A30406

Calculators permitted provided they are not capable of being used to store alphabetical information other than hexadecimal numbers.

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Final Year - BSc Artificial Intelligence and Computer Science
Undergraduate Affiliate Computer Science/Software Engineering
Final Year – BSc Computer Science
Third Year – MSci Computer Science
Final Year – MEng Computer Science/Software Engineering
Final Year – BSc Computer Science with Study Abroad
Fourth Year – BSc Mathematics with Year in Computer Science
Third Year – MSci Computer Science with Industrial Year
Final Year – BSc Computer Science with Industrial Year
Final Year – MEng Computer Science/Software Engineering with Industrial Year
Final Year – BSc Artificial Intelligence and Computer Science with Industrial Year

06 26428

Machine Learning

Summer May/June Examinations 2017

Time allowed: 1 hour 30 minutes

[Answer ALL Questions]

TURN OVER

1. (a) Describe the difference between a naïve Bayes classifier and a typical Bayes classifier. [3%]

(b) Consider the following example data available about what kind of pizza your friends would find delicious:

Pizza ID	Toppings	Spicy	Temperature	Crust	Delicious
1	Yes	Yes	Mild	Thick	Yes
2	Yes	Yes	Hot	Thin	Yes
3	No	No	Hot	Thin	Yes
4	No	Yes	Mild	Thick	No
5	Yes	No	Hot	Thick	Yes
6	No	No	Hot	Thick	No

Using the maximum a posteriori (MAP) estimate with naïve Bayes assumption, predict the outcome for pizza 7.

Pizza ID	Toppings	Spicy	Temperature	Crust	Delicious
7	No	No	Mild	Thick	?

Show all the working. [15%]

(c) Explain how you would solve the above if instead of discrete-valued attributes, we are given some continuous valued attributes for a large number of training examples. [4%]

2. (a) Briefly explain the main similarity and difference between k-means and k-nearest neighbour algorithms. [4%]

(b) Consider the following data set with two input attributes (i.e. the x and y coordinates of the points) and one binary output t (taking values + or -). We want to use k-nearest neighbours (K-NN) with Euclidean distance to predict target t. What is the average leave-one-out cross-validation error of 3-NN on this data set? Show all the working. [8%]

Hint: put the points on the Cartesian grid as a first step.

x	y	t
-3	-2	+
-1	-2	+
-3	0	+
-1	0	+
3	1	+
-2	-1	–
-2	1	–
2	-2	–
2	0	–
3	-1	–
4	-2	–
4	0	–

3. (a) When is it possible for a support vector machine (SVM) classifier to classify non-linearly separable data points. Justify your answer. [3%]

(b) Consider the 2-dimensional data below for learning a binary classification model:

Class 1: (2,2), (3,3), (-2,-3), (-3,-2)
Class 2: (-2,2), (-3,3), (2,-3), (3,-2)

These two classes are not linearly separable. Is it possible to transform these points from 2-dimensions to a different 2-dimensions such that the transformed points are linearly separable? Justify with reasoning. [6%]

(c) Given the confusion matrix below as a measure of classifier's performance to predict class labels for a 3-class problem, answer the following: [5%]

		Actual class label		
Predicted label		Renal disease	Metabolic disease	Healthy
	Renal disease	24	5	2
	Metabolic disease	3	29	2
	Healthy	3	2	39

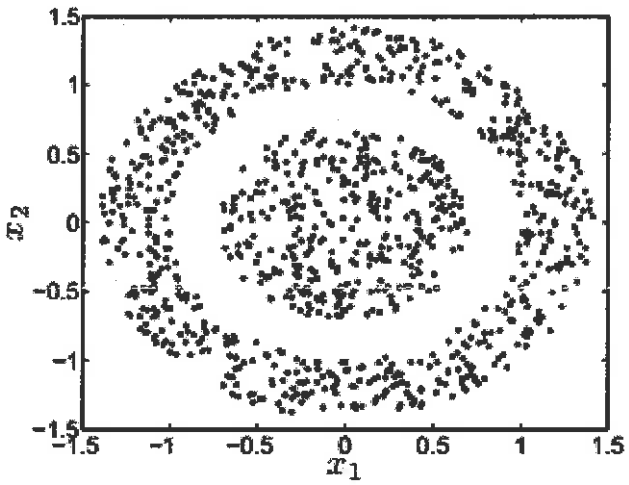
(i) Compute the overall classification accuracy.
(ii) For renal disease, compute the true positive, true negative, false positive, false negative.

4. (a) Consider the following two-dimensional data which has been divided into two clusters as shown:

Cluster 1: (2,-3), (2,1), (3,2)
Cluster 2: (-3,1), (-3,-2)

Given the above information, determine which cluster a new point (0,0) belongs to by using k-means clustering algorithm. Show all the working/reasoning to support your answer. Describe any assumptions you may consider. [5%]

(b) Given the 2-dimensional dataset below, can conventional k-means algorithm successfully determine the two apparent clusters? Justify your comments with reasoning. If yes, describe how it will be able to determine the means of both clusters. If no, suggest why not and propose any variations in k-means algorithm to cluster this dataset. [6%]



(c) Explain, with the aid of suitable diagrams, three ways in which the similarity between groups of points can be evaluated in hierarchical agglomerative clustering. [3%]

Question continues over the page

Question 4 continued

(d) The distance matrix below provides the distance between each pair of four objects. Draw a dendrogram plot showing how the objects will be clustered by hierarchical agglomerative clustering algorithm using: (i) min/single link, and (ii) max/complete link. Note: the height of each "junction" in the dendrogram represents the distance between the pair of clusters. [9%]

	A	B	C	D
A		2	5	5
B	2		3	6
C	5	3		4
D	5	6	4	

5. (a) Describe the steps to calculate the principal components of a dataset and perform dimensionality reduction with these components. [7%]

(b) A dataset contains 100 measurements, each of which is of 10 variables. The ten eigenvalues of the covariance matrix of this dataset are shown below:

1382.0, 25.5, 1.3, 0.46, 508.4, 9.3, 3.4, 187.0, 68.8, 0.17

Based upon the eigenvalues, comment about the underlying dimensionality of this dataset. [5%]

(c) Describe the regularised least squares method. What advantages does it offer over the standard least squares method? [5%]

6. Consider the means and covariance matrices of a two-class data set below.

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & -1.5 \\ -1.5 & 2 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix}$$

Make a rough drawing of the point cloud (i.e. spread of points) of each class, considering that samples are modelled with a multivariate Gaussian probability density function.

For this two-class data set, draw the shape of the decision boundary modelled with a Bayesian classifier with Gaussian class-conditional likelihood, in the case of with naïve assumption and without naïve assumption. [12%]

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.