CSC311 HOMEWORK 4

Q1.

(1)

By Bayes' rule

$$\mathrm{p}(\mathrm{y}=\mathrm{k}\mid \mathbf{x},\boldsymbol{\mu},\boldsymbol{\sigma}) = \frac{\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k},\boldsymbol{\mu},\boldsymbol{\sigma})p(\mathrm{y}=\mathrm{k},\boldsymbol{\mu},\boldsymbol{\sigma})}{\mathrm{p}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\sigma})}$$
$$= \frac{\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k},\boldsymbol{\mu},\boldsymbol{\sigma})p(\mathrm{y}=\mathrm{k}\mid \boldsymbol{\mu},\boldsymbol{\sigma})p(\boldsymbol{\mu},\boldsymbol{\sigma})}{\mathrm{p}(\mathbf{x}\mid \boldsymbol{\mu},\boldsymbol{\sigma})\mathrm{p}(\boldsymbol{\mu},\boldsymbol{\sigma})}$$
$$= \frac{\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k},\boldsymbol{\mu},\boldsymbol{\sigma})p(\mathrm{y}=\mathrm{k}\mid \boldsymbol{\mu},\boldsymbol{\sigma})}{\mathrm{p}(\mathbf{x}\mid \boldsymbol{\mu},\boldsymbol{\sigma})}$$

By the law of total probability,

$$\mathrm{p}(\mathbf{x}\mid \boldsymbol{\mu},\boldsymbol{\sigma}) = \sum_{k'} \mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k}',\boldsymbol{\mu},\boldsymbol{\sigma})\cdot \mathrm{p}(\mathrm{y}=\mathrm{k}') = \sum_{k'} \alpha_{k'}\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k}',\boldsymbol{\mu},\boldsymbol{\sigma})$$

And since $p(\mathrm{y}=\mathrm{k}) = \alpha_k$ which doesn't depend on $\boldsymbol{\mu},\boldsymbol{\sigma}$, $\mathrm{y}=\mathrm{k}$ and $\boldsymbol{\mu},\boldsymbol{\sigma}$ independent. So $p(\mathrm{y}=\mathrm{k},\ \boldsymbol{\mu},\boldsymbol{\sigma}) = p(\mathrm{y}=\mathrm{k})\cdot \mathrm{p}(\boldsymbol{\mu},\boldsymbol{\sigma})$.

$$p(\mathrm{y}=\mathrm{k}\mid \boldsymbol{\mu},\boldsymbol{\sigma}) = \frac{p(\mathrm{y}=\mathrm{k},\ \boldsymbol{\mu},\boldsymbol{\sigma})}{p(\boldsymbol{\mu},\boldsymbol{\sigma})} = \frac{p(\mathrm{y}=\mathrm{k})\cdot \mathrm{p}(\boldsymbol{\mu},\boldsymbol{\sigma})}{p(\boldsymbol{\mu},\boldsymbol{\sigma})} = p(\mathrm{y}=\mathrm{k})$$

So

$$\mathrm{p}(\mathrm{y}=\mathrm{k}\mid \mathbf{x},\boldsymbol{\mu},\boldsymbol{\sigma}) = \frac{\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k},\boldsymbol{\mu},\boldsymbol{\sigma})p(\mathrm{y}=\mathrm{k}\mid \boldsymbol{\mu},\boldsymbol{\sigma})}{\mathrm{p}(\mathbf{x}\mid \boldsymbol{\mu},\boldsymbol{\sigma})} = \frac{\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k},\boldsymbol{\mu},\boldsymbol{\sigma})p(\mathrm{y}=\mathrm{k})}{\sum_{k'}\alpha_{k'}\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k}',\boldsymbol{\mu},\boldsymbol{\sigma})}$$
$$= \frac{\alpha_k \mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k},\boldsymbol{\mu},\boldsymbol{\sigma})}{\sum_{k'}\alpha_{k'}\mathrm{p}(\mathbf{x}\mid \mathrm{y}=\mathrm{k}',\boldsymbol{\mu},\boldsymbol{\sigma})}$$

(2)

$$\ell(\boldsymbol{\theta};\mathrm{D}) = -\log p\left(y^{(1)},\boldsymbol{x}^{(1)},\ldots,y^{(N)},\boldsymbol{x}^{(N)}\middle|\boldsymbol{\theta}\right)$$

since the data are iid,

$$\ell(\boldsymbol{\theta};\mathrm{D}) = -\log \prod_{i=1}^{N} p\left(y^{(i)},\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right)$$

By Bayes' rule

$$= -\log \prod_{i=1}^{N} \frac{p\left(\boldsymbol{x}^{(i)}\middle|y^{(i)},\boldsymbol{\theta}\right)p\left(y^{(i)},\boldsymbol{\theta}\right)}{p(\boldsymbol{\theta})} = -\log \prod_{i=1}^{N} \frac{p\left(\boldsymbol{x}^{(i)}\middle|y^{(i)},\boldsymbol{\theta}\right)p\left(y^{(i)}\middle|\boldsymbol{\theta}\right)p(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}$$

$$= -\log \prod_{i=1}^{N} p\left(\boldsymbol{x}^{(i)}\middle|y^{(i)},\boldsymbol{\theta}\right)p\left(y^{(i)}\middle|\boldsymbol{\theta}\right) = -\sum_{i=1}^{N} \log p\left(\boldsymbol{x}^{(i)}\middle|y^{(i)},\boldsymbol{\theta}\right) + \log p\left(y^{(i)}\middle|\boldsymbol{\theta}\right)$$

