COM2028/16/Semr2 16/17

UNIVERSITY OF SURREY ©

Faculty of Engineering and Physical Sciences

Department of Computer Science

Undergraduate Programmes in Computing
Undergraduate Programmes in Mathematics

Module COM2028: 15 credits

Artificial Intelligence

FHEQ Level 5 Examination

Time allowed: Two hours

Semester 2 2016/17

Answer **ALL** questions

Approved calculators ARE permitted

Questions 1-4 carry 20, 35, 30, and 15 marks respectively

Where appropriate the mark carried by an individual part of a question is indicated in square brackets [ ]

COM2028/16/Semr2 16/17

Solutions

1. This question is about *Image Processing.*

(a) Explain how to transform the images on the top to the their corresponding images on the bottom of the same column in Figure 1, via the respective transformation function shown in the middle image. In your explanation, compare the left transformation with one on the right.

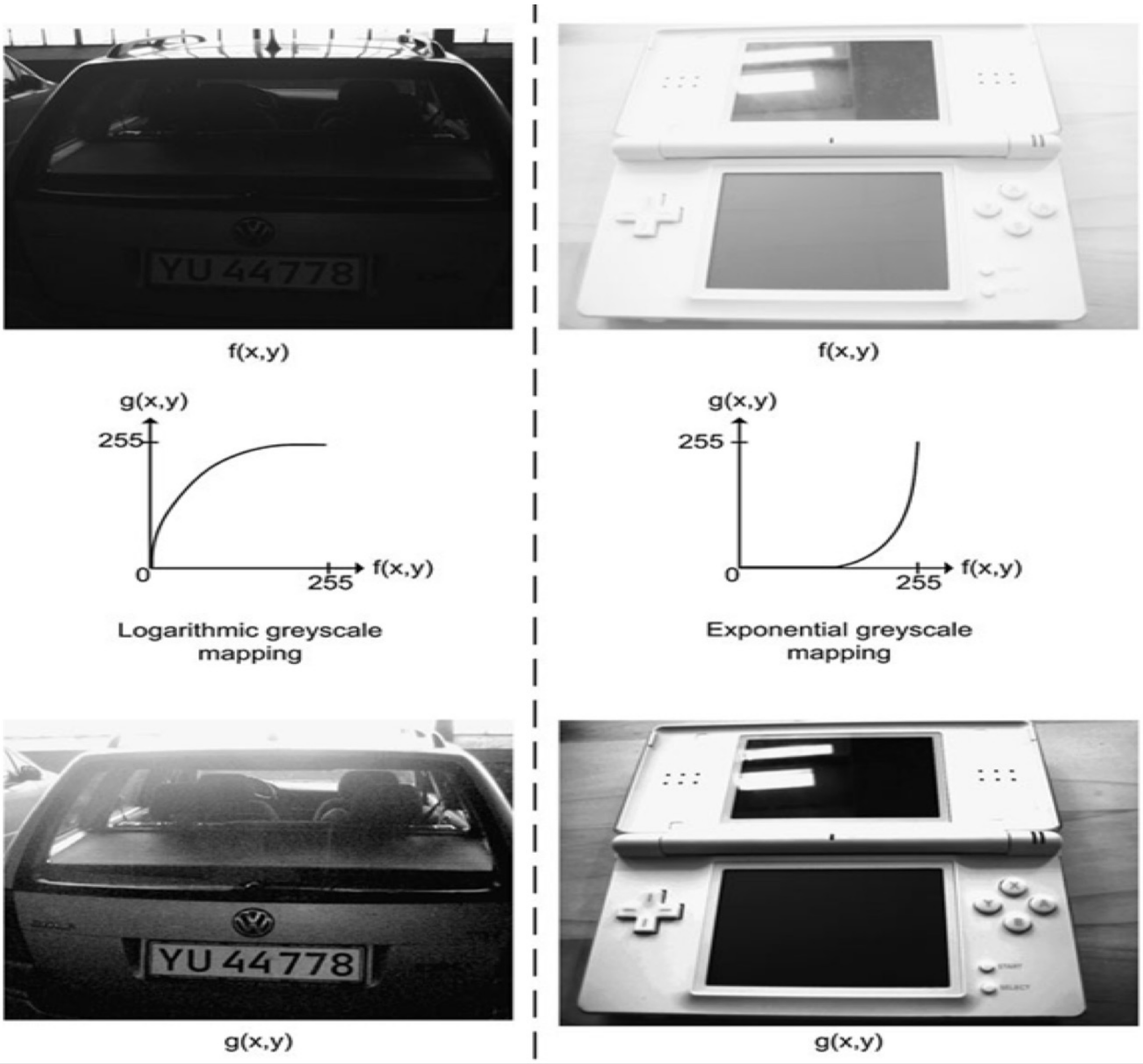

Figure 1: Image transform using two types of mapping functions

[5 marks]

Generally, logarithmic mapping is to enhance details in the darker regions of the image, at the expense of detail in the brighter regions. Exponential mapping has a reverse effect, contrast in the brighter parts of an image is increased at the expense of contrast in the darker parts.

(b) The following is a algorithm for implementing histogram equalisation using the cumulative histogram of a grey image as a mapping function.

```
Given $f(x, y)$ representing the grey level intensity for pixel $(x, y)$ in an image
Compute a scaling factor, $a = 255/\text{number of pixels}$
Calculate the histogram of the grey image, histogram $[i]$, where $0 < i < 255$
$c[0] = a{*}\text{histogram}[0]$
for all remaining grey levels, $i$, do
    $c[i] = c[i-1] + a{*}\text{histogram}[i]$
end for
for all pixel coordinates, $x$ and $y$, do
    $g(x, y) = c[f(x, y)]$
end for
```

Explain when applying this algorithm to an input image, what effect will be on the image.

[5 marks]

The cumulative histogram of the image has such properties of having a steep slope ($a > 1$) at grey levels that occur frequently, and a gentle slope ($a < 1$) at unpopular grey levels. Thus when using the cumulative histogram as a mapping function, it will increase contrast for the most frequently occurring grey levels and reduce contrast in the less popular part of the grey level range.

