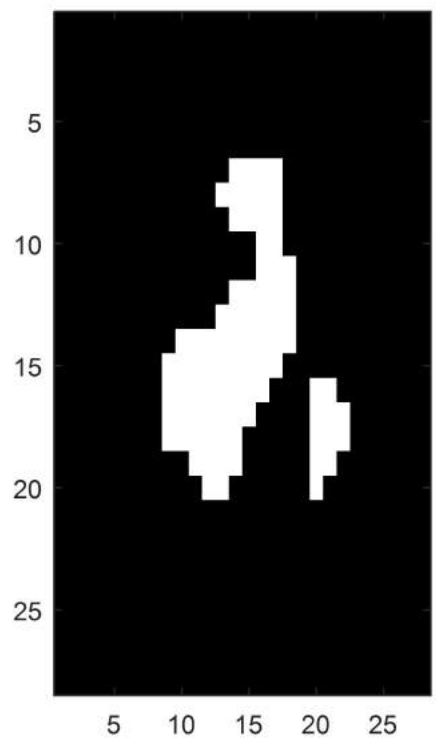

5
10
15
20
25
5
10
15
20
25

- Show three misclassified examples as images and show the probability distribution on the 10 digits learned by the Bayes classifier for each one.

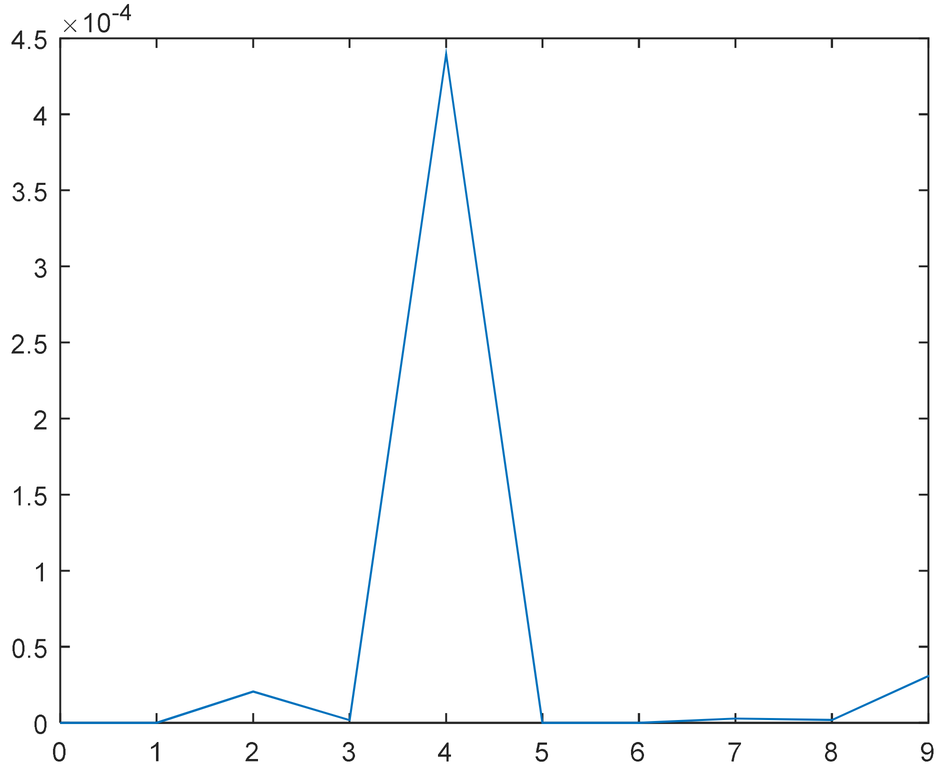
Solution

n= 130
Actual digit = 2 ; Predicted digit =4



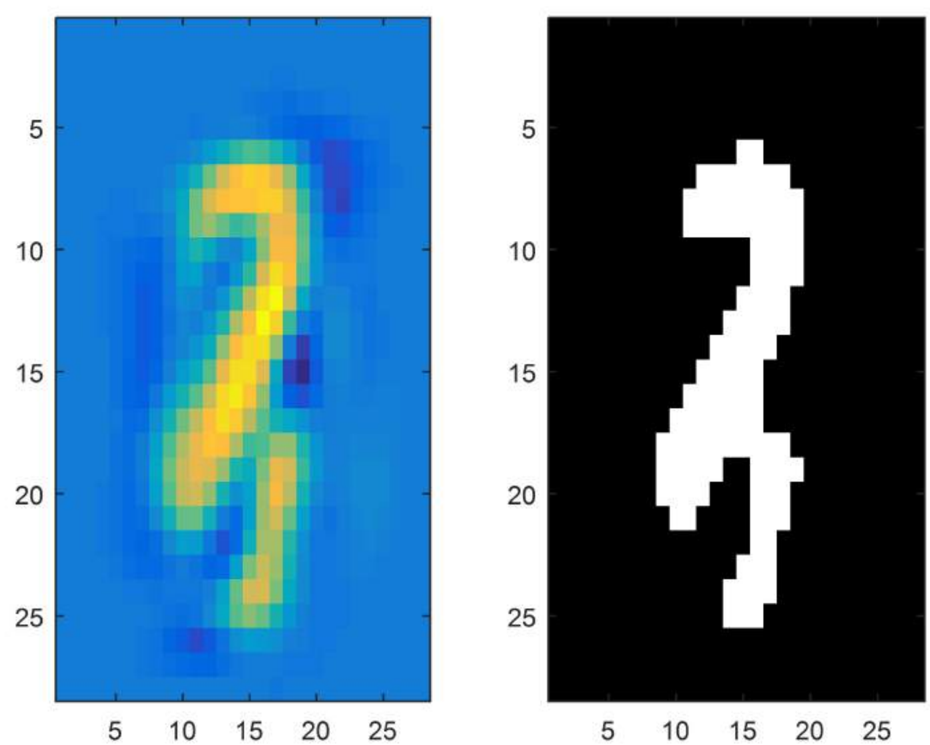


Plot of the probability distribution:

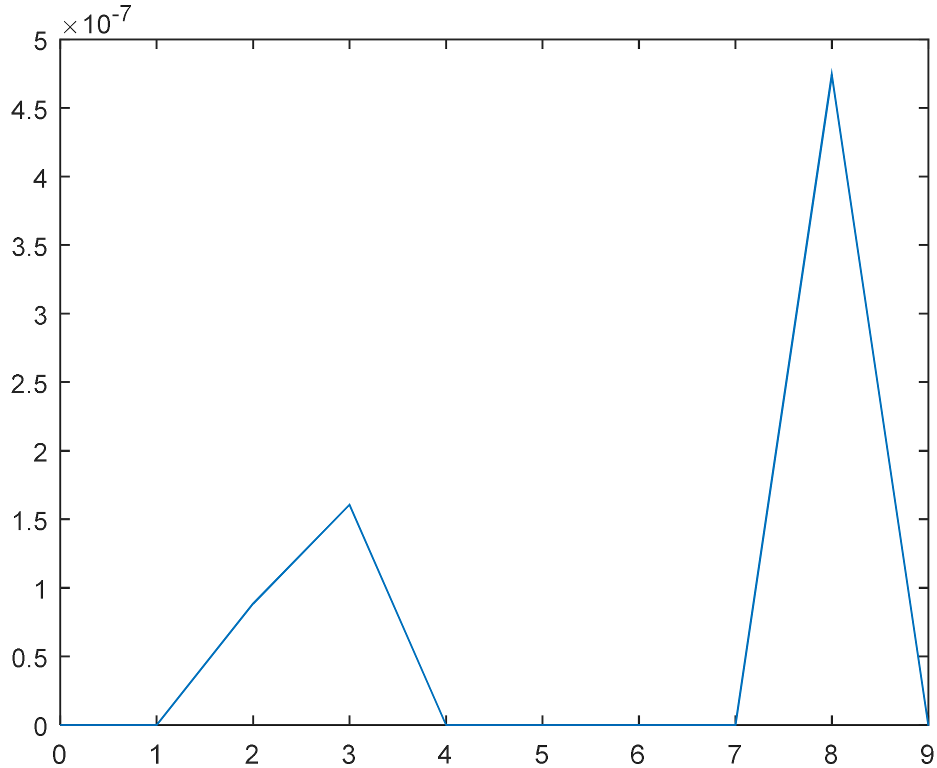


n=180

Actual digit = 3 ; Predicted digit =8

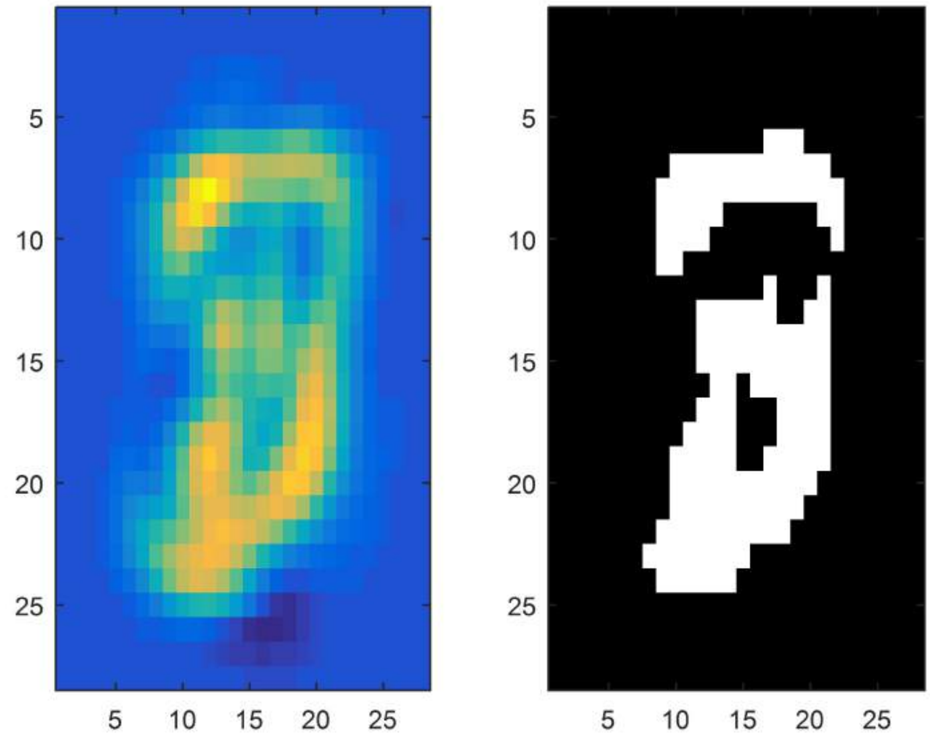


Plot of the probability distribution:

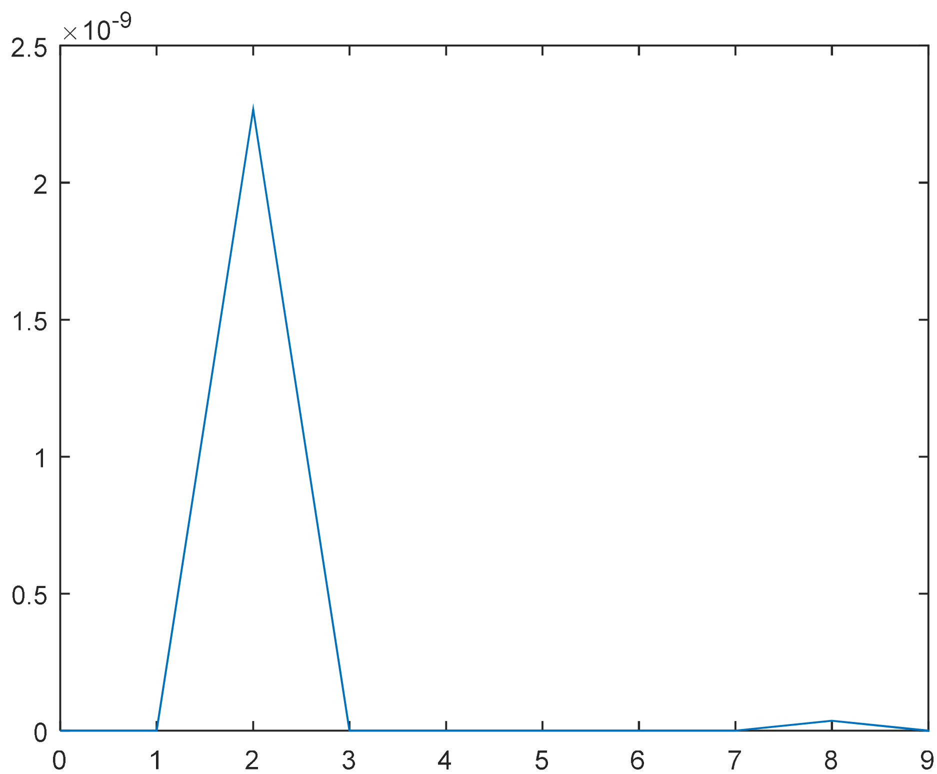


n=422

Actual digit = 8 ; Predicted digit =2



Plot of the probability distribution:

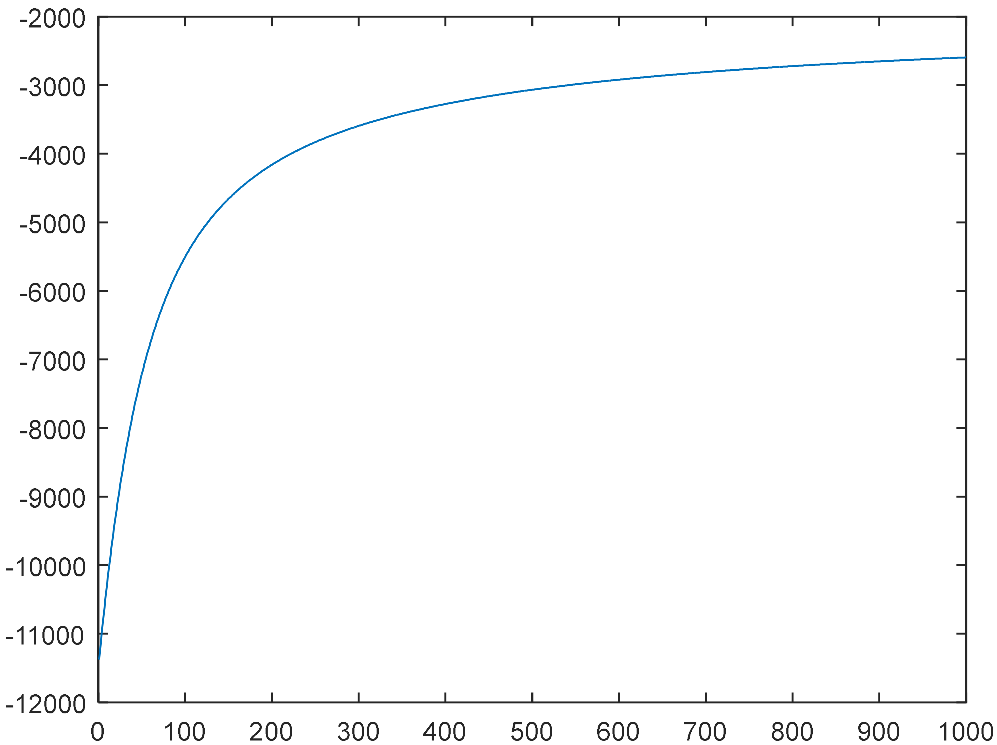


Problem 3c

- After making an update of each $w_0, \ldots, w_9$, calculate $\mathcal{L}$ (see Problem 1) and plot as a function of iteration. Run your algorithm for 1000 iterations.

Solution

Plot of the log likelihood:



- Show the confusion matrix in a table. Indicate the prediction accuracy by summing along the diagonal and dividing by 500.

