CSC411: Assignment 2

Due on Sunday, November 12th, 2017

Student Name: **Gokul K. Kaushik**

Student Number: **999878191**

Table of Contents

1 - Class Conditional Gaussians

Q1-1 Use Bayes Rule to derive an expression for $p(y=k|\vec{x},\vec{\mu},\vec{\sigma})$

Using Bayes Rule for general probabilities:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

We know $p(\vec{x}|y=k,\vec{\mu},\vec{\sigma})$ from the question above.

Therefore

$$p(y=k|\vec{x},\vec{\mu},\vec{\sigma}) = \frac{P(y=k,\vec{x},\vec{\mu},\vec{\sigma})}{P(\vec{x},\vec{\mu},\vec{\sigma})} \quad (I)$$

Similarly, for $p(\vec{x}|y=k,\vec{\mu},\vec{\sigma})$, we get:

$$p(\vec{x}|y=k,\vec{\mu},\vec{\sigma}) = \frac{P(y=k,\vec{x},\vec{\mu},\vec{\sigma})}{P(y=k,\vec{\mu},\vec{\sigma})} \quad (II)$$

Therefore, from re-writing (II), we get

$$p(\vec{x}|y=k,\vec{\mu},\vec{\sigma}) \times P(y=k,\vec{\mu},\vec{\sigma}) = P(y=k,\vec{x},\vec{\mu},\vec{\sigma}) \quad (III)$$

Therefore, substituting the LHS of (III) with the numerator in the RHS of (I), we obtain:

$$P(y=k|\vec{x},\vec{\mu},\vec{\sigma}) = \frac{p(\vec{x}|y=k,\vec{\mu},\vec{\sigma}) \times P(y=k,\vec{\mu},\vec{\sigma})}{P(\vec{x},\vec{\mu},\vec{\sigma})} \quad (IV)$$

M1

Expanding this equation's subparts using Bayes' Theorem leads to:

$$P(y=k|\vec{x},\vec{\mu},\vec{\sigma}) = \frac{P(\vec{x}|y=k,\vec{\mu},\vec{\sigma}) \times P(y=k,\vec{\mu},\vec{\sigma})}{P(\vec{x},\vec{\mu},\vec{\sigma})} \quad (V)$$

But

$P(\vec{x},\vec{\mu},\vec{\sigma}) = P(\vec{x}|\vec{\mu},\vec{\sigma}) \times P(\vec{\mu}|\vec{\sigma}) \times P(\vec{\sigma})$ from Bayes Thm.

and

$P(y=k,\vec{\mu},\vec{\sigma}) = P(y=k|\vec{\mu},\vec{\sigma}) \times P(\vec{\mu}|\vec{\sigma}) \times P(\vec{\sigma})$ from Bayes Thm.

Therefore eqn (V) can be expanded to:

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

Cancelling out like terms gets us:

$$P(y=k|\vec{x},\vec{\mu},\vec{\sigma}) = \frac{P(\vec{x}|y=k,\vec{\mu},\vec{\sigma}) \times P(y=k|\vec{\mu},\vec{\sigma})}{P(\vec{x}|\vec{\mu},\vec{\sigma})} \quad (VI)$$

Now, using the hint provided (to use law of total probability for the denominator term, we obtain:

$$P(\vec{x}|\vec{\mu},\vec{\sigma}) = \sum_K P(x|y=k,\vec{\sigma},\vec{\mu}) \times P(y=k)$$

Therefore, equation (VI) becomes:

$$P(y=k|\vec{x},\vec{\mu},\vec{\sigma}) = \frac{P(\vec{x}|y=k,\vec{\mu},\vec{\sigma}) \times P(y=k|\vec{\mu},\vec{\sigma})}{\sum_K (P(x|y=k,\vec{\sigma},\vec{\mu}) \times P(y=k))} \quad (VII)$$

Making some further substitutions:

$P(y=k \mid \vec{\mu}, \vec{\sigma}) = P(y=k)$ as $P(y=k) = a_k$ regardless of given prior information about the distribution's mean and variance (from the question above).