(c) Given an image I, and three convolution kernels, A, B and C.

$$I = \begin{pmatrix} 2 & 1 & 1 & 2 \\ 0 & 1 & 0 & 3 \\ 1 & 3 & 2 & 5 \\ 3 & 0 & 2 & 5 \end{pmatrix} A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} B = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & 1 & -1 \end{pmatrix} C = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

(i) Apply A to I, give the convolution result of I.

[5 marks]

(i)

$$I' = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$$

(ii) Between B and C, which will obtain the edge information of I and which will make I sharper?

[5 marks]

B will transform I to an edge image, while C will sharpen the image.

Total marks: 20

2. This question is about *Unsupervised Machine Learning.*

(a) By clustering blogs, it might be possible to determine if there are groups of blogs that frequently write about similar subjects or write in similar styles. Such a result could be very useful in searching, cataloging, and discovering the huge number of blogs that are currently online. Assume we have collected 5000 web blogs and we would like to cluster them into three groups using k-means clustering method. Answer the following questions:

A. Propose your approach to processing the blogs so that an appropriate data representation or features can be obtained to be fed into the clustering process. [3 marks]

B. How to train the k-means clustering? [6 marks]

C. How to assign a new blog to one of the clusters after the training? [3 marks]

A. Data processing and representation: We can first generate word counts for each blog, based on keywords that are within maximum and minimum percentages or calculate n-grams, or tf-idf to form a feature vector for each blog.

B. K-means clustering begins with k randomly placed centroid feature vectors (points in space that represent the centre of the cluster), and assigns the feature vector of each blog to the nearest centroid. After the assignment, the centroids are moved to the average location of all the nodes assigned to them, and the assignment are redone. The process repeats until the assignments stop changing.

C: Testing: When there is a new input blog coming in, we can calculate the feature vector in the same way of processing the training data, then calculate the Euclidean distance between this vector to the centroid of each cluster. The shortest distance means the blog will be assigned to that cluster.