Solution

The confusion matrix is

46	0	1	1	0	0	2	0	0	0
0	49	0	0	0	0	0	0	1	0
0	0	38	2	1	0	4	0	5	0
1	0	2	39	0	2	0	1	5	0
0	0	1	0	42	1	0	0	1	5
1	1	0	4	2	39	1	0	0	2
0	0	1	0	4	3	42	0	0	0
0	0	3	0	1	0	0	44	1	1

0	0	0	0	0	2	1	0	46	1
0	1	1	0	3	0	0	1	0	44

The prediction accuracy is

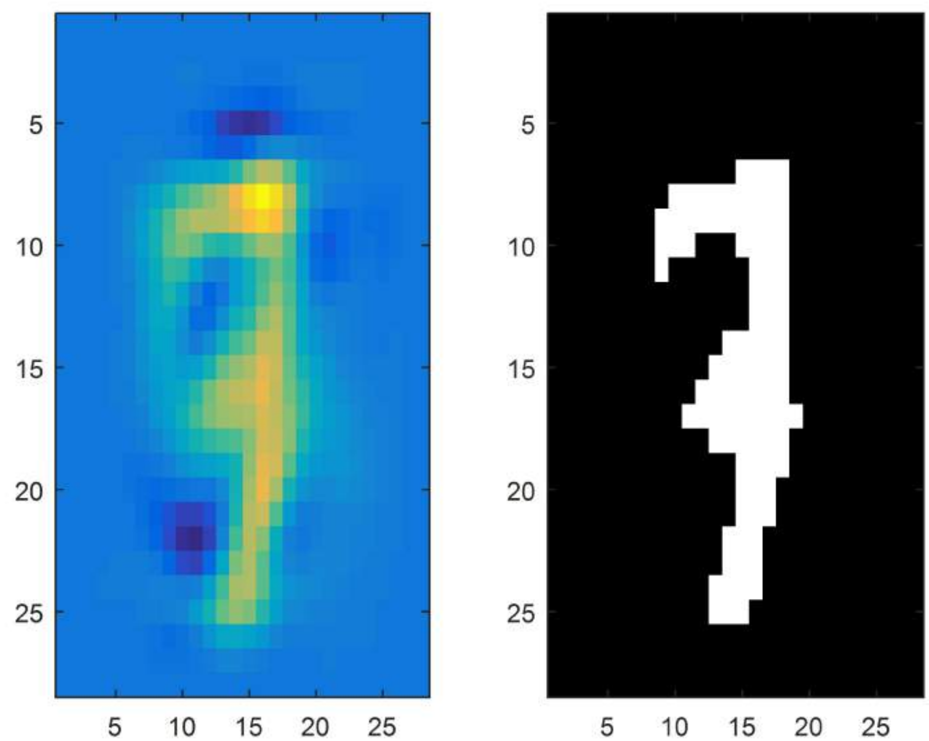
0.8580

- Show three misclassified examples as an image and show the probability distribution on the 10 digits learned by the softmax function for each one.

