Xi'an Jiaotong-Liverpool University

西交利物浦大學

PAPER CODE	EXAMINER	DEPARTMENT	TEL
INT305	Jimin XIAO	Intelligent Science	3209

Final Exam 2021/2022

BACHELOR DEGREE – Year 4

Machine Learning

TIME ALLOWED: 3 Hours 0 Minutes

INSTRUCTIONS TO CANDIDATES

1、 **Total marks available are 100.**

2、 **Answer all questions.**

3、 **The number in the column on the right indicates the approximate marks for each section.**

4、 **In general, it is particularly important to give reasons for your answer. Only partial marks will be awarded for correct answers with inadequate reasons.**

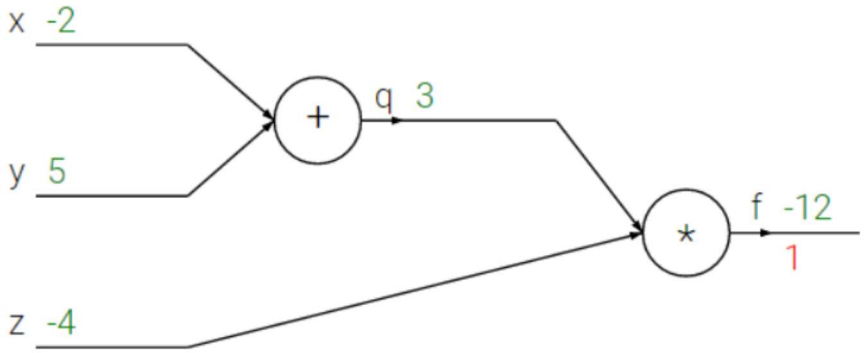
5、 **Answer should be written in the answer booklet(s) provided.**

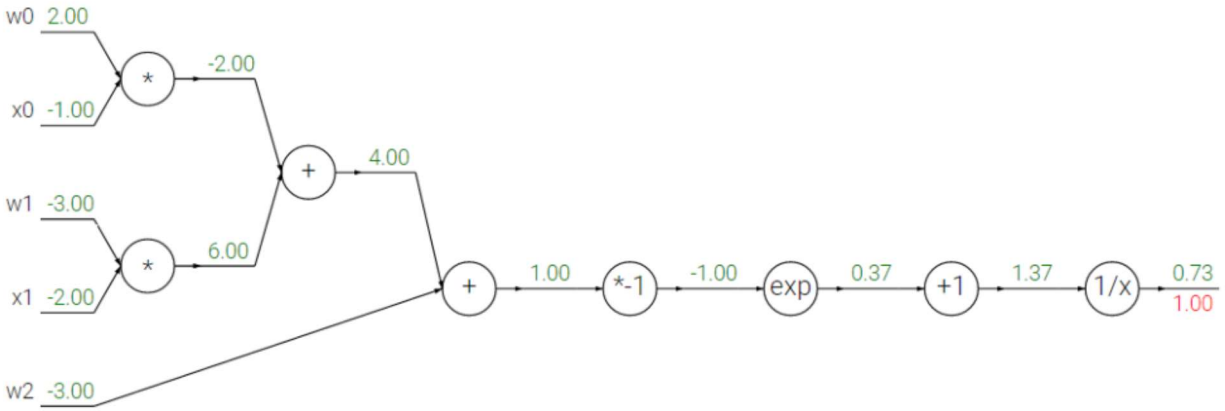
6、 **The university approved calculator - Casio FS82ES/83ES can be used.**

7、 **Only solutions written in English will be accepted.**

Answer All Questions

1. Fill in the missing gradients underneath the forward pass activations in each circuit diagram. The gradient of the output with respect to the loss is one (1.00), and has already been filled in. (The output values for each unit are represented on top of the line in green and the gradients are in red.)





1.5 marks each blank

Total 24

2. Pooling units take n values x_i, $i \in [1, n]$ and compute a scalar output whose value is invariant to permutations of the inputs.

