Module Code: COMP5611M01

Module Title: Machine Learning

© **UNIVERSITY OF LEEDS**

School of Computing

Semester 2 2022/2023

Examination Information

- There are **9** pages to this examination.
- There are **2 hours** to complete the examination.
- Answer **all 6** questions.
- The number in brackets **[ ]** indicates the marks available for each question or part question.
- You are reminded of the need for clear presentation in your answers.
- The total number of marks for this examination paper is **60**.
- You are allowed to use annotated materials.

Question 1

(a) Consider the table of probabilities given below which lists the joint probability distribution of a rose being red or white, and of having thorns or not. Give the probability of the rose having thorns conditioned on being red rounded to two decimals

P(c=r,th=y) = 0.25	P(c=r,th=n) = 0.23
P(c=w,th=y) = 0.46	P(c=w,th=n) = 0.06

[3 marks]

(b) Consider the vectors given below to represent the probability of each of the six possible outcomes from rolling two dice.
Dice 1: $p_1 = [0.2, 0.2, 0.2, 0.1, 0.1, 0.2]$
Dice 2: $p_2 = [0.3, 0.05, 0.05, 0.4, 0.1, 0.1]$
What is the joint probability of rolling a 5 with Dice 1 and then rolling a 4 with Dice 2?
[2 marks]

(c) A doctor is called to see a sick child. The doctor has prior information that 80% of sick children in that neighborhood have the flu, while the other 20% are sick with COVID. Let F stand for an event of a child being sick with flu and C stand for an event of a child being sick with COVID. Assume for simplicity that that there no other infections in that neighborhood. A well-known symptom of COVID is a dry cough (having a dry cough is denoted D). Assume that the probability of having a dry cough if one has COVID is $P(D \mid C) = 0.95$. Occasionally children with flu also develop a dry cough, and the probability of having a dry cough if one has flu is $P(D \mid F) = 0.10$. Upon examination the doctor finds the child has a dry cough. What is the probability that the child has COVID? Show your working and round to decimal places.
[3 marks]

[Question 1 Total: 8 marks]

Question 2

(a) Which of the following statements is **not** true? Justify your selection with a few lines of text.

(i) If the variables of a bivariate Gaussian distribution are uncorrelated, then they are independent.

(ii) Any bivariate Gaussian distribution can be written as the product of two univariate Gaussian distributions.

(iii) We can use a linear transformation to decorrelate the two variables of a bivariate Gaussian distribution.

(iv) Eigenvalues of the covariance matrix are always positive.

[3 marks]

(b) A coin is modelled with a Bernoulli process. As a prior, a Beta distribution with parameters $a = 6, b = 4$ is chosen, given by, Beta$(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, where, a and b correspond to outcomes 1 and 0, respectively. The coin is tossed 10 times. The following sequence is observed: '1010100101'. The posterior distribution is also a Beta distribution. Infer the parameters of the posterior distribution, show your workings.

[5 marks]

(c) Consider the scatter plot below comprising points sampled from a bivariate Gaussian distribution with mean $\mu = [10.0, -10.0]$ and covariance matrix $\Sigma = \begin{bmatrix} 79.0 & 20.0 \\ 20.0 & 37.0 \end{bmatrix}$.

Select from the statements below, the correct way to transform these set of points to those generated from the product of two centered, uncorrelated one-dimensional Gaussian distributions of equal variance.

(i) Set the off-diagonal entries in Σ to zero and resample the points to generate the desired transformed points

(ii) Translate the points to the global origin [0,0] by subtracting the mean and set the off-diagonal entries in Σ to zero before resampling the desired transformed points

(iii) Translate the points to the global origin [0,0] by subtracting the mean. Calculate the eigenvectors and eigenvalues of Σ using eigendecomposition and rotate the mean-centred samples using the estimated eigenvectors.

(iv) Translate the points to the global origin [0,0] by subtracting the mean. Calculate the eigenvectors and eigenvalues of Σ using eigendecomposition and rotate the mean-centred samples using the estimated eigenvectors. Finally scale the points by dividing by the square root of the eigenvalues.