(b) Given a dataset of 6000 handwritten digits as shown in Figure 2, describe how to use Self Organizing Map (SOM) to catagorise such data into 10 groups. Each digit image is of $8x8$ in size. Include in your solution the following components:

A. How to represent the digit data? [3 marks]

B. How to train the SOM? [12 marks]

C. How to assign a new digit image to one of the groups after the training? [3 marks]

SEE NEXT PAGE

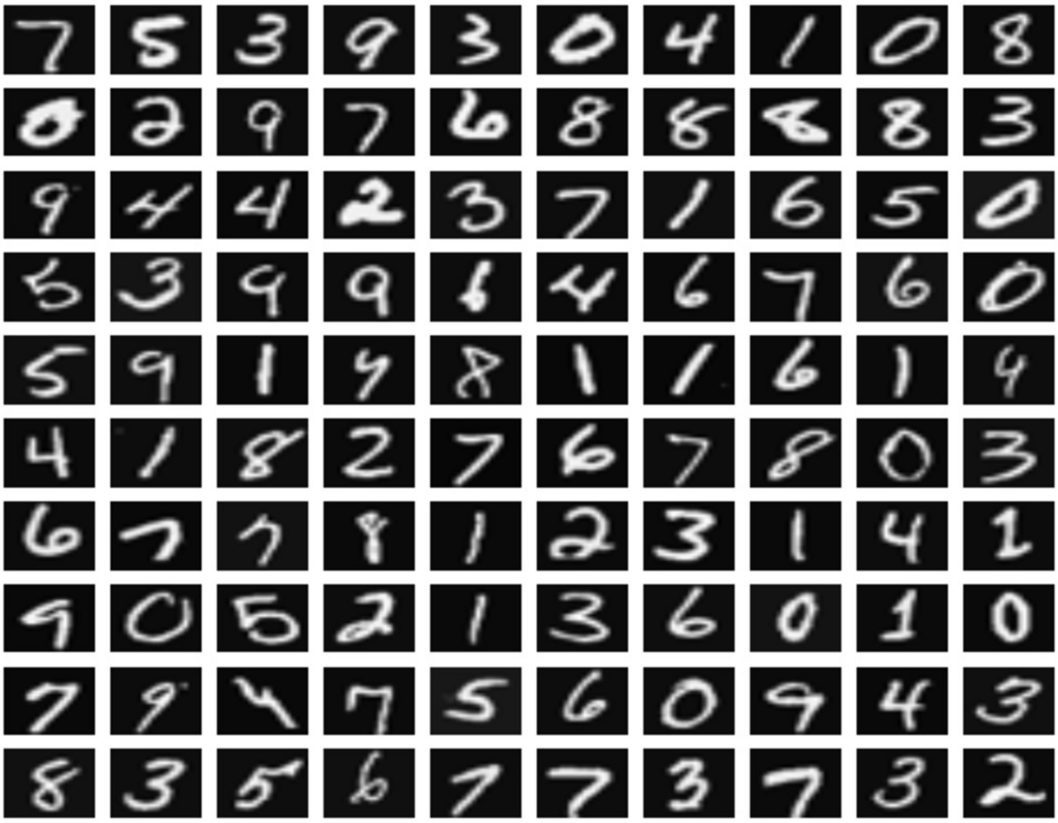

Figure 2: The digit image dataset

A. Data representation: Each digit image is represented in a feature vector (64 in size) hold the intensities of the pixels.

B. On SOM, student can either describe using plain English or in the form of algorithm.

Training:

Initialization (Randomize the map's nodes' weight vectors, which have the same dimension of the feature vector of training data.) (3 marks)

Loop:

{Given an input vector of an digit image, measure the similarity between the input vector and the map's node's weight vector using Euclidian distance. (3 marks)

Track the node that produces the smallest distance (BMU) at iteration p.

Update the nodes in the neighbourhood of BMU by pulling them closer to the input vector. The degree of such pulling is decided by a learning rate. (3 marks)

Go back to the beginning of the loop to train with another image until the minimum distance is satisfied, or the map is stabilized. (3 marks) }

C. Testing: Given any testing datum, calculate the Euclidean distance between the testing datum and the weight neurons. The closest neuron represent the group that the testing datum belongs to.

(c) Why both above approaches, k-means clustering and SOM, are considered to be unsupervised machine learning?

[5 marks]

This is because when we train the algorithm, there are no labels associated to the data. The algorithms establish the relationship between the data to form groups or discover patterns.

Total marks: 35

3. This question is about *Multilayer Perceptron Neural Networks (MLP), Optimisation and Support Vector Machines (SVM).*

(a) Design a simple Multilayer Perceptron Neural Network for an Optical Character Recognition (OCR) system, as shown in Figure 3.

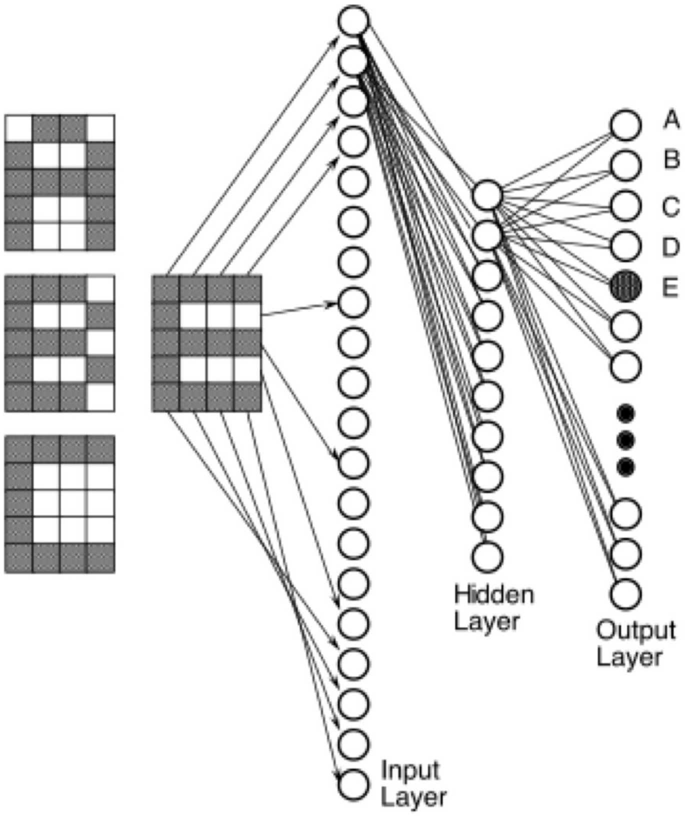

Figure 3: The architecture of the neural network for an OCR system

Answer the following questions:

A. What is the size of the input dimension and what are in the feature vector? [3 marks]

B. How are the hidden neurons and output neurons calculated? Use an example if that helps the explanation. [6 marks]

C. Explain the backpropagation process through an example. [7 marks]

A. 20. The feature vector contains the pixel intensities of the input image either for training or testing.

B. It is based on feedforward, ie, the multiplication of the vector of previous layer and corresponding weight vector is first calculated, the result will go through an activation function, then feed to the next layer.

C. For example, if the current training sample is a letter E, the corresponding neuron at the output layer (5th neoron) should be fired or having the highest value. If not, the error between the correct output and the current output will be calculated and the error gradient will be propagated back through the output layer first then the hidden layer to update the weights. The same procedure is repeated when inputing other training samples. This is to minimise the overall error across all train samples.

(b) If we use an optimisation technique such as Hill Climbing, Simulated Annealing, or Genetic Algorithms, to optimise the architecture of the above neural network.

A. What in the neural network could be optimised? Give at least two examples. [4 marks]

B. What could be the objective function (cost function) for such optimisation. [5 marks]

C. Give at least one key difference between a genetic algorithm and the other two optimisation approaches listed above? [3 marks]

A. number of nodes in the hidden layer, the number of hidden layer, initial weights.

B. The objective function can be the accuracy of the validation data.

C. GA uses a population of solutions while the rest will consider one solution at a time.

(c) If we use Support Vector Machines (SVM) to implement the OCR system, discuss the following key elements in the SVM system.

A. SVMs are inherently two-class classifiers. How to deal with this multi-class problem through SVMs? [4 marks]

B. What does the maximum margin hyperplane mean? [3 marks]

A. SVMs are inherently two-class classifiers. One way to do multiclass classification with SVMs to build one-versus-rest classifiers and to choose the class which classifies the test datum with greatest margin. Another strategy is to build a set of one-versus-one classifiers, and to choose the class that is selected by the most classifiers.

B. We need to select the hyperplane which separates the two classes better. Compared with any other hyperplane, maximum margin hyperplane has the longest distance between the nearest data point (from either class) and the hyperplane.

Total marks: 35

4. This question is about *the use of a Gaussian Classifier and Naïve Bayesian Classifier.*

In Gaussian classifier, the Gaussian probability density function for category w_i is given as below:

$$P(x \mid w_i) = [(2\pi)^d \mid \sum_i |]^{-\frac{1}{2}} \exp[-\frac{1}{2}(x-u_i)^T \sum_i^{-1} (x-u_i)] \tag{1}$$

where

u_i — the mean vector of class w_i,

$\sum_i$ — the ith class covariance matrix.

x is the feature vector.

d is the dimension of x.

u_i and $\sum_i$ are calculated from training samples belonging to category w_i

$u_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_j, x_j \in w_i$, where N_i is the number of training samples from class w_i.

The covariance matrix as

$$\sum_i = \frac{1}{N_i} \sum_{j=1}^{N_i} (x_j - u_i)(x_j - u_i)^T \tag{2}$$

A. Describe how a Gaussian Classifier should be trained and how to evaluate if it has learned the patterns in the data. [8 marks]

B. What is the common approach that both a Gaussian Classifier and a Naïve Bayesian Classifier have? [4 marks]

C. What is the key difference between a Gaussian Classifier and a Naïve Bayesian Classifier? [3 marks]

A, During the training processing, u_i and $\sum_i$ are calculated for each category using the data from each group. C, Testing stage

Based on Bayes Theorem, we have

$$\begin{aligned} P(x, w_i) &= P(x) * P(w_i \mid x) = P(w_i) * P(x \mid w_i) \\ \Rightarrow P(w_i \mid x) &= \frac{P(w_i) * P(x \mid w_i)}{P(x)} \end{aligned} \tag{3}$$

To simplify the problem here, we suppose $P(x)$ and $P(w_i)$ are scale factors or of the same constants value, so

$$P(w_i \mid x) = \lambda \, P(x \mid w_i), \tag{4}$$

where λ is a constant to make sure $\sum\limits_i P(w_i \mid x) = 1$

After the training stage, we have obtained Gaussian function for each category $w_i : P(x \mid w_i)$. When testing on an unknown sample x, from the above theoretical inference, we get the posterior probability: $P(w_i \mid x)$.

Assign x to w_k where $P(w_k \mid x)$ is the largest value among $P(w_i \mid x)$.

B. In testing stage, both use Bayes Theorem in order to infer the posterior knowledge based on prior knowledge, i.e., given any testing datum x, we would like to know which classes or category it belongs to, and this can be done through:

$$P(w_i \mid x) = \frac{P(w_i) * P(x \mid w_i)}{P(x)} \tag{5}$$

where $p(x \mid w_i)$ has been known through learning. And often $P(w_i)$ and $P(x)$ are treated as scale factors.

C. In training phase, a Gaussian classifier measure the mean and covariance of each class based on training samples, while in a Naïve Bayesian Classifier, the training process takes the training data and their class labels to calculate the probabilities that each feature in the feature vector is associated with a particular class label based on all the training samples in each class, generating the probability about a certain class will contain a given feature, $p(x^{(k)} \mid w_i)$, where k is the kth feature in the feature vector.

The feature probabilities need to be combined into a single probability for the entire piece of data x.

$$P(x \mid w_i) = p(x^{(1)} \mid w_i) * p(x^{(2)} \mid w_i) * \cdots \tag{6}$$

Total marks: 15

END OF PAPER

INTERNAL EXAMINER: DR. H.L.TANG
EXTERNAL EXAMINER: PROF. DAVID MARSHALL