Module Code: COMP561101

Module Title: Machine Learning

School of Computing

Semester 2 Practice Version

Calculator instructions:

- You **are** allowed to use a non-programmable calculator only from the following list of approved models in this examination: **Casio FX-82**, **Casio FX-83**, **Casio FX-85**.

Dictionary instructions:

- You are **not** allowed to use your own dictionary in this examination. A basic English dictionary is available to use: raise your hand and ask an invigilator, if you need it.

Examination Information

- There are **9** pages to this examination.
- There are **2 hours** to complete the examination.
- Answer **all 5** 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 **66**.
- You are allowed to use annotated materials.

with solutions

Question 1

HMRC has been seizing illegal wine imports and collated statistics in Table 1, where wines are split out by type and nationality.

(a) Consider the statistics in Table 1 on illegal wine imports. According to the table what are the marginal probabilities of seizing a French wine for white, rose and red? Give results precise to 2 decimals.

Solution:

The three probabilities are given by:

$$\frac{0.05}{0.05 + 0.22 + 0.17} = \frac{0.05}{0.44} = 0.11$$

and

$$\frac{0.22}{0.44} = 0.5$$

and

$$\frac{0.17}{0.44} = 0.39$$

[6 marks]

(b) What is probability of seizing a red wine, conditioned on it being Italian?

Solution:

The probability of a wine being Italian is 0.56. Therefore the probability of wine being red, conditioned on it being Italian is

$$\frac{0.23}{0.56} = 0.41$$

:

[4 marks]

[Question 1 Total: 10 marks]

	white	rose	red
French	0.05	0.22	0.17
Italian	0.18	0.15	0.23

Table 1: HMRC has compiled statistics on fraudulent wine imports. The table represents joint probabilities on seized contraband.

Question 2

A test for melanoma detects 95 % of all true melanoma cases. When applied to people who don't have melanoma, 4 % of the test results is positive. The prevalence of melanoma in the general population is 0.1 %.

(a) Use Bayes' law to estimate what the probability is that a person has melanoma upon receiving a positive test result.

Solution:

Bayes' law states:

$$p(+ \mid D) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid \not D)P(\not D)}$$

Here D signifies the subject actually having melanoma and $/D$ the subject not having melanoma. $P(+ \mid D) = 0.5$ and $P(+ \mid \not D) = 0.04$. $P(\not D) = 1 - P(D)$.

Substitution gives:

$$\frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.04 \cdot 0.999} \approx 0.23$$

[6 marks]

[Question 2 Total: 6 marks]

Question 3

The cross entropy is defined as: $\mathbb{E}_p[\log q]$. Assume a state space with three states and probabilities $p_1 = 0.05, p_2 = 0.45, p_3 = 0.5$.

(a) Calculate the cross-entropy between p and q for the two following distributions of q: $q_1 = 0.01, q_2 = 0.48, q_3 = 0.51$ and $q_1 = 0.001, q_2 = 0.48, q_3 = 0.519$.

Solution:

Using the definition gives: 0.8972169147173776 and 1.0035995906433213

[4 marks]

(b) Comment on whether you expect to see a similar difference between the two q distributions if you would have used the mean squared error (it should not be necessary to perform the actual calculation)? Explain why a measure that is sensitive to the difference between the two q distributions can be important for some problems.

Solution:

The mean squared error would deliver a small difference between both cases. But the frequency of event 1 happening is a factor of 10 higher for the second q-distribution, which is much harder to detect from the MSE than from the cross-entropy.

[4 marks]

[Question 3 Total: 8 marks]

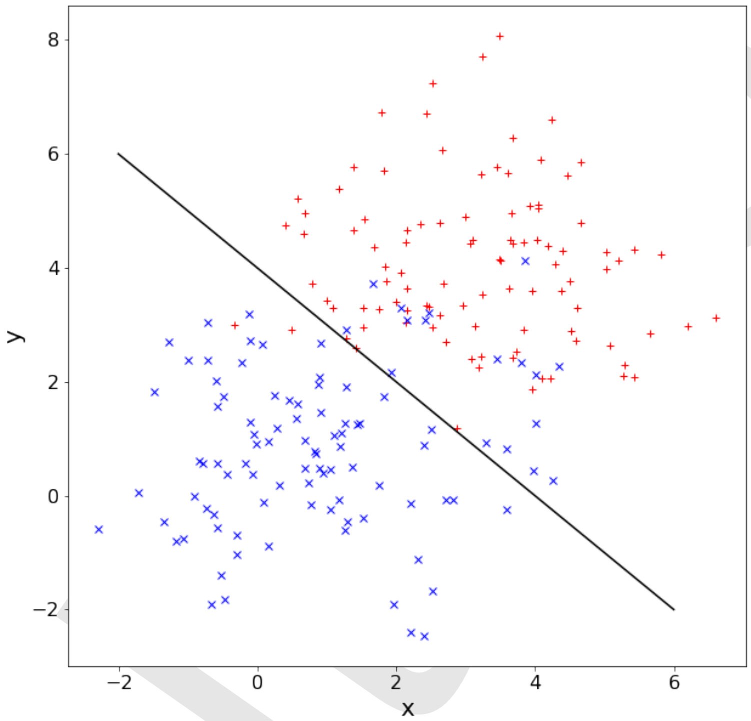

Figure 1: A dataset consisting of points that have been atrributed to two different classes: one indicated by '+' and one by 'x'.

Question 4

In Figure 1 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.