[3 marks]

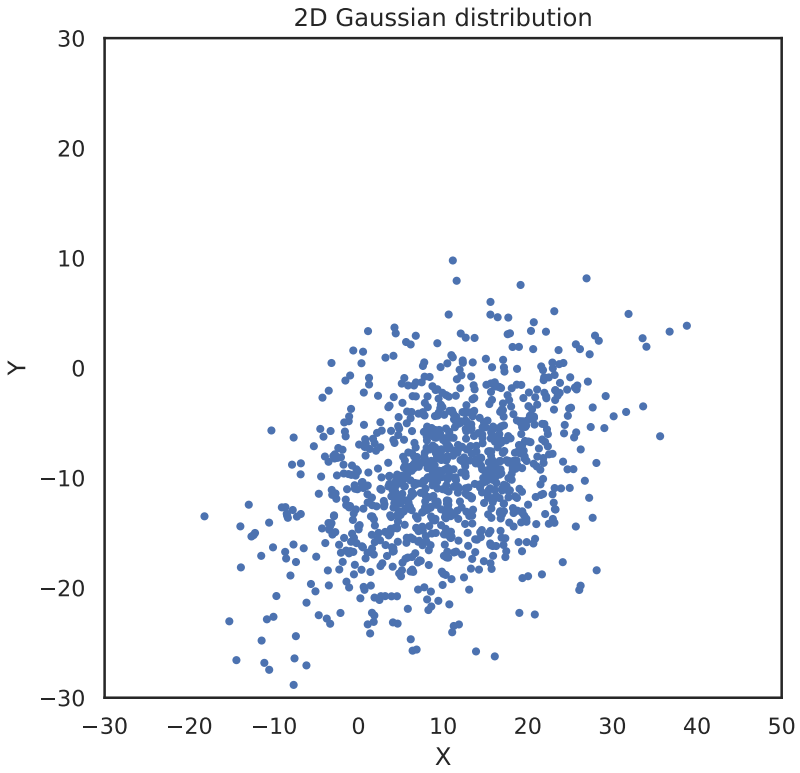

Figure 1: Sample points from a bivariate Gaussian distribution.

(d) Consider a Bernoulli process with $\mu = 0.3$. Simulate a sequence of 100 outcomes. Interpret the simulated sequence as follows: place a walker at position 0. A '0' is interpreted as a move of size -1, a '1' is interpreted as a move of size 1. Our simulated sequence can thus be interpreted as a random walk. Record the end position of our random walk. Repeat this 10000 times. The appropriately binned histogram of these end positions should approximately resemble a Gaussian according to the central limit theorem. What is the mean μ and variance σ of this distribution.

(i) $\mu = 60, \sigma = 1.1$

(ii) $\mu = 60, \sigma = 21$

(iii) $\mu = 30, \sigma = 4.6$

(iv) $\mu = 30, \sigma = 21.1$

[5 marks]

[Question 2 Total: 16 marks]

Question 3

(a) Consider the following four samples: $x_1 = 6, x_2 = 4, x_3 = 2, x_4 = -5$. Here samples $x_{n=1..4}$ represent the features/predictors to be used to fit a linear regression model using the ordinary least squares approach. The chosen regression function for this task is a degree 3 polynomial, i.e. a cubic polynomial. Provide the design matrix D for this problem and express D as an $N \times M$ matrix, where N=4 represents the number of samples and M=3 denotes the degree of the chosen polynomial regression function. Show your workings. (Note: for simplicity you may represent a row vector as: $x = [x_1, x_2, x_3]$ (i.e. using comma separated values), a column vector as $x = [x_1; x_2; x_3]$ (i.e. using semi-colon separated values), and a $N \times M$ matrix as: $X = [x_{11}, x_{12}, ..., x_{1M}; x_{21}, x_{22}, ..., x_{2M}; ...; x_{N1}, x_{N2}, ..., x_{NM}]$.

[5 marks]

(b) You fit a Bayesian linear regression model to your data, assuming your unexplained noise is Gaussian distributed with a known, fixed variance and unknown mean. You assume a prior distribution over your unknown regression weights such that it is conjugate to the likelihood function, given by a zero-centered Gaussian distribution with variance α. This allows you to derive the posterior distribution for your regression weights given the data and derive the maximum-a-posteriori estimate for the same. Answer the following questions about the resulting regression model:

What is the difference between Bayesian and classical linear regression (here classical linear regression model is a frequentist model that assumes the residuals are Gaussian distributed)? Explain in a few sentences.

[2 marks]

(c) Which of the following statements is **true** about fitting a Bayesian linear regression model and a classical linear regression model (as described above) to a data set with few training samples?

(i) The number of training data points is irrelevant, both models provide the same result.

(ii) With few training data points Bayesian linear regression model overfits to the data more than a classical linear regression model.

(iii) With few training data points there is high uncertainty in the Bayesian linear regression model and no uncertainty in the weights estimated by the linear regression model.

(iv) Both the Bayesian linear regression model and the classical linear regression model underfit to the training data.

[3 marks]

[Question 3 Total: 10 marks]

Question 4

(a) What is the difference between the CART and ID3 algorithms? Select the most suitable option.

(i) ID3 doesn't capture curvature in decision boundaries but CART does

(ii) ID3 uses Information Gain based on entropy while CART uses Gini Gain based on Gini impurity for splitting attributes.

