AI Research Labs

Deep Learning Foundations · Practice test 1 · 20 questions

Foundations of Deep Learning

How to use this test. Answer all 20 questions without notes, then mark yourself with the answer key at the back. Each question has exactly one best answer. Score 75% or more and you've mastered the material; below that, reread the lessons for the questions you missed.

1.1 Machine Learning and Deep Learning

Q1 concept In the model $\hat{y} = mx + c$, which statement correctly separates the model from learning?

A. The line is the model; finding m and c from data is learning
B. Learning is choosing the straight-line form; the model is the fitted values
C. The model and the learning algorithm are the same thing in machine learning
D. The data points are the model, and drawing the line through them is the learning

Q2 calculate Learning produces $\hat{y} = 5x + 30$. What does the model predict for $x = 6$?

A. 60
B. 35
C. 180
D. 41

Q3 concept Which statement about machine learning (ML) and deep learning (DL) is correct?

A. DL is the part of ML that uses multi-layered neural networks
B. DL and ML are unrelated fields that happen to share some vocabulary
C. Every ML model is a neural network, and DL is just the largest ones
D. ML is the part of DL that uses models without any neural networks in them at all

Q4 apply Why does the learning loop include an evaluation step before going round again?

A. Because the model has to be re-chosen from scratch after every round of training
B. Because evaluation is where the loss gradient for the next round gets computed
C. Because the parameters can only be updated once the evaluation step has finished
D. Because fitting training data doesn't guarantee good predictions on new data

Q5 interpret Which milestone made training multi-layer networks practical, by giving a systematic way to compute their gradients?

A. Backpropagation, popularised in 1986
B. The McCulloch–Pitts neuron model of 1943
C. The ImageNet competition result of 2012
D. The ADALINE echo canceller of 1959

1.2 Types of Learning

Q6 apply A model predicts a student's final mark (0–100) from hours of study. What kind of task is this?

- A Supervised classification
- B Unsupervised clustering
- C Reinforcement learning
- D Supervised regression

Q7 apply You have 10,000 customer records with no labels and want to find natural groups of similar customers. Which type of learning fits?

- A Unsupervised learning
- B Reinforcement learning
- C Supervised learning
- D Semi-supervised regression with pseudo-labels

Q8 concept In the notation $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, what does the superscript (i) mean?

- A It picks out training example number i
- B It is the i-th feature of a single input vector
- C It is the layer number in the network
- D It raises x to the power i

Q9 concept What distinguishes reinforcement learning from supervised learning?

- A It always uses much deeper neural networks than supervised learning does
- B It learns from rewards for actions, not from labelled input–output pairs
- C It is simply supervised learning that is applied to time-series data only
- D It needs every training example to carry a correct numerical target label

Q10 interpret A photo app asks you to name one face, then names the same person across hundreds of other photos. Which type of learning does this illustrate?

- A Pure unsupervised learning
- B Reinforcement learning
- C Semi-supervised learning
- D Fully supervised learning

1.3 Gradient Descent and Momentum

Q11 calculate Minimise $f(w) = (w - 3)^2$ by gradient descent from $w_0 = 5$ with $\eta = 0.1$. What is w_1?

- A 4.6
- B 4.8
- C 5.4
- D 3.0

Q12 interpret Your training loss shoots up and becomes NaN within a few iterations. What is the most likely cause?

A The learning rate is too small, so training has not had enough time
B The loss function is convex, so gradient descent cannot be used
C The learning rate is too large, so the updates overshoot and diverge
D The model has reached a local minimum where the gradient is zero

Q13 concept At some parameters the gradient is exactly zero. What can you conclude?

A The point is a local minimum, though perhaps not the global one
B The point is the global minimum of the objective function
C The learning rate is too small to make any further progress
D The point is stationary: a minimum, maximum, saddle point or plateau

Q14 concept What does momentum add to the gradient descent update $x_{k+1} = x_k - \eta \nabla f(x_k)$?

A A fraction of the previous step, so steady directions build speed
B A penalty on large parameters, so the learned weights stay small
C A second-derivative term, so the step size adapts to the curvature
D A random perturbation at every step, so it can escape local minima

Q15 calculate For $f(w) = (w-3)^2$, each gradient step multiplies the distance to 3 by $(1 - 2\eta)$. What happens with $\eta = 1$, starting at $w = 5$?

A It converges slowly towards 3
B It reaches 3 in a single step
C It moves steadily away from 3
D It bounces between 5 and 1

1.4 Notation, Probability and Jargon

Q16 calculate X is 0 with probability 0.75 and 10 with probability 0.25. What are $\mathbb{E}[X]$ and $\mathrm{Var}(X)$?

A 2.5 and 18.75
B 5.0 and 25.00
C 2.5 and 4.33
D 2.5 and 6.25

Q17 concept Which of these is a hyperparameter rather than a parameter?

A A hidden layer's weight matrix
B A single weight w_{12}
C The learning rate η
D An output unit's bias term

Q18 `interpret` A lecturer says a model "has high bias". In a discussion of underfitting, what do they mean?

- A Its bias terms b have unusually large values
- B Its model class makes systematic errors
- C It treats some groups of people unfairly
- D It changes a lot from one training sample to another

Q19 `concept` If w is a column vector (w_1, w_2, w_3), what is $w^\top x$ for $x = (x_1, x_2, x_3)$?