Solution:

A logistic classsifier in this case is defined by:

$$o = \frac{1}{1 + e^{-(w_0 + w_1 x + w_2 y)}}$$

The use of the minus sign is unimportant. Its output is a number between 0 and 1, which is typically interpreted as the probability of a data point to one of the two classes. A decision is usually made based on whether the probability is smaller or larger than a half.

[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.

Solution:

The line passing through the points (-2, 6), (6,-2) is $x + y = 4$, so $x + y - 4 = 0$ is the line of 50 % probability. This means e.g. $w_0 = -4, w_1 = 1, w_2 = 1$.

[6 marks]

(c) Present a geometrical argument for the statement that a loss function for logistic regression has a global minimum. You are allowed to base your argument on Figure 1 and to generalise it. Explain why the existence of a global minimum simplifies finding a minimum for the cost function.

Solution:

It is clear that the optimal line for separation must be close to the one drawn because any other line would give many more misclassifications. It is clear that depending on the loss function, one line will be the best, so there is a global maximum. In a higher dimensional space the argument remains the same. [4 marks]. This helps with techniques such as steepest gradient descent because there are no local minima where the technique can get stuck.

[6 marks]

(d) On a much larger dataset, someone has estimated means and covariance for both classes. The results are as follows:

$$\mu_{'+'} = (3,4)^T, \mu_{'\mathsf{x}'} = (1,1)^T$$

and

$$\Sigma_{'+'} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = \Sigma_{'\mathsf{x}'}.$$

Based on this information, give the equation for the 50 % probability line. Explain whether the line drawn in the figure is accurate or not.

Solution:

Students have been given a formula for calculating this. However, they may conclude on the basis of a geometrical argument that the line equidistant between the points (1,1) and (3,4) gives the right answer. This line is perpendicular to the direction (2,3) and must pass through the halfway point between them. So $2x + 3y = c$, for c such that it passes $(2, 2.5)$, which gives $c = 11.5$. In standard form this is $y = -2/3x + 3.83$. [7] marks for the calculation. [3] marks for concluding that the line is not quite correct because its gradient is off.

[10 marks]

[Question 4 Total: 26 marks]

Sample No.	Regularity of tumour boundary	Tumour Diameter	Classification
Sample 1	Regular	Large	Benign
Sample 2	Irregular	Small	Malignant
Sample 3	Irregular	Small	Benign
Sample 4	Regular	Small	Benign
Sample 5	Irregular	Large	Malignant
Sample 6	Irregular	Large	Malignant
Sample 7	Regular	Small	Benign
Sample 8	Regular	Large	Benign
Sample 9	Regular	Small	Benign
Sample 10	Irregular	Small	Malignant

Table 2: A small dataset of tumour attributes and their classification as health risk.

Question 5

It is well known that tumours with regular boundaries are often benign and those with irregular boundaries are often malignant. However, it is also fairly common to find that malignant tumours tend to be bigger and have larger diameters than benign tumours. Given the data set (of N = 10 samples) described in Table 2, one can find the optimal order for splitting attributes/features (i.e. regularity of boundary and tumour diameter) in a binary decision tree for classifying these data into two classes: Malignant tumour vs Benign tumour.

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

Solution:

ID3:
The entropy of a set S is given by $H(S) = -\sum_i p_i \log_2(p_i)$

$P(B) = 0.6, P(M) = 0.4$, denote set of samples (S).

Entropy of full set:

$$H(S) = -0.6 * \log_2(0.6) – 0.4 * \log_2(0.4) = 0.9710$$

$$H(S \mid \mathit{Attribute}1) = 0.5 * (0) + 0.5 * (-1/5 * \log_2(1/5) - 4/5 * \log_2(4/5)) = 0.3609$$

Information Gain: $IG(\mathit{Attribute}1) = H(S) – H(S \mid \mathit{Attribute}1) = 0.9710 - 0.3609 = 0.6101$

$$H(S \mid \mathit{Attribute}2) = 0.6 * (-4/6 * \log_2(4/6) – 2/6 * \log_2(2/6)) +$$

$$0.4 * (-2/4 * \log_2(2/4) – 2/4 * \log_2(2/4)) = 0.9510$$

Information Gain: $IG(\mathit{Attribute}2) = H(S) – H(S \mid \mathit{Attribute}2) = 0.9710 – 0.9510 = 0.0200$.

Therefore, split first on Attribute 1 (regularity of tumour boundary) as information gain is higher.

[8 marks]

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

Solution:

CART:
The Gini impurity of a set S is given by $G(S) = \sum_i p_i(1 - p_i)$
Gini impurity of full set:

$$G(S) = 0.6 * (1 - 0.6) + 0.4 * (1 - 0.4) = 0.48$$

$$G(S \mid Attribute1) = 0.5 * (0) + 0.5 * (1/5 * (1 - 1/5) + 4/5 * (1 - 4/5)) = 0.16$$

Gini Gain: $GG(Attribute1) = G(S)–G(S \mid Attribute1) = 0.48$

$$H(S \mid Attribute2) = 0.6 * (4/6 * (1 - 4/6) + 2/6 * (1 - 2/6))+$$

$$0.4 * (2/4 * (1 - 2/4) + 2/4 * (1 - 2/4)) = 0.4667$$

Gini Gain: $GG(Attribute2) = G(S)–G(S \mid Attribute2) = 0.48–0.4667 = 0.0133$

Therefore, split first on Attribute 1 (regularity of tumour boundary) as resulting Gini gain is higher.

[8 marks]

[Question 5 Total: 16 marks]

[Grand Total: 66 marks]