UNIVERSITY OF BIRMINGHAM

School of Computer Science

Artificial Intelligence 2

Main Summer Examinations 2021

Artificial Intelligence 2

Exam paper

Question 1 (Probabilistic AI/ML)

As a data scientist in a telecommunication company, your task is to analyse a customer dataset to predict whether a customer will terminate his/her contract. The dataset consists of around 8000 customer records, each consisting of one binary dependent variable Y, indicating whether the customer terminates the contract ($Y = 1$) or not ($Y = 0$), and 19 independent variables, which include the customer's information, e.g., age, subscription plan, extra data plan, etc., and the consumer behaviour such as average numbers of calls and hours per week. Since your boss needs some actionable insights to retain customers, you decided to use interpretable machine learning methods. Design your interpretable machine learning method by answering the following questions:

(a) You have implemented a feature selection algorithm based on mutual information to select the most informative features from the 19 independent variables. To validate the implementation of your mutual information calculation function, you use a small subset of the data to calculate mutual information manually. You select one independent variable 'subscription plan', denoted as S, which takes two values, $S \in \{1, 2\}$. Please use the following Probability Mass Function table

$p(S, Y)$	$S = 1$	$S = 2$
$Y = 0$	$\frac{2}{12}$	$\frac{5}{12}$
$Y = 1$	$\frac{2}{12}$	$\frac{3}{12}$

to calculate

- Entropies $H(S)$ and $H(Y)$
- Conditional entropies $H(S|Y)$ and $H(Y|S)$
- Joint entropy $H(S, Y)$
- Mutual information $I(S; Y)$

Show all your working. Discuss what mutual information means and whether this feature will be selected or not. **[6 marks]**

(b) After applying your algorithm you selected two variables: 1) extra data plan E, which is a binary random variable that indicates whether the customer subscribes to the extra data plan ($E = 1$) or not ($E = 0$); and 2) averaged hours used per week H, which is a continuous random variable. You then built a logistic regression model

to classify customers into 'low risk' or 'high risk' of terminating the contract. The fitted model is

$$\log\left(\frac{p}{1-p}\right) = -0.77 + 0.23H - 1.18E$$

- Given a customer who has the extra data plan ($E = 1$) and spent on average 0.5 hours per week, calculate the odds and the probability the customer will terminate the contract ($Y = 1$). **[4 marks]**
- Using this fitted model, explain to your boss what actions should be taken to retain customers. **[10 marks]**

Question 2 (Deep Learning / Artificial Neural Networks)

(a) Consider that we have a 64x64x5 dimensions medical image (i.e. 64x64 pixels and 5 channels) which we are processing through a convolutional layer of a deep convolutional neural network. Answer the following for this image:

(i) Consider that the convolutional layer has 32 receptive fields each of size 7x7x5, without zero padding and a stride size of 1. What will be the output dimensions of the image produced from this convolutional layer? Justify your answer briefly with reasons through arguments and/or diagrams. **[6 marks]**

(ii) Consider that you pass the output obtained from the convolutional layer in (i) above to a channel-wise maximum pooling layer with a pool size 2x2 and a stride size of 2. Note that channel wise max pooling layer operates on each channel separately. What will be the output dimensions of the image produced from this max pooling layer? Justify your answer briefly with reasons through arguments and/or diagrams.

[4 marks]

(b) Consider that a cancer hospital intends to develop an AI solution for the automatic prediction of cancer from the magnetic resonance image (MRI) of a patient's brain. Formulate this problem as a deep machine learning problem and suggest how its solution can be developed. Your answer should include: problem formulation (e.g. what are the data & labelling needs), identification of relevant deep neural network that can be used for solution development (i.e. just the name of relevant network), the training/learning process of such network (i.e. how weights/parameters will be determined/updated) and the performance evaluation of the solution. Note: you are not required to explain a learning algorithm and neither are you required to provide mathematical formulae for the algorithm or performance evaluation.

[10 marks]

Answer for Question 1 :

a)

1.

