exercisesweek47Answer

November 28, 2025

1 Exercise week 47-48

November 17-28, 2025

Date: **Deadline is Friday November 28 at midnight**

2 Overarching aims of the exercises this week

The exercise set this week is meant as a summary of many of the central elements in various machine learning algorithms we have discussed throught the semester. You don't need to answer all questions.

2.1 Linear and logistic regression methods

2.1.1 Question 1:

Which of the following is not an assumption of ordinary least squares linear regression?

NOT an assumption -> There is a linearity between predictors/features and target/outout

IS an assumption -> The inputs/features distributed according to a normal/gaussian distribution

2.1.2 Question 2:

The mean squared error cost function for linear regression is convex in the parameters, guaranteeing a unique global minimum. True or False? Motivate your answer.

True, as we have an analytical solution for the minimum of the OLS method.

2.1.3 Question 3:

Which statement about logistic regression is false?

Logistic regression is used for binary classification.

It uses the sigmoid function to map linear scores to probabilities.

False -> It has an analytical closed-form solution.

Its log-loss (cross-entropy) is convex.

2.1.4 Question 4:

Logistic regression produces a linear decision boundary in the input space. True or False? Explain.

True, as it separates based on a probability if it is true ot not. i.e. we have a linear output which we then create a decision threshold on which is a line.

2.1.5 Question 5:

Give two reasons why logistic regression is preferred over linear regression for binary classification.

It is preferred over linear regression as it has a steep curve near the boundary and gives two classes 0 or 1 based on the boundary and the squish of everything within these bounds. In a more condensed form: - Robustness to outliers - Clear output and interpretation

2.2 Neural networks

2.2.1 Question 6:

Which statement is not true for fully-connected neural networks?

- Without nonlinear activation functions they reduce to a single linear model.
- Training relies on backpropagation using the chain rule.
- A single hidden layer can approximate any continuous function on a compact set.
- False -> The loss surface of a deep neural network is convex.

2.2.2 Question 7:

Using sigmoid activations in many layers of a deep neural network can cause vanishing gradients. True or False? Explain.

True, as with numbers not close to the middle of the sigmoid function has a gradient that is close to 0 creating a sequence of small numbers multiplied together.

2.2.3 Question 8:

Describe the vanishing gradient problem: Why does it occur? Mention one technique to mitigate it and explain briefly.

It occurs because of the multiple multiplication of small numbers because of the chain rule. One can mitigate it by having an optimization rule that increases the movement downwards for each update to the parameters. Such as ADAM or RMSPROP.

Or one could use a different activation function such as LeakyReLU, which has a constant gradient at some size.

2.2.4 Question 9:

Consider a fully-connected network with layer sizes n_0 (the input layer) ,n_1 (first hidden layer), $..., n_L$, where n_L is the output layer. Derive a general formula for the total number of trainable parameters (weights + biases).

$$(n_0 * n_1 + n_1) + (n_1 * n_2 + n_2) + \cdots = (\Sigma n_i n_{i+1} + \Sigma_{i=1} n_i)$$

2.3 Convolutional Neural Networks

2.3.1 Question 10:

Which of the following is not a typical property or advantage of CNNs?

- Local receptive fields
- Weight sharing
- This is not a property -> More parameters than fully-connected layers
- Pooling layers offering some translation invariance

2.3.2 Question 11:

Using zero-padding in convolutional layers can preserve the input spatial dimensions when using a 3×3 kernel/filter, stride 1, and padding $P = 1$. True or False?

True, as we we will have the same output as input.

2.3.3 Question 12:

Given input width W, kernel size K, stride S, and padding P, derive the formula for the output width $W_{\text{out}} = \frac{W-K+2P}{S} + 1$.

2.3.4 Question 13:

A convolutional layer has: C_{in} input channels, C_{out} output channels (filters) and kernel size $K_h \times K_w$. Compute the number of trainable parameters including biases.

One filter is equal to $K_h * K_w * C_{in}$, so the amount for all filters will be $K_h * K_w * C_{in} * C_{out}$

Then we have the bias which is equal to C_{out}

Total: $(K_h * K_w * C_{in} + 1) * C_{out}$

2.4 Recurrent Neural Networks

2.4.1 Question 14:

Which statement about simple RNNs is false?

- They maintain a hidden state updated each time step.
- They use the same weight matrices at every time step.
- They handle sequences of arbitrary length.
- False -> They eliminate the vanishing gradient problem.

2.4.2 Question 15:

LSTMs mitigate the vanishing gradient problem by using gating mechanisms (input, forget, output gates). True or False? Explain.

True, as it fails via the chainrule multiplication it is allevated or fixed via having and additive memory of the cost function, with a forgetting parameter and input paramter.

2.4.3 Question 16:

What is Backpropagation Through Time (BPTT) and why is it required for training RNNs?

As an RNN loops compared to the standard FFNN, one needs to make some changes for the backpropogation to work. BPTT fixes this by seeing the NN as a long chain of copies of the cell for each time, here the weights are identical. Calculate the Gradient for each step, sum them up and the change the Weights for this cell using this gradient.

2.4.4 Question 17:

What does a sliding window do? And why would we use it?