(1) The Lp-pooling module takes positive inputs and computes $y = (\sum_i x_i^p)^{\frac{1}{p}}$, assuming we know that $y' = \frac{\partial L}{\partial y}$, what is $x_i' = \frac{\partial L}{\partial x_i}$? **(8 marks)**

(2) The log-average module computes $y = \frac{1}{\beta} ln(\frac{1}{n}\sum_i exp(\beta x_i))$, assuming we know that $y' = \frac{\partial L}{\partial y}$, what is $x_i' = \frac{\partial L}{\partial x_i}$? **(8 marks)**

Total 16

3. Suppose binary-valued random variables X and Y have the following joint distribution:

	Y = 0	Y = 1
X = 0	2/8	4/8
X = 1	1/8	1/8

Determine the information gain IG(Y|X). You may write your answer as a sum of logarithms.

Total 15

4. We showed that the Support Vector Machine (SVM) can be viewed as minimizing hinge loss:

$$\min_{w,b} \sum_{i=1}^{N} \mathcal{L}_H(y,t) + \frac{1}{2\gamma}\|\mathbf{w}\|_2^2$$

where hinge loss is defined as:

$$\mathcal{L}_H(y,t) = max(0, 1-ty)$$

Here $t \in \{-1, 1\}$ is the label, $y = \mathbf{w}^T\mathbf{x} + b$ is the predicted score.

(a) TRUE or FALSE: if the total hinge loss is zero, then every training example must be classified correctly. Justify your answer. **(5 marks)**

(b) Suppose we replace the hinge loss with the following:

$$\mathcal{L}(y,t) = max(0, -ty)$$

and otherwise keep the soft-margin SVM objective the same. What would go wrong? **(5 marks)**

Total 10

5. Recall two linear classification methods we considered:

Model 1:

$$y = \mathbf{w}^T\mathbf{x} + b$$

$$\mathcal{L}_{SE}(y,t) = \frac{1}{2}(y - t)^2$$

Model 2:

$$z = \mathbf{w}^T\mathbf{x} + b$$

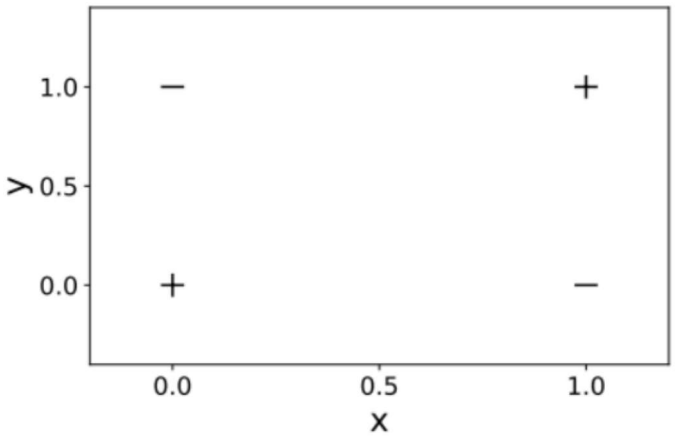
$$y = \sigma(z)$$

$$\mathcal{L}_{SE}(y,t) = \frac{1}{2}(y - t)^2$$

Here, σ denotes the logistic function (sigmoid function), and the target label t take values in $\{0, 1\}$. Briefly explain our reason for preferring Model 2 to Model 1.

Total 10

6. The drawing below shows a dataset. Each example in the dataset has two inputs features x and y, and maybe classified as a positive example (labelled +) or a negative example (labelled −). Draw a decision tree which correctly classifies each example in the dataset.



Total 10

7. The Laplace distribution, parameterized by μ and β, is defined as follows:

$$Laplace(x;\mu,\beta) = \frac{1}{2\beta} exp(-\frac{|x - \mu|}{\beta})$$

Consider a variant of Bayesian linear regression where we assume the prior over the weights $\mathbf{w}$ consists of an independent zero-centered Laplace distribution for each dimension (w_j), with shared parameter β:

$$w_j \sim Laplace(0,\beta)$$
$$t \mid \mathbf{w} \sim \mathcal{N}(t;\ \mathbf{w}^T\psi(\mathbf{x}),\sigma)$$

For reference, the Gaussian PDF is:

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

(1) Suppose you have a labelled training set $\{(x^{(i)}, t^{(i)})\}_{i=1}^{N}$. Please give the cost function you would minimize to find the MAP estimate of w. (It should be expressed in terms of mathematical operations.)

(10 marks)

(2) Based on your answer to part (1), how might the MAP solution for a Laplace prior differ from the MAP solution if you use a Gaussian prior? **(5 marks)**

Total 15

The end of the paper