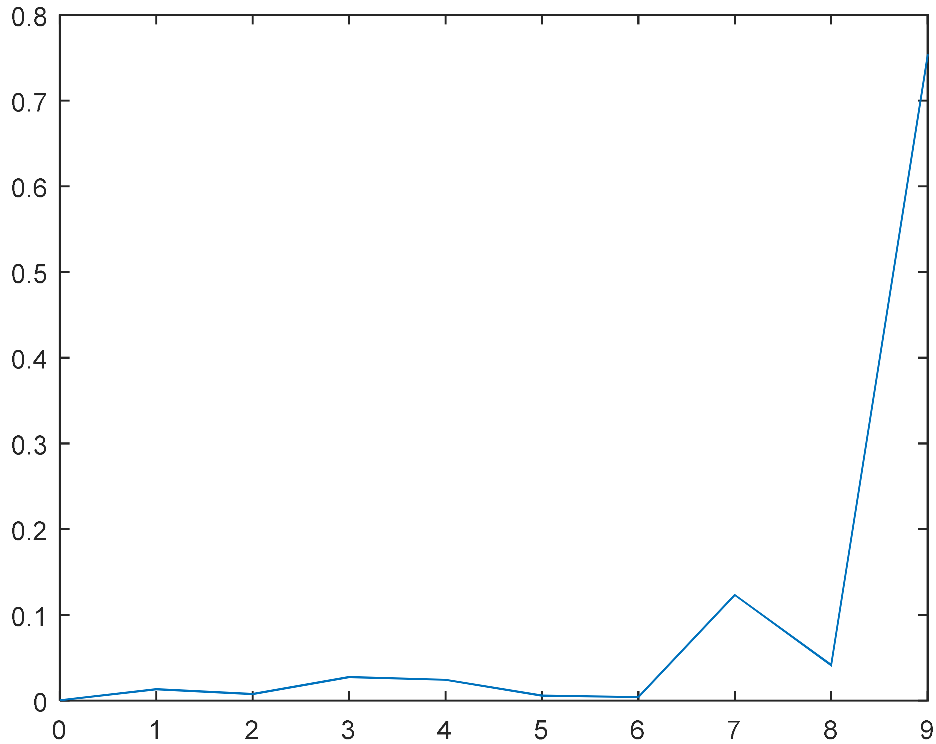
Solution

n=384

Actual digit = 7 ; Predicted digit =9

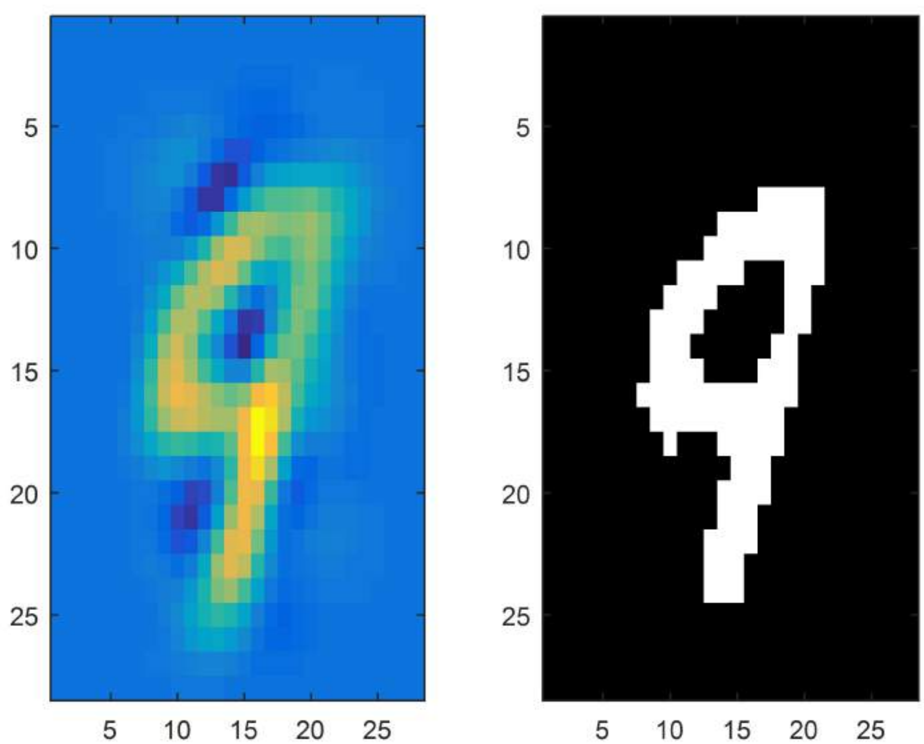


Plot of the probability distribution:

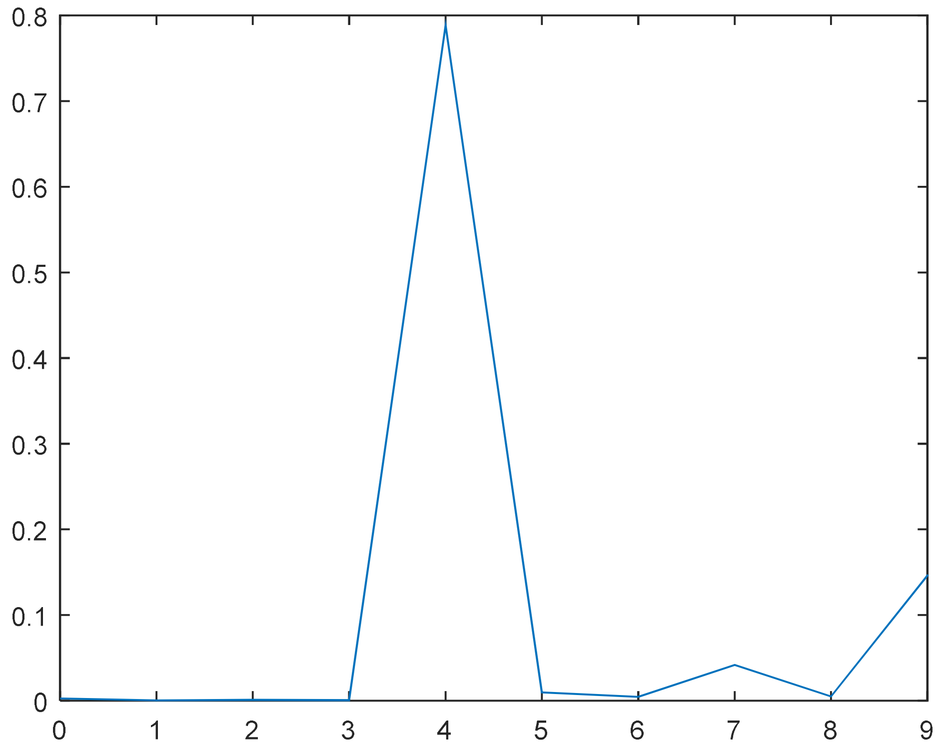


n=493

Actual digit = 9 ; Predicted digit = 4

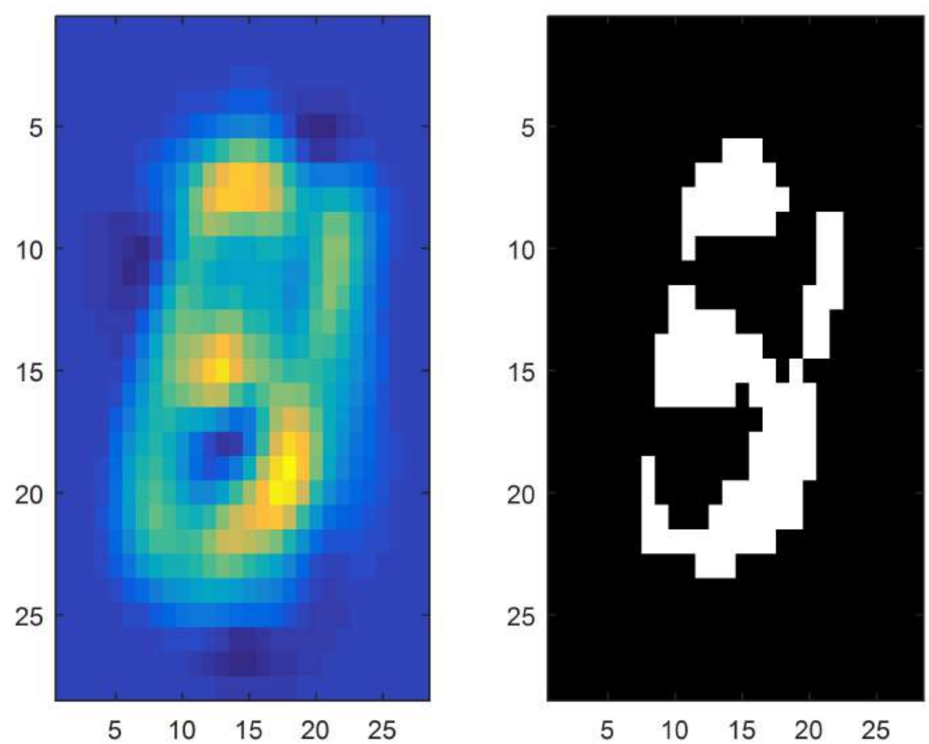


Plot of the probability distribution:



n=448

Actual digit = 8 ; Predicted digit =5



Plot of the probability distribution: