Information

Preliminaries

Restrictions/Allowances

- You are allowed a calculator
- This test is closed book
- You may **not** code
- Access to any other Moodle course or page is forbidden and will be logged

Rounding

- All numerical answers must be rounded to **3 decimal places**.
- Unless otherwise stated, when your calculations involve multiple steps, round any intermediate results to 5 decimal places or use fractions before using them in the next step.

Question 1

Correct

Mark 1.00 out of 1.00

Which function maps a real-valued number to a probability between 0 and 1 in logistic regression?

○ a. ReLU

○ b. Softmax

● c. Sigmoid ✓

○ d. Linear

The correct answer is: Sigmoid

Question 2

Incorrect

Mark 0.00 out of 1.00

Regularisation adds a penalty term proportional to the:

● a. Sum of weights. ✗

○ b. Sum of squared weights.

○ c. Sum of absolute weights.

○ d. Number of non-zero weights.

The correct answer is: Sum of squared weights.

Question 3

Correct

Mark 1.00 out of 1.00

A key difference between discriminative and generative models for classification is that:

○ a. Discriminative models learn P(x|y), while generative models learn P(y|x).

● b. Generative models directly model the data distribution P(x), discriminative models do not model P(x) directly. ✓

○ c. Discriminative models are always linear, while generative models are always Gaussian.

○ d. Generative models cannot handle continuous features, while discriminative models can handle continuous features.

The correct answer is: Generative models directly model the data distribution P(x), discriminative models do not model P(x) directly.

Question 4

Correct

Mark 1.00 out of 1.00

What does the term $\phi(x)$ represent in the context of $h_\theta(x) = \sigma(\theta^T\phi(x))$?

○ a. The model parameters (weights).

○ b. The activation function.

○ c. The output probability.

● d. A vector of input features. ✓

The correct answer is: A vector of input features.

Question 5

Correct

Mark 1.00 out of 1.00

In the context of gradient descent, the learning rate α controls:

○ a. The direction of the weight update.

● b. The magnitude (step size) of the weight update. ✓

○ c. The number of iterations required for convergence.

○ d. The complexity of the model.

The correct answer is: The magnitude (step size) of the weight update.

Question 6

Correct

Mark 1.00 out of 1.00

The Perceptron Learning Algorithm is guaranteed to converge if:

● a. The data is linearly separable. ✓

○ b. A sigmoid activation is used.

○ c. The learning rate is sufficiently small.

○ d. Regularisation is applied.

The correct answer is: The data is linearly separable.

Question 7

Correct

Mark 1.00 out of 1.00

The decision boundary $\theta^T x = 0$ in logistic regression corresponds to where the predicted probability $h_\theta(x)$ equals 0.5.

● True ✓

○ False

The correct answer is 'True'.

Question 8

Correct

Mark 1.00 out of 1.00

Using polynomial basis functions in logistic regression allows it to model non-linear decision boundaries.

● True ✓

○ False

The correct answer is 'True'.

Question 9

Correct

Mark 2.00 out of 2.00

Match the concept (A-B) with its description (i-ii).

A. Cross-Entropy Cost	ii. Measures the performance of a classification model whose output is a probability value between 0 and 1.	✓
B. Softmax Function	i. Generalises logistic regression to multi-class classification, producing probabilities that sum to 1.	✓

The correct answer is: A. Cross-Entropy Cost → ii. Measures the performance of a classification model whose output is a probability value between 0 and 1., B. Softmax Function → i. Generalises logistic regression to multi-class classification, producing probabilities that sum to 1.

Information

Relevant Formulae

- $\sigma(z) = \frac{1}{1+e^{-z}}$
- $h_\theta(\mathbf{x}) = \sigma(\theta^T \phi(\mathbf{x}))$
- $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$
    - where m is the number of data points and log is the natural logarithm (ln).
- $\frac{\partial J}{\partial \theta_i} = (h_\theta(x) - y)x_i$
- $\theta_i \leftarrow \theta_i - \alpha\frac{\partial J}{\partial \theta_i}$

Question 10

Correct

Mark 4.00 out of 4.00

Consider a logistic regression model for a single training example $(\mathbf{x}, y)$. The weights are $\theta = [\theta_0, \theta_1, \theta_2]^T = [-1, 2, -1]^T$.

The input feature vector is $\mathbf{x} = [x_1, x_2]^T = [1.5, 0.5]^T$. The true label is $y = 1$.

Calculate the predicted probability $h_\theta(\mathbf{x})$.

Answer: 0.818 ✓

The correct answer is: 0.818

Question 11

Correct

Mark 3.00 out of 3.00

For a single training example a logistic regression model predicts $h_\theta(\mathbf{x}) = 0.7$, while the true label is $y = 1$. Compute the non-regularised cross-entropy cost of this single example.

Answer: 0.357 ✓

The correct answer is: 0.357

Question 12

Correct

Mark 3.00 out of 3.00

Consider a logistic regression model for a single training example $(\mathbf{x}, y)$. The weights are $\theta = [\theta_0, \theta_1, \theta_2]^T = [0, 1, -2]^T$. The input feature vector (excluding bias) is $\mathbf{x} = [x_1, x_2]^T = [1.5, 0.5]^T$. The true label is $y = 1$ and the predicted value was $h_\theta(\mathbf{x}) = 0.622$.

Compute the new values of θ after one gradient step with $\alpha = 1$.

θ_0 = 0.378 ✓

θ_1 = 1.567 ✓

θ_2 = -1.811 ✓

Question 13

Correct

Mark 1.00 out of 1.00

Which of the following is commonly used as an activation function in hidden layers of neural networks designed for general function approximation or classification?

○ a. Softmax

● b. ReLU ✓

○ c. Cross-Entropy

○ d. Gradient Descent

The correct answer is: ReLU

Question 14

Correct

Mark 1.00 out of 1.00

What does $a_i^{(l)}$ represent in neural network notation?

Select one:

○ a. The weight connecting neuron i to layer l.

● b. The activation of neuron i in layer l ✓

○ c. The pre-activation value of neuron i in layer l

○ d. The bias term added to layer l.

The correct answer is: The activation of neuron i in layer l

Question 15

Correct

Mark 1.00 out of 1.00

Which activation function outputs the input directly if it's positive, and zero otherwise?

○ a. Sigmoid

○ b. Softmax

● c. ReLU ✓

○ d. Linear

The correct answer is: ReLU

Question 16

Correct

Mark 1.00 out of 1.00

A feed-forward network means that connections generally flow:

○ a. From later layers back to earlier layers.

○ b. Within the same layer only.

● c. From earlier layers forward to later layers. ✓

○ d. Randomly between any two neurons.

The correct answer is: From earlier layers forward to later layers.

Question 17

Correct

Mark 1.00 out of 1.00

A neural network has layer sizes (excluding bias units) $s_1 = 10$ (input), $s_2 = 20$, $s_3 = 5$ (output). What is the dimension of the weight matrix $\Theta^{(2)}$ connecting layer 2 to layer 3?

Select one:

○ a. 5×20

● b. 5×21 ✓

○ c. 21×5

○ d. 20×5

The correct answer is: 5×21

Question 18

Incorrect

Mark 0.00 out of 1.00

The Universal Approximation Theorem implies that neural networks are powerful but does NOT guarantee:

○ a. The ability to represent complex functions.

○ b. Efficient learning of the function from data.

○ c. The need for non-linear activation functions.

● d. That a single hidden layer might suffice theoretically. ✗

The correct answer is: Efficient learning of the function from data.

Question 19

Correct

Mark 1.00 out of 1.00

In a classification problem the number of neurons in the output layer of an MLP is typically determined by the number of classes.

● True ✓

○ False

The correct answer is 'True'.

Question 20

Correct

Mark 1.00 out of 1.00

Bias units ($a_0^{(l)}$) typically have fixed activation values (e.g., 1) and do not apply the layer's main activation function $g(z)$.

● True ✓

○ False

The correct answer is 'True'.