Therefore, eqn (VII) becomes:

$$P(y=k \mid \vec{x}, \vec{\mu}, \vec{\sigma}) = \frac{P(\vec{x} \mid y=k, \vec{\mu}, \vec{\sigma}) \times P(y=k)}{\sum_K P(x \mid y=k, \vec{\mu}, \vec{\sigma}) \times P(y=k)}$$

$$\boxed{P(y=k \mid \vec{x}, \vec{\mu}, \vec{\sigma}) = \frac{P(x \mid y=k, \vec{\mu}, \vec{\sigma}) \times a_k}{\sum_K \left(P(x \mid y=k, \vec{\mu}, \vec{\sigma}) \times a_k \right)}} \quad \text{(VIII)}$$

A further representation can be performed by replacing the respective probabilities with their Gaussian Equations:

$$P(y=k \mid \vec{x}, \vec{\mu}, \vec{\sigma}) = \left(\prod_{i=1}^{i=D} 2\pi\sigma_i^2 \right)^{-1/2} \times \exp\left\{ -\sum_{i=1}^{i=D} \frac{1}{2\sigma_i^2} (x_i - \mu_{ki})^2 \right\} \times a_k$$

$$\frac{\left(\prod_{i=1}^{i=D} 2\pi\sigma_i^2 \right)^{-1/2} \times \exp\left\{ -\sum_{i=1}^{i=D} \frac{1}{2\sigma_i^2} (x_i - \mu_{ki})^2 \right\} \times a_k}{\sum_{k=1}^{K=K} \left[\left(\prod_{i=1}^{i=D} 2\pi\sigma_i^2 \right)^{-1/2} \times \exp\left\{ -\sum_{i=1}^{i=D} \frac{1}{2\sigma_i^2} (x_i - \mu_{ki})^2 \right\} \times a_k \right]} \quad \text{(IX)}$$

"capital K" "small k"

The Bayes formula/version of the solution is eqn (VIII).
The full Gaussian representation of the solution is eqn (IX).

A further simplification can be performed by removing the $\left(\prod_{i=1}^{i=D} 2\pi\sigma_i\right)^{-1/2}$ out of the main equation (in the denominator).

This cancels the same value in the numerator. Giving us:

$$p(y=k|\vec{x},\vec{\mu},\vec{\sigma}) = \frac{\exp\left\{-\sum_{i=1}^{i=D}\frac{1}{2\sigma_i^2}(x_i-\mu_{ki})^2\right\} * a_k}{\sum_{k=1}^{k=K}\left(\exp\left\{-\sum_{i=1}^{i=D}\frac{1}{2\sigma_i^2}(x_i-\mu_{ki})^2\right\} * a_k\right)}$$

Q1-2 Write down an expression for the negative log likelihood

NLL (negative log likelihood) = $\ell(\vec{\theta}; D) = -\log p(y^{(1)}, \vec{x}^{(1)}, y^{(2)}, \vec{x}^{(2)}, \ldots, y^{(N)}, \vec{x}^{(N)} | \vec{\theta})$

for $\vec{\theta} = \{\vec{\alpha}, \vec{\mu}, \vec{\sigma}\}$

Note: taking log base as e (i.e ln)

Therefore

$\ell(\vec{\theta}; D) = -\log p(y^{(1)}, \vec{x}^{(1)}, \ldots y^{(N)}, \vec{x}^{(N)} | \vec{\alpha}, \vec{\mu}, \vec{\sigma})$

Since the data is i.i.d, it can be treated as the product of each sample:

$$= -\log \left(\prod_{n=1}^{n=N} p(y^{(n)}, \vec{x}^{(n)}, \vec{\alpha}, \vec{\mu}, \vec{\sigma}) \right)$$

By Bayes Rule, this becomes:

$$= -\log \left(\prod_{n=1}^{n=N} p(y^{(n)} | \vec{x}^{(n)}, \vec{\alpha}, \vec{\mu}, \vec{\sigma}) \times p(\vec{x}^{(n)} | \vec{\alpha}, \vec{\mu}, \vec{\sigma}) \right)$$

From applying $\log(A \times B \times C) = \log A + \log B + \log C$ (identity), we can remove the multiplication operator and replace it w/ a sigma (Σ).

$$= \sum_{n=1}^{n=N} \left(\log \left(p(y^{(n)} | \vec{x}^{(n)}, \vec{\alpha}, \vec{\mu}, \vec{\sigma}) \times p(\vec{x}^{(n)} | \vec{\alpha}, \vec{\mu}, \vec{\sigma}) \right) \right)$$

From the previous question (Q1-1) we know what each value is.