Substituting terms given in question, we get

$$\ell(\boldsymbol{\theta}; \mathrm{D}) = -\sum_{i=1}^{N} \log\left(\left(\prod_{j=1}^{D} 2\pi\sigma_j^2\right)^{-\frac{1}{2}} \exp\left\{-\sum_{j=1}^{D} \frac{1}{2\sigma_j^2}\left(x_j^{(i)} - \mu_{y^{(i)}j}\right)^2\right\}\right) + \log \alpha_{y^{(i)}}$$

$$= -\sum_{i=1}^{N} -\frac{1}{2}\sum_{j=1}^{D} \log 2\pi\sigma_j^2 - \sum_{j=1}^{D} \frac{1}{2\sigma_j^2}\left(x_j^{(i)} - \mu_{y^{(i)}j}\right)^2 + \log \alpha_{y^{(i)}}$$

$$= \frac{N}{2}\sum_{j=1}^{D} \log 2\pi\sigma_j^2 + \sum_{i=1}^{N}\sum_{j=1}^{D} \frac{1}{2\sigma_j^2}\left(x_j^{(i)} - \mu_{y^{(i)}j}\right)^2 - \sum_{i=1}^{N} \log \alpha_{y^{(i)}}$$

$$= \frac{N}{2}\sum_{j=1}^{D} \log 2\pi + \frac{N}{2}\sum_{j=1}^{D} \log \sigma_j^2 + \sum_{i=1}^{N}\sum_{j=1}^{D} \frac{1}{2\sigma_j^2}\left(x_j^{(i)} - \mu_{y^{(i)}j}\right)^2 - \sum_{i=1}^{N} \log \alpha_{y^{(i)}}$$

(3)

From (2) we know that

$$\ell(\boldsymbol{\theta}; \mathrm{D}) = \frac{N}{2}\sum_{j=1}^{D} \log 2\pi + \frac{N}{2}\sum_{j=1}^{D} \log \sigma_j^2 + \sum_{i=1}^{N}\sum_{j=1}^{D} \frac{1}{2\sigma_j^2}\left(x_j^{(i)} - \mu_{y^{(i)}j}\right)^2 - \sum_{i=1}^{N} \log \alpha_{y^{(i)}}$$

Therefore,

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \mu_{ki}} = \frac{\partial}{\partial \mu_{ki}} \sum_{j=1}^{N}\sum_{i\prime=1}^{D} \frac{1}{2\sigma_{i\prime}^2}\left(x_{i\prime}^{(j)} - \mu_{y^{(j)}i\prime}\right)^2$$

(here I change the i and j used in the previous summations to j and i' to avoid name confusion)

Since only terms that satisfy $y^{(j)} = k$ and i' = i contribute to the derivative,

$$= \sum_{j=1}^{N} 1\left(y^{(j)} = k\right) \frac{-2}{2\sigma_i^2}\left(x_i^{(j)} - \mu_{ki}\right) \; where \; 1(y^{(i)} = k) \; is \; an \; indicator \; function$$

$$= -\sum_{j=1}^{N} 1\left(y^{(j)} = k\right) \frac{1}{\sigma_i^2}\left(x_i^{(j)} - \mu_{ki}\right)$$

Similarly,

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \sigma_i^2} = \frac{\partial}{\partial \sigma_i^2}\left(\frac{N}{2}\sum_{i\prime=1}^{D} \log \sigma_{i\prime}^2 + \sum_{j=1}^{N}\sum_{i\prime=1}^{D} \frac{1}{2\sigma_{i\prime}^2}\left(x_{i\prime}^{(j)} - \mu_{y^{(j)}i\prime}\right)^2\right)$$

(here I change the i and j used in the previous summations to j and i' to avoid name confusion)

Since only terms that satisfy i' = i contribute to the derivative,

$$= \frac{\mathrm{N}}{2\sigma_i^2} - \sum_{j=1}^{N} \frac{1}{2\sigma_i^4}\left(x_i^{(j)} - \mu_{y^{(j)}i}\right)^2$$

Hence,

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \mu_{ki}} = -\sum_{j=1}^{N} 1(y^{(j)} = k) \frac{1}{\sigma_i^2} \left(x_i^{(j)} - \mu_{ki} \right)$$

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \sigma_i^2} = \frac{\mathrm{N}}{2\sigma_i^2} - \sum_{j=1}^{N} \frac{1}{2\sigma_i^4} \left(x_i^{(j)} - \mu_{y^{(j)}i} \right)^2$$

(4)

To find the maximum likelihood estimate for $\boldsymbol{\mu}$, we set $\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \mu_{ki}} = 0$ to find the maximum likelihood estimate for μ_{ki}.

