CS5103: Applied Machine Learning – Answer Outlines for the Mid-Sem Exam

Exam Held on 13th September 2025 at the Indian Institute of Science Education and Research Kolkata

Duration: 1.5 Hours Full Marks: 30

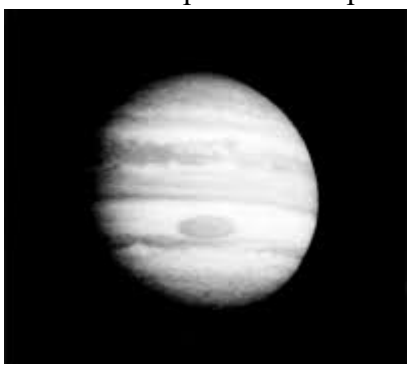
Q1. (a) Which attribute of Python packages differentiate them from non-package modules?
Ans. The '__path__' attribute.

(b) When we run the command 'import numpy as np', where does the Python interpreter look for the 'numpy' module?
Ans. In the directories specified by the PYTHONPATH environment variable.

(c) When investigating a trained decision tree classifier, we found that the tree has only a single node. What does it imply?
Ans. It implies that the tree could not separate out the training samples belonging to different classes.

(d) If the scatter plot of your data has the appearance of a sphere, what can you say about the linear relationships between the data dimensions? At the same time, what can you say about their non-linear relationships?
Ans. They do not have any linear relationship. However, we can not say anything about their non-linear relationship from the plot.

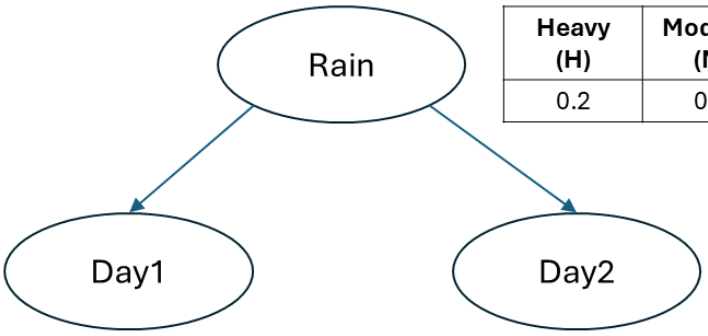
(e) The following 1974 photograph of Jupiter was taken by NASA's Pioneer 10 spacecraft, the first spacecraft to fly past Jupiter. Can we use a QDA classifier to separate out Jupiter from its surroundings?



Ans. Yes, we can use a QDA classifier because the contour of the Jupiter looks like a circle which can be represented by a quadratic equation.

[1+1+1+2+1= 6 marks]

Q2. (a) Suppose, we want to forecast whether there will be heavy rain, moderate rain, or no rain on a particular day. For the forecasting, we will utilise the information about rain (heavy, moderate or none) for the previous two days, referred to as 'Day1' and 'Day2'. Assuming we choose to use the naïve Bayes classifier, please draw the naïve Bayes model along with the model parameters (i.e. the conditional probability tables or CPTs). Initialise the CPTs with random values.
Ans.



Rain

Heavy (H)	Moderat (M)	None (N)
0.2	0.4	0.4

Day1

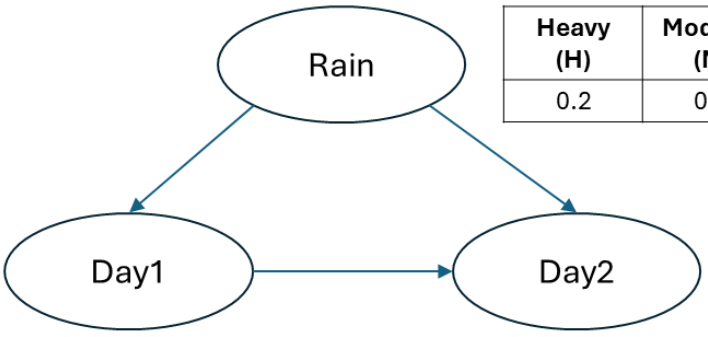
Rain	H	M	N
H	0.4	0.3	0.3
M	0.3	0.4	0.3
N	0.3	0.3	0.4

Day2

Rain	H	M	N
H	0.5	0.3	0.2
M	0.3	0.5	0.2
N	0.3	0.2	0.5

(b) Suppose, we relax the naïve assumption and assume that rain on Day2 also depends on rain on Day1. In that case, please redraw the naïve Bayes model along with its parameters. Again, please randomly initialise the model parameters. [3+2 = 5 marks]

Ans.