Therefore, substituting it:

$$= \sum_{n=1}^{n=N} \log \left(\frac{\left(\prod_{i=1}^{i=D} 2\pi\sigma_i^2\right)^{-1/2} \times \exp\left\{ -\sum_{i=1}^{i=D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2 \right\} \times a_k}{\sum_{k=1}^{k=K} \exp\left\{ -\sum_{i=1}^{i=D} \frac{1}{2\sigma_i^2}(x_i \dots \right\}} \right)$$

"kay"

can't fit in the page

$$\times \left(\prod_{i=1}^{i=D} 2\pi\sigma_i^2 \right)^{-1/2} \sum_{k=1}^{k=K} \left[a_k \times \exp\left\{ -\sum_{i=1}^{i=D} \frac{1}{2\sigma_i^2}(x_i - \mu_i)^2 \right\} \right] \Bigg)$$

By cancelling out like terms, we obtain

$$= \sum_{n=1}^{n=N} \log \left[\left(\prod_{i=1}^{i=D} 2\pi\sigma_i^2 \right)^{-1/2} a_k^{(d)} \exp\left\{ \sum_{i=1}^{i=D} \frac{1}{2\sigma_i^2}\left(x_i - \hat{\mu}_{k^{(d)}i}\right)^2 \right\} \right]$$

This gives us the answer of:

$$\ell(\vec{\theta}; D) = \sum_{n=1}^{n=N} \left[\sum_{i=1}^{i=D} \left(\log(2\pi\sigma_i^2) + \frac{1}{2\sigma_i}\left(x_i - \hat{\mu}_{k^{(d)}i}\right)^2 \right) - \log\left(a_{k^{(d)}}\right) \right]$$

"kay"

Q1 – 3 Partial derivatives w.r.t μ_{ki} and σ_i^2 of log likelihood

$NLL = L(\vec{\theta}; D)$

$$\frac{\partial NLL}{\partial \mu_{ki}} = \frac{\partial}{\partial \mu_{ki}} \left(\sum_{n=1}^{n=N} \left(\sum_{i=1}^{i=D} + \left(\log(2\pi\sigma_i^2)\right) \right. \right. \left(\frac{1}{2\sigma_i^2} \left(x_i^{(n)} - \mu_{k^{(n)}i} \right)^2 (2)(-1) \right)$$

$$\left. \left. + \frac{1}{2\sigma_i^2} \left(x_i^{(n)} - \mu_{k^{(n)}i} \right)^2 \right. \right.$$

$$\left. - \left(\log(\pi_{k^{(n)}})\right) \right)$$

$\mu_{k^{(n)}i}$

$$\frac{\partial NLL}{\partial \mu_{ki}} = \sum_{n=1}^{n=N} \sum_{i=1}^{i=D} \frac{1}{\sigma_i^2} \left(x_i^{(n)} - \mu_{k^{(n)}i} \right)$$

However for μ_{ki} only exists when $k^{(n)} = k$. Therefore, this becomes

↳ i.e. 1 else 0

$$\frac{\partial NLL}{\partial \mu_{ki}} = \sum_{n=1}^{n=N} \left(\mathbb{1}(k^{(n)}, k) \frac{1}{\sigma_i^2} \left(x^{(n)} - \mu_{ki} \right) \right)$$

↳ 1 when $k^n == k$ else 0.