Question 21

Correct

Mark 2.00 out of 2.00

Match the term with its correct description

Activation $a_i^{(l)}$	The final output value of neuron i in layer l	✓
Pre-activation $z_i^{(l)}$	The weighted sum of inputs (plus bias) to neuron i in layer l	✓

The correct answer is: Activation $a_i^{(l)}$
→ The final output value of neuron i in layer l, Pre-activation $z_i^{(l)}$
→ The weighted sum of inputs (plus bias) to neuron i in layer l

Information

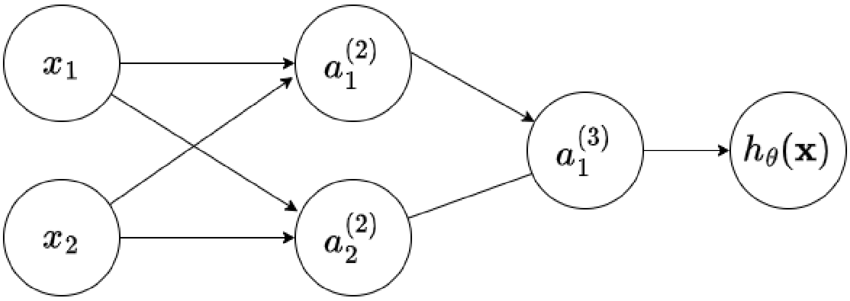
Relevant Formulae

- **ReLU** function: $g(z) = max(0, z)$
- $\mathbf{z}^{(\mathbf{l})} = \Theta^{(l-1)}\mathbf{a}^{(l-1)}$ for all $l \in [2, L]$

Information

Question 22

Consider a neural network with 2 input neurons (x_1, x_2), 1 hidden layer with 2 neurons $(a_1^{(2)}, a_2^{(2)})$, and 1 output neuron $(a_1^{(3)})$. The neural network is visualised below. Bias units are implicit and are not visualised.



The activation function for all neurons is the **ReLU** function: $g(z) = max(0, z)$.

The weight matrices are:

$\Theta^{(1)} = \begin{bmatrix} -1 & 2 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \Theta^{(2)} = \begin{bmatrix} -0.5 & 1 & -1 \end{bmatrix}$

Question **22.1**

Correct

Mark 4.00 out of 4.00

Given $\mathbf{x} = [x_1, x_2]^T = [2, 1]^T$, calculate the pre-activation vector for the hidden layer $\mathbf{z}^{(2)} = [z_1^{(2)}, z_2^{(2)}]^T$.

$z_1^{(2)}$ = 2 ✓

$z_2^{(2)}$ = -2 ✓

Question **22.2**

Correct

Mark 2.00 out of 2.00

Given some *other* $\mathbf{x}$ that gives $\mathbf{z}^{(2)} = [z_1^{(2)}, z_2^{(2)}]^T = [-10, 10]$, calculate the activation vector for the hidden layer $\mathbf{a}^{(2)} = [a_1^{(2)}, a_2^{(2)}]^T$.

$a_1^{(2)}$ = 0 ✓

$a_2^{(2)}$ = 10 ✓

Question **23.3**

Correct

Mark 3.00 out of 3.00

Given some other $\mathbf{x}$ assume that $\mathbf{a}^{(2)} = [a_1^{(2)}, a_2^{(2)}] = [3, 4]$. What is the value of z_1^3 and $h_\theta(\mathbf{x})$?

z_1^3 = -1.5 ✓

$h_\theta(\mathbf{x})$ = 0 ✓

Question 24

Correct

Mark 1.00 out of 1.00

Backpropagation is fundamentally an application of which mathematical rule?

○ a. Bayes' Theorem

● b. The Chain Rule ✓

○ c. Linear Algebra Matrix Inversion

○ d. The Central Limit Theorem

The correct answer is: The Chain Rule

Question 25

Correct

Mark 1.00 out of 1.00

What does the cost function $J(\Theta)$ quantify in neural network training?

○ a. The number of layers in the network.

○ b. The speed of convergence.

● c. The difference between the network's predictions and the true labels. ✓

○ d. The computational complexity of the forward pass.

The correct answer is: The difference between the network's predictions and the true labels.

