COM2028/14/Semr2 15/16

UNIVERSITY OF SURREY ©

Faculty of Engineering and Physical Sciences

Department of Computing

Undergraduate Programmes in Computing
Undergraduate Programmes in Mathematics

Module COM2028: 15 credits

Artificial Intelligence

FHEQ Level 5 Examination

Time allowed: Two hours

Semester 2 2015/16

Answer **ALL** questions

Approved calculators ARE permitted

Questions 1-4 carry 35, 25, 15, and 25 marks respectively

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

COM2028/14/Semr2 15/16

Solutions

1. This question is about *general understanding of the subject area.*

   Note : Some of the questions have multiple choices. Please indicate in your answer book the correct answers, noting that some questions may have more than one correct answers. Give a brief justification for your answers whenever appropriate.

   (a) Examining the image in Figure 1, what is the best option to remove both salt and pepper noise (i.e. those tiny white and black pixels with intensity values of 255 and 0)?

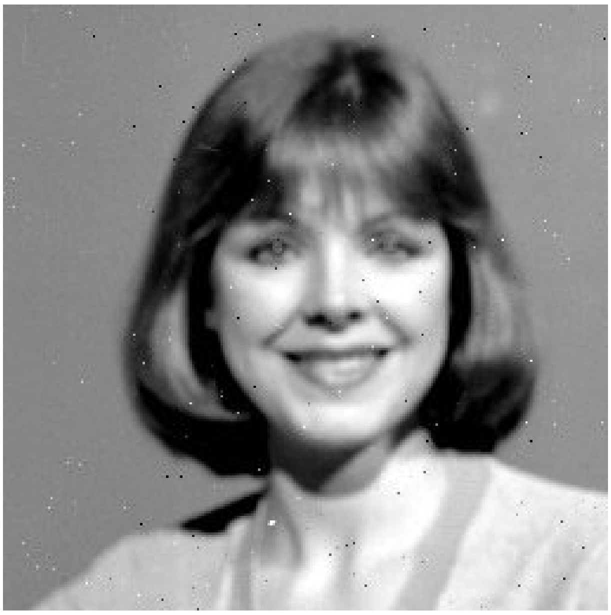

Figure 1: A picture with salt and pepper noise

A. A mean filter

B. A median filter

C. Brightness enhancement

D. A Gaussian filter

[3 marks]

B, median filter

Justification: median filter will remove salt and pepper noise as it only chooses the intensity ranked in the middle among all pixels specified by the filter.

(b) Which one or more of the following are true about a cumulative histogram?

A. It has a steep slope at intensity levels that occur less frequently.

B. It has a steep slope at intensity levels that occur more frequently.

C. The cumulative histogram of an image can be used as a mapping function to make the histogram spread more evenly.

D. It can be used as a feature for classification.

[4 marks]

B, C, D

Justification: A is not true because a steep slope means a sharp rise of the number of pixels which have that intensity.

(c) Examine the following convolution kernels, which can be used to remove some noise from an image?

Convolution kernels:

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix} B = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} C = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & 1 & -1 \end{pmatrix} D = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

[4 marks]

B

Justification: B is a mean filter that averages the neighbouring pixels thus smooth the image.

(d) Examine the following applications, choose those that will benefit from supervised learning.

A. Discover similar groups from 20,000 web pages.

B. Design an autonomous vehicle that can steer on its own based on the road images captured through the camera.

C. Automatic generating a portion of yourself.

QUESTION 1 CONTINUES ON NEXT PAGE

D. Classify pictures of human and animals with 1,000,000 labelled training samples already available.

E. Face recognition.

F. Spam filtering.

[6 marks]

B, D, E, F

Justification: A needs unsupervised learning while C needs optimisation algorithms.

(e) Given an image with dimensions 100 x 100. The parameters for its feature descriptor using Histogram of Oriented Gradients (HOG) are: orientations=8, pixels per cell = 10 x 10, cells per block = 2 x 2. Estimate the dimension of its HOG feature descriptor.

A. 32

B. 2500

C. 2592

D. 12800

[3 marks]

C, 9x9x4x8= 2592

(f) Which of the following can be included in a genetic algorithm?

A. a mutation operation

B. a cost function

C. a hill climbing neighbouring search

D. a crossover operation

E. a solution representation

F. a temperature control parameter

[6 marks]

A, B, D, E

Justification: C is a different optimisation approach and F is related simulated annealing.

(g) Which of the following can describe k-means clustering algorithm?

A. There are k nearest neighbours that can help to make classification decision.

B. The k cluster centre is updated each time when the cluster members are changed.

C. The top k closest clusters are chosen and the distances between each two clusters are calculated.

D. The algorithm will stop when all centres merge into one.

[4 marks]

B

Justification: B is true as it describes how the cluster centre is updated when re-calculating membership for each group.

(h) Which of the following can describe Support Vector Machines (SVMs) and k-nearest mean algorithm?

A. They both belong to supervised learning category.

B. SVMs need to find the support vectors that determine maximum-margin hyperplane.

C. They both need to find the centre of a class.

D. In SVMs, the data samples are usually transformed to another feature space so that the data of different classes are divided by a minimum-margin hyperplane.

E. In SVM, the data samples are usually mapped to another feature space so that the data of different classes are divided by a clear gap that is as wide as possible.

[5 marks]

A, B, E

Justification: C is not true as SVMs doesn't need to find the centre of a class. D is wrong because E is correct.

Total marks: 35

2. This question is about *Neural Networks.*

(a) The MNIST database (Mixed National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems. Each digit image has 28x28 dimensions. A typical neural network to solve the MNIST task might have 784 inputs (pixels) connected to 1,000 hidden neurons, which are in turn connected to 10 output targets (one for each digit). Each layer is fully connected to the layer above (so each input pixel has a set of weights that connect it to each neuron, etc.).

A. Draw the architecture of the neural network. [5 marks]

B. How many weights does this network have? [4 marks]

C. What are in the feature vector? [2 marks]

D. Explain the learning process. [5 marks]

E. Explain the testing process, i.e. given an unlabelled testing digit image which has not been seen by the neural network before, how does the network decide a target class for this digit image? [5 marks]

(b) Could this system be implemented using another machine learning technique or a different neural network architecture? Provide a brief discussion on any alternative approach. [4 marks]

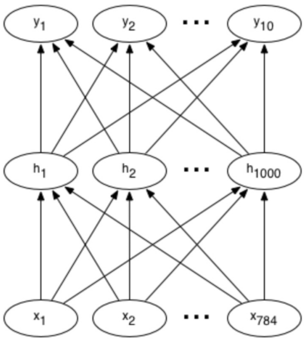

Figure 2: The architecture of the neural network

A. The architecture of the system.

See Figure 2

B. The weights:

784x1000+1000x10 =794000

C. The feature vector: pixel intensity values.

D. The learning process.

Training images together with their labels as pairs are fed into the neural network for training. The input layer has 794 node, the hidden layer has 1000 hidden nodes. There are 10 output nodes each representing a digit class. Initially the weights on all connections are randomly assigned. Through feeding forward, the output from each neuron in hidden layer and output layer is calculated based on the current weight. Back propagation is carried out based on the error between the current output and desired output and the weights are adjusted. The training is repeated as such until converged.

E. The testing process, i.e. when inputting an unlabelled digit image, the system will use the raw pixel values as feature vector, feed them into the neural network, calculate the value at each neuron based on learned weights, and check the most activated output node for corresponding digit class.

(b) kNN could be used for this task. Another neural network is convolutional neural network which should lend better classification accuracy. (note, students can propose different techniques as long as it is justifiable.)

Total marks: 25

3. This question is about *the use of a Gaussian Classifier.*

Assume there are a set of wearable devices to collect daily data from people, such as body temperature, heart beat, blood pressure, total hours of sleep, alcohol intake level, smoking behaviour, physical movement patterns etc. Design a Gaussian classifier to first learn healthy and unhealthy categories based on existing 5000 labelled data then classify newly collected 1000 unknown data.

The Gaussian probability density function for category w_i is given as below:

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

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

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

Describe your proposed system with regards to the following aspects:

A. How to process and prepare the data for training? [3 marks]

B. Describe the training process, and explain, after the system is trained, how does the system hold the knowledge about each category? [6 marks]

C. How to test the system? [6 marks]

A, The collected information such as temperature, blood pressure etc. will first be converted to values forming feature vectors. Data labelled as healthy and unhealthy are in different groups as training data for the two categories.

B, During the training processing, u_i and $\sum_i$ are calculated for each category using the data from each group. $P(x \mid w_1)$ for healthy class w_1 and $P(x \mid w_2)$

for unhealthy class w_2 according to equation (1) will be obtained as knowledge about each class.

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 constants, and $P(w_1) = P(w_2)$, so

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

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

After the training stage, we have obtained one Gaussian function for category $w_1 : P(x \mid w_1)$, and the another Gaussian function for category $w_2 : P(x \mid w_2)$. When testing on an unknown sample x, from the above theoretical inference, we get the posterior probability: $P(w_1 \mid x)$ and $P(w_2 \mid x)$.

If $P(w_1 \mid x) > P(w_2 \mid x)$, assign x to w_1; otherwise assign x to w_2.

Total marks: 15

4. This question is about *Machine Learning and Optimisation.*

   Design a system that classify all the tweets posted on Twitter each day under the same news trend. Each tweet is classified into positive or negative category. Large number of historical tweets have been collected regardless whichever trend they belong to, but have been labelled as positive and negative.

   Propose a Naïve Bayesian Classifier that is able to perform such classification task for twitter data. Answer the questions and complete the descriptions as below.

   A. Describe your choice of the features to be used in the system. [3 marks]

   B. Describe the training process. [6 marks]

   C. Describe the classification process, i.e.. when a new tweet posted, classify it as a positive or negative tweet. [6 marks]

   D. If you are unsure whether your choice of features is effective, it is possible to improve it using optimisation techniques. Describe the details of your approach for such optimisation process. This should include

      (i). your design of an objective function, [3 marks]

      (ii). and the key steps of your optimisation algorithm. [7 marks]

   A. features: bag of keywords, or 2 or 3-gram together with tf-idf can be used to generate features for each tweet.

   B. Describe the training process.

      Those historical tweets with positive and negative labels are first processed in order to extract the features for each tweet.

      Training stage: Like all supervised methods, a Bayesian classifier is trained with examples. Each example is a list of the features for each tweet and the label for that tweet. We can create a pure histogram of n-gram (or a combination of tf-idf and n-gram) as the feature vector for each tweet.(2 marks) The training process takes a tweet and its label. It then calculate the probabilities that the features are associated with a particular classification based on all the training samples in each category, generating the probability that a tweet about a certain category will contain a given feature. (4 marks)

   C. Describe the classification process. After a Bayesian classifier has been trained, it can be used to automatically classify new items (new tweet).

The feature probabilities need to be combined into a single probability for the entire item.

$$Pr(Category \mid tweet) = Pr(tweet \mid Category) * Pr(Category)/Pr(tweet) \tag{5}$$

(2 marks) Where

$$Pr(tweet \mid Category) = Pr(Feature1 \mid Category) * Pr(Feature2 \mid Category) * \cdots \tag{6}$$

Where $Pr(Category)$ is the overall frequency of the category and $Pr(tweet)$ is a scale factor. Whichever category gets a higher score for $Pr(Category \mid tweet)$ is the predicted category (positive, or negative). (4 marks)

D. Describe the objective function and the optimisation algorithm.

(i) Objective function can be measured using a set of training samples with ground truth based on certain choice of features such as just a bag of keywords, or n-gram with a random n, with or without tf-idf. Run through a Naïve Bayesian Classifier to check the classification results. The objective function is to calculate the number of time that the classifier correctly classifies the tweet. The score indicates the quality/cost of the classifier with its current choice of features.

(ii) Genetic Algorithm (or other optimization algorithm)can be used. The chromosome for each solution can be current choice of features. (2 marks) A number of solutions can form the initial population, then for each solution, the set of testing samples are used to get scores through the objective function. The solutions are then ranked. Top solutions are selected based on predefined percentage of such selection or a subset of solutions are selected based on Roulette wheel selection process. Such subset of solutions will be used in the next generation.(3 marks) Some new solutions are generated through crossover and mutation of the top solutions, as well as those from random generation, and these make up the full set of solutions for the next generation. The objective function will apply to the next generation again and the whole process repeats until a satisfactory the top score cannot improved any more. (2 marks)

Total marks: 25

END OF PAPER

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