- A The vector $(w_1 x_1, w_2 x_2, w_3 x_3)$
- B A 3×3 matrix of all products
- C The vector w written as a row
- D The single number $w_1 x_1 + w_2 x_2 + w_3 x_3$

Q20 `concept` Why is the standard deviation often reported instead of the variance?

- A It is easier to differentiate than the variance
- B It is in the same units as the variable itself
- C It is always smaller than the variance is
- D It ignores the extreme values in a data set

Answer key

Each answer comes with the reasoning, including why the most tempting wrong option is wrong.

Q1 **A** · lesson 1.1

The model is the mathematical mapping with parameters; learning is the procedure (such as gradient descent) that estimates those parameters from examples. Choosing the straight-line form is a design decision made before learning, not the learning itself.

Q2 **A** · lesson 1.1

$5 \times 6 + 30 = 60$. 35 forgets to multiply by x, 41 adds the inputs, and 180 multiplies by 30 instead of adding it.

Q3 **A** · lesson 1.1

ML is the broad field of learning model parameters from data; DL is the subset using many-layered neural networks that learn their own representations. Plenty of ML models (decision trees, linear regression) are not neural networks.

Q4 **D** · lesson 1.1

Evaluation checks the model on data it hasn't learned from, which is the real goal (Section 8). It isn't where gradients come from (that's the learning step), and the loop refines the model rather than restarting it.

Q5 **A** · lesson 1.1

Backpropagation computes the gradient of the loss for every weight efficiently, which is what multi-layer training needs. The 1943 model was an abstraction of one neuron, ADALINE was a single-layer application, and 2012 came after backpropagation had long been in use.

Q6 **D** · lesson 1.2

There are labelled examples (hours, mark), so it's supervised, and the target is a number, so it's regression. Classification would predict a category such as pass/fail; clustering has no labels.

Q7 **A** · lesson 1.2

With no labels, the goal is to discover structure: unsupervised learning (clustering). Supervised learning needs target labels, and reinforcement learning needs an agent receiving rewards from actions.

Q8 **A** · lesson 1.2

The bracketed superscript indexes the training example. A power would be written without brackets, a feature is usually a subscript (x_j), and layer numbers use their own notation. Mixing these up is a classic source of errors.

Q9 **B** · lesson 1.2

A reinforcement learner acts in an environment and receives rewards or penalties, learning a policy. Supervised learning needs correct labels. Network depth and data type don't define either one.

Q10 **C** · lesson 1.2

A few labels (the one face you named) are combined with lots of unlabelled data (the clustered faces): semi-supervised learning. It isn't fully supervised, since most photos were never labelled, and no rewards are involved.

Q11 **A** · lesson 1.3

$f'(w) = 2(w - 3) = 4$ at $w = 5$, so $w_1 = 5 - 0.1 \times 4 = 4.6$. Adding the gradient (5.4) moves uphill; 4.8 forgets the factor 2; jumping straight to 3 would need $\eta = 0.5$.

Q12 **C** · lesson 1.3

An exploding loss is the signature of a learning rate that overshoots the minimum by more each step. A too-small rate makes progress slow but steady, a local minimum makes the loss flatten, and convexity helps rather than hurts.

Q13 **D** · lesson 1.3

A zero gradient is necessary for an interior minimum but not sufficient: maxima, saddle points and flat plateaus also have zero gradient. It says nothing about the learning rate.

Q14 **A** · lesson 1.3

Momentum adds $\alpha\Delta x_k = \alpha(x_k - x_{k-1})$: steps that agree accumulate, zig-zags cancel. Random noise, second derivatives and weight penalties are different ideas (noise injection, Newton-type methods, regularisation).

Q15 **D** · lesson 1.3

$1 - 2\eta = -1$, so $w - 3$ flips sign but keeps size 2: $5 \to 1 \to 5 \to \ldots$. One-step convergence needs $\eta = 0.5$ (factor 0), steady convergence needs $0 < \eta < 0.5$, and divergence needs $\eta > 1$.

Q16 **A** · lesson 1.4

$\mathbb{E}[X] = 0(0.75) + 10(0.25) = 2.5$. $\mathrm{Var}(X) = 0.75(0 - 2.5)^2 + 0.25(10 - 2.5)^2 = 4.6875 + 14.0625 = 18.75$. 5.0 and 25.00 treat both values as equally likely, 6.25 is only $(\mathbb{E}[X])^2$, and 4.33 is the standard deviation $\sqrt{18.75}$, not the variance.

Q17 **C** · lesson 1.4

Hyperparameters are set outside gradient-based learning (learning rate, layer sizes, dropout rate) and chosen on validation data. Weights and biases are parameters learned by gradient descent.

Q18 **B** · lesson 1.4

In the bias–variance sense, bias is systematic error from a too-simple model: underfitting. The word "bias" also names a neuron's b and unfair treatment of groups, so context decides; here it's about underfitting. High variance is the opposite problem.

Q19 **D** · lesson 1.4

$w^\top x$ is a row times a column: a dot product, one number. The element-wise product would be a vector, $xw^\top$ (column times row) would be a 3×3 matrix, and $w^\top$ alone is the row vector.

Q20 **B** · lesson 1.4

Variance is in squared units; its square root, the standard deviation, is in the original units, so it's easier to interpret. It isn't always smaller (a variance below 1 has a larger square root), and it is just as sensitive to extremes.