Question 26

Correct

Mark 1.00 out of 1.00

If weights are initialised to zero, what problem occurs during the first backpropagation step?

○ a. Division by zero in the activation function.

● b. All neurons in a layer will compute the same gradient. ✓

○ c. The forward pass cannot be computed.

○ d. The cost function becomes infinite.

The correct answer is: All neurons in a layer will compute the same gradient.

Question 27

Correct

Mark 1.00 out of 1.00

Feature scaling or normalisation aims to put input features:

○ a. Into a $\{0, 1\}$ range only.
● b. Onto similar scales or ranges. ✓
○ c. Into an orthogonal basis.
○ d. Into a higher dimensional space.

The correct answer is: Onto similar scales or ranges.

Question 28

Correct

Mark 1.00 out of 1.00

Momentum helps gradient descent by:

○ a. Adding noise to escape local minima.
○ b. Decreasing the learning rate automatically.
● c. Penalising large changes in direction. ✓
○ d. Temporarily removing neurons from the network according to probability p.

The correct answer is: Penalising large changes in direction.

Question 29

Correct

Mark 1.00 out of 1.00

Dropout is primarily used as a technique to:

○ a. Speed up computation.
● b. Reduce overfitting. ✓
○ c. Handle missing data.
○ d. Automatically determine the number of hidden units.

The correct answer is: Reduce overfitting.

Question 30

Correct

Mark 1.00 out of 1.00

Backpropagation calculates the error $\delta^{(l)}$ starting from the output layer and moving backward towards the input layer.

● True ✓

○ False

The correct answer is 'True'.

Question 31

Correct

Mark 1.00 out of 1.00

Using gradient descent on a typical neural network cost function is guaranteed to find the set of weights that gives the lowest possible error.

○ True

● False ✓

The correct answer is 'False'.

Question 32

Correct

Mark 2.00 out of 2.00

Match the backpropagation term with its role.

Error term $\delta_i^{(l)}$ | Represents how much a neuron i in layer l contributed to the errors in the subsequent layer(s) ✓

Activation derivative $g'(z_j^{(l)})$ | Represents how much a change in the neuron’s pre-activation impacts its activation, used in error backpropagation. ✓

The correct answer is: Error term $\delta_i^{(l)}$
→ Represents how much a neuron i in layer l contributed to the errors in the subsequent layer(s), Activation derivative $g'(z_j^{(l)})$
→ Represents how much a change in the neuron’s pre-activation impacts its activation, used in error backpropagation.

Information

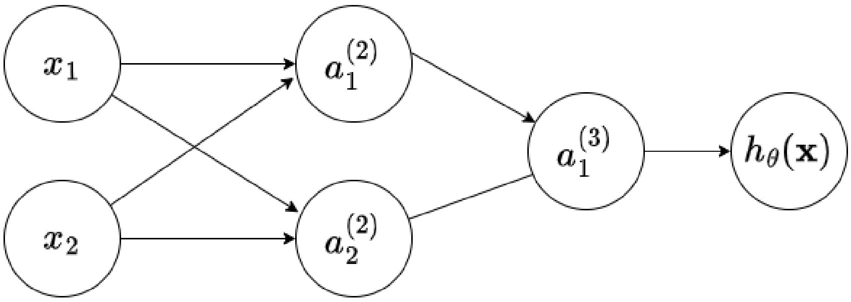
Relevant Formulae

- $\odot$ represents Hadamard product (or otherwise known as element wise multiplication)
- $\sigma'(\mathbf{z}^{(l)}) = \mathbf{a}^{(l)} \odot (1 - \mathbf{a}^{(l)})$
- $\delta^{(L)} = \mathbf{a}^{(L)} - \mathbf{y}$
- $\delta_j^{(l)} = \left(\sum_m \delta_m^{(l+1)} \Theta_{mj}^{(l)}\right) g'(z_j^{(l)})$ for all $l \in [2, L-1]$
- $\delta^{(l)} = ((\tilde{\Theta}^{(l)})^T \delta^{(l+1)}) \odot g'(z^{(l)})$ for all $l \in [2, L-1]$
  - $\tilde{\Theta}^{(l)}$ refers to the weight matrix for layer l excluding bias connections.
- $\frac{\partial J}{\partial \Theta_{ij}^{(l)}} = a_j^{(l)} \delta_i^{(l+1)}$
- $\Theta^{(l)} \leftarrow \Theta^{(l)} - \alpha D^{(l)}$

