The University Of Sheffield.

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE AUTUMN SEMESTER 2018

Machine Learning and Adaptive Intelligence **2 hours**

Answer ALL the questions.

Figures in square brackets indicate the marks allocated to each part of a question, out of 100.

This page is blank.

1. Machine Learning and Probability [Total: 25 marks]

   a) Give two examples of supervised learning problems and two examples of unsupervised learning problems. Indicate which ones are the inputs and which ones are the outputs in each problem. [5 marks]

   b) Let X and Y be two random variables with joint probability distribution $P(X, Y)$. Show that the mean of their sum satisfies

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

   [10 marks]

   c) We are building a probabilistic model for predicting whether a patient has Meningitis or not based on some descriptive features related to the symptoms that the patient exhibits. Our dataset is shown in the following table

ID	Headache (H)	Fever (F)	Vomiting (F)	Meningitis (M)
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

   Symptoms include the presence or absence of headache (column Headache in the table above), the presence or absence of fever (column Fever in the table above), and the presence or absence of vomiting (column Vomiting in the Table above). What is the probability that a patient has Meningitis if they exhibit the following features: Headache=true, Fever=false, and Vomiting= true? [10 marks]

2. Linear Regression and Basis Functions [Total: 25 marks]

   Consider a regression problem for which each observed output y_i has an associated weight factor $r_i > 0$, such that the sum of squared errors is given as

$$E(\mathbf{w}) = \sum_{i=1}^{n} r_i (y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2,$$

   where $\mathbf{w} = [w_0, \ldots, w_m]^\top$ is the vector of parameters, and $\boldsymbol{\phi}_i = [\phi_0(\mathbf{x}_i), \ldots, \phi_m(\mathbf{x}_i)]^\top$ is a vector-valued function of basis functions.

   a) Starting with the expression above, write the sum of squared errors in matrix form. You should include each of the steps necessary to get the matrix form solution. [HINT: a diagonal matrix is a matrix that only has entries different from zero in the main diagonal. The weight factors $r_i > 0$ can be written in the main diagonal of a diagonal matrix $\mathbf{R}$ of size $n \times n$.]. [15 marks]

   b) Find the optimal value of $\mathbf{w}$, $\mathbf{w}^*$, that minimises the sum of squared errors. The solution should be in matrix form. Use matrix derivatives. [10 marks]

3. Bayesian Regression and Naive Bayes [Total: 25 marks]

a) Use one or two sentences to briefly describe the purpose of each of the four components in Bayes' rule and write down the Bayes' rule in these components as a formulae: i) *prior*, ii) *likelihood*, iii) *marginal likelihood*, iv) *posterior*. [8 marks]

b) What does the term *marginalise* mean in relation to probability distributions? Write down your answer with description, and then consider the discrete and continuous settings SEPARATELY by writing down mathematical formulae to support your answer. [5 marks]

c) Explain the naive Bayes assumption that lets us simplify the expression

$$P(X_1 = v_1, \cdots, X_d = v_d | C = c)P(C = c).$$

[3 marks]

d) Assume we have a random sample that is Bernoulli distributed $X_1, \cdots, X_n \sim Bernoulli(\theta)$. We are going to derive the Maximum Likelihood Estimation (MLE) for θ. Recall that a Bernoulli random variable X takes values in $\{0, 1\}$ and has probability mass function given by

$$P(X; \theta) = \theta^X (1 - \theta)^{1-X}.$$

Derive the likelihood denoted as $L(\theta; X_1, \cdots, X_n)$ and log likelihood denoted as $\ell(\theta; X_1, \cdots, X_n)$. Show your steps. [6 marks]

e) Practical implementations of a Naive Bayes classifier often use log probabilities. Explain why. [3 marks]

4. Principal Component Analysis (PCA) and Logistic Regression [Total: 25 marks]

a) You are given a dataset $\mathbf{X}_{trn}$ with the following properties:

$$\text{Covariance matrix: } \mathbf{C} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$\text{Eigenvectors of } \mathbf{C}\text{: } \mathbf{w}_1 = \begin{bmatrix} -0.55 \\ 0.83 \end{bmatrix}, \ \mathbf{w}_2 = \begin{bmatrix} 0.83 \\ 0.55 \end{bmatrix}$$

$$\text{Eigenvalues of } \mathbf{C}\text{: } \lambda_1 = 5, \lambda_2 = 51.$$

(i) Explain the criterion function that PCA optimises. [2 marks]

(ii) Which direction vector above gives the first principal component for this dataset? Why? [2 marks]

(iii) How much of the variance in the data is explained by the first principal component? [3 marks]

(iv) What geometrical relation is there between the first and the second principal component? [1 marks]

(v) Given a test data sample $\mathbf{x}_{tst}$, you are required to get its PCA representation $\mathbf{y}_{tst}$. Write down, as specifically as possible, a mathematical formulae that will allow you to obtain $\mathbf{y}_{tst}$. [5 marks]

(vi) For the training dataset $\mathbf{X}_{trn}$, the PCA-transformed dataset is $\mathbf{Y}_{trn}$. What is the covariance matrix for the PCA-transformed data? (*Hint: no computation is needed to get the answer.*) [4 marks]

b) This question is about *binary* logistic regression with two output classes.

An experiment is conducted on the toxicity of doses of an insecticide on the tobacco budworm moth. In the experiment batches of 20 male moths were exposed for 3 days to the insecticide and the number in each batch that were dead or knocked down was recorded. The data are given below.

Dose level	1	2	4	8	16	32
$\log_2$(Dose level)	0	1	2	3	4	5
Dead or down	1	4	9	13	18	20

(i) What is the definition of the *odds* in binary logistic regression. [2 marks]

(ii) What are the maximal and minimal possible values of the odds? [2 marks]

(iii) From the table above, what are the observed odds of dead or down at dose level 16? [2 marks]

(iv) How would you build a classification model that can handle three output classes using binary logistic regression? [2 marks]

END OF QUESTION PAPER