Similarly, for $\frac{\partial NLL}{\partial \sigma_i^2}$, we get

$$\frac{\partial NLL}{\partial \sigma_i^2} = \frac{\partial}{\partial \sigma_i^2} \sum_{n=1}^{n=N} \sum_{i=1}^{i=D} \frac{1}{2\pi\sigma_i^2} (2\pi) + \log(2\pi\sigma_i^2) \frac{-1}{2\sigma_i^4} \left(x_i^{(n)} - \mu_{k^{(n)}i} \right)^2$$

$$+ \frac{1}{2\sigma_i^2} \left(x_i^n - \mu_{k^n i} \right)^2$$

$$- \log(\alpha_k n)$$

0

$$= \sum_{n_i=1}^{n=N} \left(\frac{1}{2\sigma_i^2} - \frac{1}{2\sigma_i^4} \left(x_i^{(n)} - \mu_k{}^{n}{}_i \right)^2 \right)$$

$$\frac{\partial NLL}{\partial \sigma_i^2} = \frac{N}{2\sigma_i^2} - \frac{1}{2\sigma_i^4} \sum_{n=1}^{n=N} \left(x_i^n - \mu_k{}^{n}{}_i \right)^2$$

$x_i^n \in$ (x)

Q1-4 To get the MLEs for $\vec{\mu}$ and $\vec{\sigma}$, we have to set the partial derivatives of μ_{ki} and σ_i^2 obtained from the previous question to 0).

Note that since the MLE behaves well under transformations, and since the mapping of S.D σ to variance σ^2 is such a one-to-one function, the MLE (σ) will be the same as the MLE (σ^2).

Therefore:

$$MLE_{\mu_{ki}} = \frac{\partial NLL}{\partial \mu_{ki}} = 0$$

$$\Rightarrow \sum_{n=1}^{n=1} \left[\mathbb{1}(k^n, k) \frac{1}{\sigma_i^2} (x_i^n - \mu_{ki}) \right] = 0$$

$$\Rightarrow \boxed{\mu_{ki} = \frac{\sum_{n=1}^{n=N} \left(\mathbb{1}(k^n, k)\, x_i^n \right)}{\sum_{n=1}^{n=N} \left(\mathbb{1}(k^n, k) \right)}}$$

Similarly,

$$\frac{\partial MLE_{\sigma^2}}{\partial \sigma^2} = \frac{\partial NLL}{\partial \sigma^2} = 0$$

$$\frac{1}{2\sigma_i^2} - \frac{1}{2\sigma_i^2} \sum_{n=1}^{n=N} \left(x_i^n - \mu_{k^n i} \right)^2 = 0$$

$$\therefore \boxed{\sigma_i^2 = \frac{1}{N} \sum_{n=1}^{n=N} \left(x_i^n - \mu_{k^n i} \right)}$$

Therefore

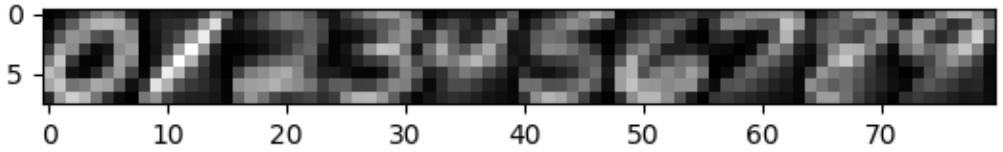
$$MLE_{\sigma} = \frac{1}{N} \sum_{n=1}^{n=N} \left(x_i^n - \mu_{k^n i} \right)$$

D1

2 - Handwritten Digit Classification

0 - Loading the data and Plotting the Feature Means

The means (from 700 samples per digit) for each feature (64 features in total for an 8-by-8 pixel image) for 10 digits (digit 0 to digit 9) are plotted below:



1 - K-NN Classifier

2.1.1 Train and Test Classification Accuracy for K=1 and K=15

Training Data Classification Accuracy for K=1: 1.0
Training Data Classification Accuracy for K=15: 0.961
Test Data Classification Accuracy for K=1: 0.96875
Test Data Classification Accuracy for K=15: 0.959

2.1.2 Tie Breaker Method

There are cases in K-Nearest Neighbours where there isn't one most frequent neighbours (there might be two neighbours that occur equally frequently). Therefore, in such cases a tie breaking decision needs to be made. I have chosen to reduce the number of nearest neighbours by one - **effectively removing the last neighbour and repeating the check until a decision can be made**.