Rain

Heavy (H)	Moderat (M)	None (N)
0.2	0.4	0.4

Day1

Rain	H	M	N
H	0.4	0.3	0.3
M	0.3	0.4	0.3
N	0.3	0.3	0.4

Day2

Rain, Day1	H	M	N
HH	0.5	0.3	0.2
HM	0.3	0.5	0.2
HN	0.3	0.2	0.5
MH	0.1	0.6	0.3
MM	0.2	0.4	0.4
MN	0.3	0.4	0.3
NH	0.2	0.5	0.3
NM	0.3	0.3	0.4
NN	0.1	0.7	0.2

Q3. (a) The following figure is a contrast-enhanced CT scan of a human brain taken from Uggetti et al., 2009, American Journal of Neuroradiology. The concerned person is suffering from a rare disorder known as AGS. We can see white spots scattered across the brain. They are abnormal deposits of calcium known as 'calcifications'. We need to build a classifier that can distinguish the calcifications from the rest of the brain. Note that the CT scan is a 2D image comprised of thousands of pixels. Each calcification is made up of one or more pixels. For preparing the training data, we will select the coordinates of some of the pixels that lie on calcifications (the 'calcification' class) and the coordinates of some of the pixels that do not lie on any calcification (the 'no calcification' class). We could observe that the calcifications are linearly non-separable in the original 2D space. Please construct the equation of a support vector classifier that increases the number of dimensions such a way that a 3D hyperplane can potentially separate out the calcifications from the rest of the brain.

Ans. Separation with a 3D hyperplane requires that the training samples are in a (3+1)=4D space. Therefore, we need to increase the number of dimensions of each training pixel from 2 to 4. Let us denote the original dimensions of the i-th training pixel as (x_i1, x_i2) which is simply its coordinates in the 2D image, assuming that the left bottom corner of the image is the origin (0, 0). There are many ways to derive two new dimensions from the two existing dimensions. One way is simply taking the squares of the existing dimensions, e.g., (x_i1, x_i2, $(x\_i1)^2$, $(x\_i2)^2$). In that case, the equation of the resultant support vector classifier would be same as Equation (9.16) of ISLP where we will replace 'p' with 2. In the equation, 'y_i' i.e. the class label of the i-th training pixel will be +1 if it lies on a calcification spot or -1 if it does not lie on a calcification spot.

(b) Increasing the number of dimensions will also increase the computational complexity of the constructed classifier. Can you suggest a 'trick' with which the computational complexity could be reduced?

[3+3 = 6 marks]

Ans. We can utilise the 'kernel trick'. Since the calcifications are small circular or elliptical spots consisting of one or more pixels, the radial basis function (RBF) kernel might be the most suitable for separating them from their surroundings. Equations (9.23) and (9.24) of ISLP can be used to explain how the separation will be performed using the RBF kernel.

Q4. Suppose, we wish to build a multilayer perceptron (MLP) to solve the same 'calcification classification' (a tongue twister!) problem. The MLP should have one hidden layer. The number of neurons in the hidden layer should be the same as the number of dimensions of the increased-dimension space that you have created in the previous question. Please explain the forward propagation and backpropagation with diagrams and equations under this setting. [5+5 = 10 marks]

Ans. Please find the diagrams and equations in the Jupyter notebook of the 'Intro to TensorFlow' lecture. The only difference will be in the dimensions. The input should be a 2-vector instead of a 784-vector. The hidden layer should have 4 neurons instead of 128 neurons. The output layer should consist of 2 neurons instead of 10 neurons.

Q5. (a) In the CT scan image, we could see that some calcification pixels are less bright than other calcification pixels. Those pixels might be harder to classify than the latter. Among random forest and AdaBoost, which classifier would you choose in this case and why?

Ans. I would choose AdaBoost so that the constituent weak classifiers gradually adapt to the harder-to-classify pixels.

(b) There is no free lunch i.e. we have to sacrifice something to gain something. What would you have to sacrifice for your chosen classifier? [2+1 = 3 marks]

Ans. I would have to sacrifice computational cost since AdaBoost is more computationally intensive than random forest.