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \mu_{ki}} = -\sum_{j=1}^{N} 1(y^{(j)} = k) \frac{1}{\sigma_i^2} \left(x_i^{(j)} - \widehat{\mu_{ki}} \right) = 0$$

$$\sum_{j=1}^{N} 1\left(y^{(j)} = k\right) \left(x_i^{(j)} - \widehat{\mu_{ki}} \right) = 0$$

$$\sum_{j=1}^{N} 1\left(y^{(j)} = k\right) x_i^{(j)} - \sum_{j=1}^{N} 1\left(y^{(j)} = k\right) \widehat{\mu_{ki}} = 0$$

Let n be the number of $y^{(j)}$s that are equal to k i.e. the number of data points whose class is assigned as k,

$$\sum_{j=1}^{N} 1\left(y^{(j)} = k\right) x_i^{(j)} - n\widehat{\mu_{ki}} = 0$$

$$\widehat{\mu_{ki}} = \frac{1}{n} \sum_{j=1}^{N} 1\left(y^{(j)} = k\right) x_i^{(j)}$$

Therefore,

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{j=1}^{N} 1\left(y^{(j)} = k\right) \boldsymbol{x}^{(j)}$$

where k is each class and n is the number of data points whose class label is k

To find the maximum likelihood estimate for $\boldsymbol{\sigma}$, we set $\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \sigma_i^2} = 0$ to find the maximum likelihood estimate for σ_i^2.

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathrm{D})}{\partial \sigma_i^2} = \frac{\mathrm{N}}{2\widehat{\sigma_i^2}} - \sum_{j=1}^{N} \frac{1}{2\widehat{\sigma_i^4}} \left(x_i^{(j)} - \mu_{y^{(j)}i} \right)^2 = 0$$

$$\mathrm{N}\widehat{\sigma_i^2} - \sum_{j=1}^{N} \left(x_i^{(j)} - \mu_{y^{(j)}i} \right)^2 = 0$$

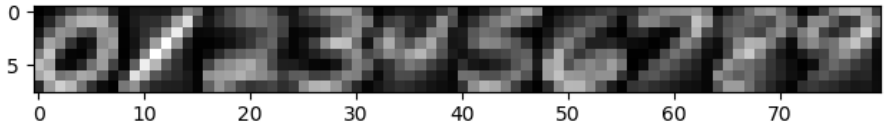
$$\widehat{\sigma_i^2} = \frac{1}{N} \sum_{j=1}^{N} \left(x_i^{(j)} - \mu_{y^{(j)}i} \right)^2$$

Therefore,

$$\widehat{\boldsymbol{\sigma}} = \frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{x}^{(j)} - \boldsymbol{\mu}_{y^{(j)}}\right)^2$$

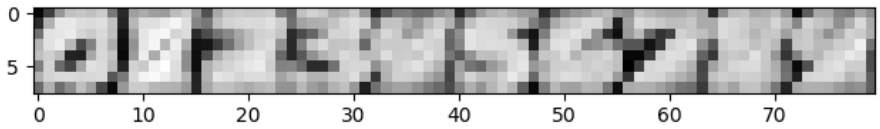
Q2.

2.0



2.1

(1)



(2)

The average conditional log-likelihood for train set:

-0.12462443666862928

The average conditional log-likelihood for test set:

-0.19667320325525448

(3)

The accuracy on the train set:

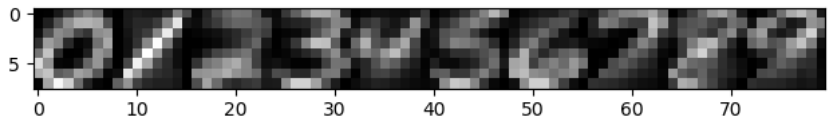
0.9814285714285714

The accuracy on the test set:

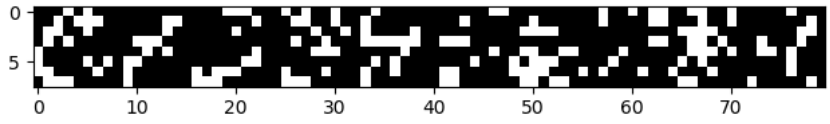
0.97275

2.2

(3)



(4)



(5)

The average conditional log-likelihood for train set:

-0.9437538618002541

The average conditional log-likelihood for train set:

-0.9872704337253588

(6)

The accuracy on the train set:

0.7741428571428571

The accuracy on the test set:

0.76425

2.3

Conditional Gaussian Classifier performs better than Naive Bayes Classifier which matches my expectation.

The accuracy of Naive Bayes Classifier highly depends on the assumption that the features are conditionally independent given the class. If the assumption is not correct, the accuracy of this classifier will be very low.

For this dataset, this assumption can't be hold. For example, for class 1, if a pixel is in the diagonal line and is set to ON, then its neighbors are also highly possible to be ON since that what a "1" looks like: pixels in the diagonal line is set to ON.

Therefore, the assumption can be matched. It is expected that Conditional Gaussian Classifier performs better than Naive Bayes Classifier.