This method was chosen because:

1. The tie remains until the *most distant* neighbour from one of the most frequent neighbours is removed. This rewards the neighbour with the closest value in such cases.
2. This decision making method is intuitive to understand and easy to implement.

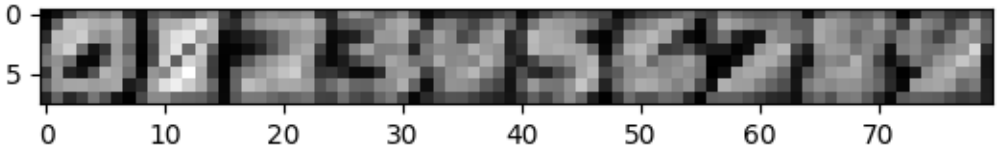
2.1.3 Optimal K

The optimal K from applying 10-folds on the data set is K=3.

Training Data Classification Accuracy for K=3 on 10-folds: 0.9865
Test Data Classification Accuracy for K=3 on 10-folds: 0.9697

2 - Conditional Gaussian Classifier Training

2.2.0 Plot of the Diagonal Elements of the Covariance Matrix



2.2.1 Average Conditional (Log) Likelihoods

The average conditional (log) likelihoods were compute for the train set and the test set are:

Train Data Set: -0.124624436669
Test Data Set: -0.196673203255

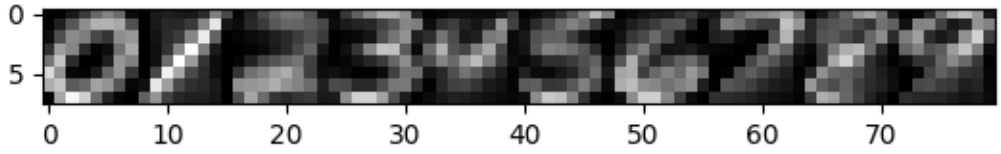
2.2.3 Accuracy for the Most Likely Posterior Class

For the most likely posterior class, the training and test set accuracies are:

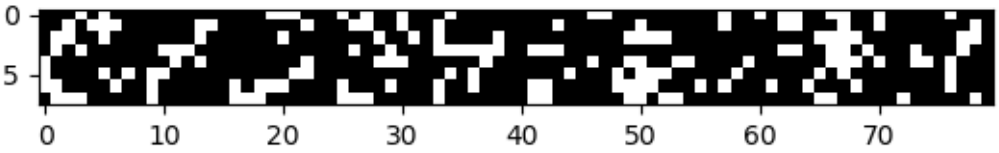
Train Data Set: 0.9814
Test Data Set: 0.9727

3 - Naive Bayes Classifier Training

2.3.3 Plot of Eta



2.3.4 Plot of Generated Binarized Data (Using a Binomial Disribution)



2.3.5 Average Conditional (Log) Likelihoods

The average conditional (log) likelihoods were compute for the train set and the test set are:

Train Data Set: -0.9437538618
Test Data Set: -0.987270433725

2.3.6 Accuracy for the Most Likely Posterior Class

For the most likely posterior class, the training and test set accuracies are:

Train Data Set: 0.7741
Test Data Set: 0.7642

4 - Model Comparison

The models performed from (best to worst):

1. Conditional Gaussian Classifier
2. K-Nearest Neighbours
3. Naive Bayes Classifier

The superior performance of the conditional Gaussian Classifier over the Naive Bayes Classifier would make sense as Naive Bayes assumes more assumptions (in terms of independence among features, which did not exist).
The K-NNs surprisingly performed worse as the number of neighbours increased. One would have assumed that increasing the size of K would smooth the decision plane boundaries. One reason could be that the number of samples per digit (700) was too small to see an improvement in performance.