(iii) ID3 can be used for classification and regression tasks, while CART can only be used for classification.

(iv) There is no difference.

[3 marks]

(b) Consider a logistic regression binary classification model that has been trained to predict whether a patient has cancer (Positive class) or not (Negative class) with a decision threshold of 0.5. Your employer has set design requirements for your model, stating that missing diagnosis of cancer is a far worse error to incur than to misdiagnose cancer. How would you go about finalising your model and delivering the final predictions? Choose one of the following statements and justify your answer in a few sentences.

(i) Do nothing, the performance of the model is what it is.

(ii) Decrease the decision threshold using the ROC curve to determine a suitable threshold

(iii) Increase the decision threshold using the ROC curve to determine a suitable threshold

(iv) Relabel your data by swapping cancer and not cancer to negative and positive classes respectively, and retrain your model

[3 marks]

[Question 4 Total: 6 marks]

Patient Name	COVID symptoms	Close contact with COVID+	Class Label
Hannah	No	No	*Don't Isolate*
Joe	No	Yes	*Don't Isolate*
Mark	Yes	Yes	*Isolate*
Cynthia	Yes	No	*Isolate*
Tom	No	No	*Don't Isolate*
Lisa	No	Yes	*Don't Isolate*

Table 1: COVID attributes data set.

Question 5

Consider the dataset in Table 1 with categorical features/attributes: Att1 - COVID symptoms, Att2 - Close contact with COVID+. Find the optimal order for splitting attributes in a binary decision tree for classifying these data into two classes: Isolate vs Don't Isolate.

(a) Do this for the ID3 algorithm. Support your answer by calculating the information gain for the chosen order of splits. Show your workings.

[5 marks]

(b) Do the same for the CART algorithm based on the GINI gain. Show your working.

[5 marks]

[Question 5 Total: 10 marks]

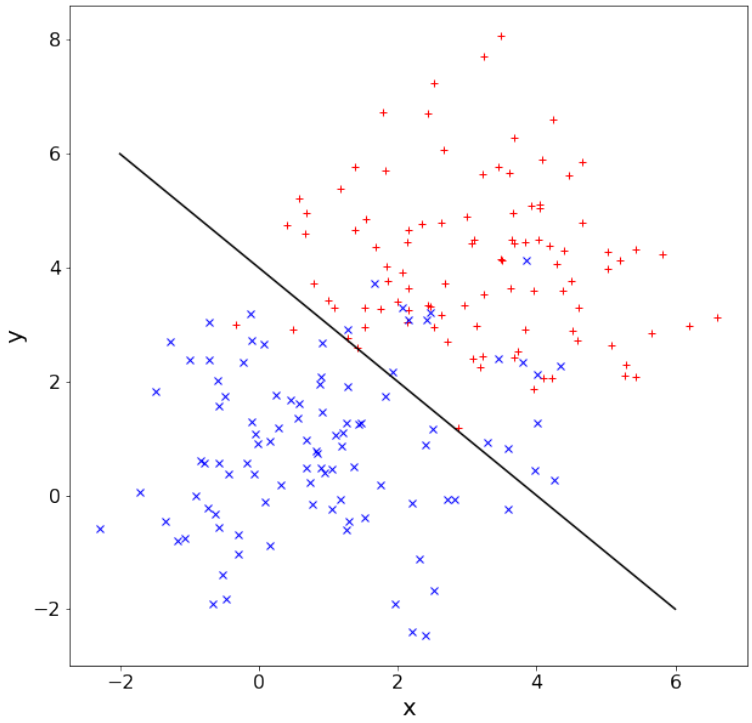

Figure 2: A dataset consisting of points that have been atrributed to two different classes: one indicated by '+' and one by 'x'.

Question 6

In Figure 2 you see a dataset with points being labeled into two classes. One class is indicated by '+', the other by 'x'. You want to be able to classify new data points using logistic regression. Someone has drawn a line by eye in the dataset with the aim of separating the classes. The line passes the points (6,-2) and (-2,6). When we mention the term 'weights' in this question, this includes all parameters, also an intercept when one is needed.

(a) Give the mathematical definition of a logistic regression classifier for this two-dimensional dataset, and explain how to interpret its output and how you can use that to decide which class a novel point belongs to. Provide an explicit formula in terms of the x and y coordinates of a point.

[4 marks]

(b) Give a logistic classifier that is based on the line drawn in the figure. Any set of weights is acceptable, as long as the line represents the set of points where a point has exactly 50 % probability of falling into the '+' class.

[6 marks]

[Question 6 Total: 10 marks]

[Grand Total: 60 marks]