$p(S,Y)$	$S=1$	$S=2$	$p(Y)$
$Y=0$	$\frac{2}{12}$	$\frac{5}{12}$	$\frac{7}{12}$
$Y=1$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{5}{12}$
$p(S)$	$\frac{4}{12}$	$\frac{8}{12}$	1

a) $H(S) = -\sum_i^n p(s_i)\log_2 p(s_i) = -(\frac{4}{12}\log\frac{4}{12} + \frac{8}{12}\log\frac{8}{12}) = 0.92$ bits

$H(Y) = -\sum_i^n p(y_i)\log_2 p(y_i) = -(\frac{7}{12}\log\frac{7}{12} + \frac{5}{12}\log\frac{5}{12}) = 0.97$ bits

c) $H(S,Y) = -\sum_{x_i\in R_X}\sum_{y_j\in R_Y} p(s_i,y_j)\log p(s_i,y_j) = 1.89$ bits

b) $H(S|Y) = H(S,Y) - H(Y) = 1.89 - 0.97 = 0.92$ bits

$H(S|X) = H(S,Y) - H(S) = 1.89 - 0.92 = 0.97$ bits

d) $I(S;Y) = H(S) - H(S|Y) = 0.92 - 0.92 = 0$ bits

b)ii)

Second question: We can investigate the effect or each variable by deriving the odd ratios by fixing the value of the other variable:

$$OR_E = \frac{\text{odds when } E=1}{\text{odds when } E=0} = \frac{\exp(-0.77+0.23-1.18)}{\exp(-0.77+0.23)} = \exp(-1.18) \approx 0.31$$

$$OR_H = \frac{\text{odds when } H=h+\Delta}{\text{odds when } H=h} = \frac{\exp(-0.77+0.23(H+\Delta)-1.18)}{\exp(-0.77+0.23H-1.18))} = (\exp(0.23))^{\Delta} \approx 1.25^{\Delta}$$

Another simpler way is to use log odd rations. You can use the fitted model

$$\log\left(\frac{p}{1-p}\right) = -0.77 + 0.23H - 1.18E$$

The extra data coefficient is 0.23 means that after holding the other factors fixed for changing from no extra data to have extra data, the log odds of the customer's risk of terminating the contract goes down by -1.18. For the average hour spent (H) coefficient is 0.23 means that all else equal, for one more hour a customer spent, the log odds of the customer's risk of terminating the contract goes up by 0.23.

From the analysis, we can suggest to the boss that, the more hours the customers spent, the more likely the customers will terminate, which means the company should improve its telecommunication service/price. However, by simply persuade them to subscribe to the extra data plan, they are more likely to stay.

b)i)

2. **First question**:

- Odds:

$$odds = \frac{p}{1-p} = \exp(-0.77 + 0.23H - 1.18E) = \exp(-0.77 + 0.115 - 1.18)$$
$$= 0.1596 \quad (1)$$

- Probability:

$$P(Y=1) = 0.137 \quad (2)$$

Answer for Question 2 :

(a) (i) The output image size will be 58x58x32. Since there is no zero padding being used, the first image pixel where the receptive field will be placed is at pixel location (4,4), thus discarding the first 3 rows and the first 3 columns in the process. Similarly, the last image pixel where the receptive field will be placed is at pixel location (61,61), thus discarding the last 3 rows and last 3 columns. As a result, we only get 58x58x32 sized output image. **[6 marks]**

(ii) The output image size will be 29x29x32. The max pooling kernel will be placed at every 2x2 block in the image and it will move/jump by every 2 pixels since the stride size is 2. From each 2x2 block on each channel, we'll get 1 output. Thus, as a result, we will only get 29x29x32 sized output image. **[4 marks]**

(b) This problem can be formulated as a binary classification problem. The problem formulation would include a collection of past example data which includes brain MRI images and the associated labels at image level. These labels would show which images contain tumour and which images don't contain tumour. We would typically need thousands of such labelled images for each class. **[4 marks]**

From this example data, the network will then learn to map the image to the tumour/non-tumour label. A convolutional neural network can be used as the learning algorithm. Such network can be trained from scratch if we've sufficient data, or a pre-trained network (e.g. ResNet network on ImageNet data) can be fine tuned through transfer learning. Both would require weights/parameters update through backpropagation. **[4 marks]**

The performance evaluation will be conducted on the basis of comparing algorithm's predictions with the actual (i.e. the ground truth) values and monitoring a count of correct/incorrect classifications which can be presented in % terms. **[2 marks]**