Information

Question 30

Consider a neural network with 2 input neurons (x_1, x_2), 1 hidden layer with 2 neurons $(a_1^{(2)}, a_2^{(2)})$, and 1 output neuron $(a_1^{(3)})$. The neural network is visualised below. This is the same network as before except using **Sigmoid** as the activation function $g(z) = \sigma(z) = \frac{1}{1+e^{-z}}$. Bias units are implicit and are not visualised.



Each of the preceding questions in this section use the same architecture but differing values for $\mathbf{x}$ and y and these values will only be given when relevant. Precomputed pre-activations, activations, and errors will be given where relevant and are specific to each question. Use this information and the information provided with each of the preceding questions in your answers.

TODO: Remove following

The activation function for all neurons is the **Sigmoid** function: $\sigma(z) = \frac{1}{1+e^{-z}}$. The weight matrices are:

$$\Theta^{(1)} = \begin{bmatrix} -1 & 2 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \Theta^{(2)} = \begin{bmatrix} -0.5 & 1 & -1 \end{bmatrix}$$

Given the data $\mathbf{x} = [x_1, x_2]^T = [3, 4]^T$, forward propagation produced the following values (rounded to 2 decimal places):

$$z^{(2)} = \begin{bmatrix} 1 & -6 \end{bmatrix}, \quad a^{(2)} = \begin{bmatrix} 0.73 \; 0.00 \end{bmatrix}, \quad z^{(3)} = [0.23], \quad a^{(3)} = [0.56]$$

Use this information to answer the subsequent questions.

Question 33.1

Correct

Mark 2.00 out of 2.00

Assume the true label for this example is $y = 1$ and the activation for the output layer was computed as $a^{(3)} = [0.56]$. Calculate the error term for the output layer $\delta^{(3)}$. Provide the value for $\delta_1^{(3)}$.

Answer: -0.44 ✓

The correct answer is: -0.44

Question 33.2

Correct

Mark 2.00 out of 2.00

Assume that the activations for the hidden layer were calculated as $a^{(2)} = \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix}$. Calculate the vector of activation derivatives for the hidden layer $g'(z^{(2)})$.

$g'(z_1^{(2)})$ = 0.25 ✓

$g'(z_2^{(2)})$ = 0.188 ✓

Question 33.3

Correct

Mark 4.00 out of 4.00

Assume the following values:

$\Theta^{(2)} = [\,-0.5 \quad 1 \quad -1\,]$

$\delta^{(3)} = [\,0.25\,]$

$g'(\mathbf{z}^{(2)}) = \begin{bmatrix} 0.5 \\ 0.3 \end{bmatrix}$

Calculate the error vector for the hidden layer $\delta^{(2)} = \begin{bmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \end{bmatrix}$.

$\delta_1^{(2)} =$ 0.125 ✓

$\delta_2^{(2)} =$ -0.075 ✓

Question 37

Correct

Mark 3.00 out of 3.00

Assume that the original value of the weight matrix for the hidden layer is $\Theta^{(2)} = \begin{bmatrix} -0.5 & 1 & -1 \end{bmatrix}$.

Assume $\delta_1^{(3)} = \begin{bmatrix} 2 \end{bmatrix}$ and $\mathbf{a}^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 0.5 \end{bmatrix}$

Use $\alpha = 0.25$.

Perform one gradient update to $\Theta^{(2)} = \begin{bmatrix} \theta_0^{(2)} & \theta_1^{(2)} & \theta_2^{(2)} \end{bmatrix}$ and provide the answers below.

$\theta_0^{(2)}$ = -1 ✓

$\theta_1^{(2)}$ = 0.5 ✓

$\theta_2^{(2)}$ = -1.25 ✓