Q&A Lessons 3,4

01 - 07 Feb 2021

Poll results

slido

Table of contents

Multiple-choice poll (Multiple answers)

What videos were engaged?

0 2 5

(1/2)

[Lesson 01] Google Colab Introduction

76 %

[Lesson 01] Google Colab Cont.

72 %

[Lesson 02] The perceptron

80 %

[Lesson 02] Building neural network

80 %

[Lesson 02] Applying neural network

80 %

[Lesson 03] Training neural network

56 %

Multiple-choice poll (Multiple answers)

What videos were engaged?

0 2 5

(2/2)

[Lesson 04] Training neural network

36 %

What are your main difficulties with this methodology and material?

0 2 0

Conciliar com estágio
Material em língua inglesa
Falta de prática com a linguagem
engajamento
falta de disciplina
Densidade do conteúdo
tempo
gestão de tempo
nenhuma
Notebook

Tempo para estudar todos os notebooks

tá acelerado, entrei depois
Carga de conhecimento anterior
Sinto falta de hands on
os primeiros notebooks são enormes
Tempo para fazer os notebooks
Dificuldade na língua inglesa
Quantidade de tópicos no notebook

Open text poll

How are you consuming the course material?

0 1 4

- Video > notebook
- video > notebook
- Assistir os videos, acompanhando os notebooks e as duvidas vou buscar na bibliografia.
- video e depois o notebook especifico
- Assisto o vídeo e depois o notebook
- Video depois notebooka
- Vídeo e depois notebook
- Assisto o vídeo e depois faço o notebook
- Vídeo depois o notebook
- Vídeos primeiro e notebook depois
- assistindo videos e indo pro notebook; tb lendo coisa por fora
- Assisto os vídeos e depois vou para o notebook
- Vídeo depois notebook
- Assisto o vídeo e logo em seguida tento fazer o notebook

Multiple-choice poll

Survey (1/12)

0 2 4

A demographic dataset with statistics on different cities' population, GDP per capita, economic growth is an example of “unstructured” data because it contains data coming from different sources.

False ✓

88 %

True

13 %

Multiple-choice poll

Survey (2/12)

0 2 3

What does a neuron compute?

A neuron computes an activation function followed by a linear function (z = Wx + b)

13 %

A neuron computes a linear function (z = Wx + b) followed by an activation function

61 %

A neuron computes a function g that scales the input x linearly (Wx + b)

17 %

A neuron computes the mean of all features before applying the output to an activation function

9 %

Multiple-choice poll (Multiple answers)

Survey (3/12)

0 2 3

Among the following, which ones are "hyperparameters"?

(1/2)

Bias

48 %

Size of the hidden layers

43 %

Learning rate (alpha)

52 %

Weights

30 %

Number of iterations

43 %

Multiple-choice poll (Multiple answers)

Survey (3/12)

0 2 3

Among the following, which ones are "hyperparameters"?

(2/2)

Number of layers L in the neural network ✓

39 %

Activation values of g(z)

22 %

Multiple-choice poll

Survey (4/12)

0 2 4

Which of the following statements is true?

The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers.

79 %

The earlier layers of a neural network are typically computing more complex features of the input than the deeper layers.

21 %

Multiple-choice poll

Survey (5/12)

0 2 4

The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False?

False

42 %

True

58 %

Multiple-choice poll

Survey (6/12)

0 2 3

You are building a binary classifier for recognizing "Tapioca" (y=1) vs. "Ginga com Tapioca" (y=0). Which one of these activation functions would you recommend using for the output layer?

Relu

22 %

Leaky Relu

9 %

Sigmoid

52 %

Tanh

17 %

Multiple-choice poll

Survey (7/12)

0 2 2

Which of these statements about mini-batch gradient descent do you agree with?

You should implement mini-batch gradient descent without an explicit for-loop over different mini-batches, so that the algorithm processes all mini-batches at the same time (vectorization)

23 %

Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.

27 %

One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.

50 %

Multiple-choice poll (Multiple answers)

Survey (8/12)

0 2 0

Why is the best mini-batch size usually not 1 and not m, but instead something in-between?

If the mini-batch size is m, you end up with batch gradient descent, which has to process the whole training set before making progress.

65 %

If the mini-batch size is m, you end up with stochastic gradient descent, which is usually slower than mini-batch gradient descent.

20 %

If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.

40 %

If the mini-batch size is 1, you end up having to process the entire training set before making any progress.

20 %

Multiple-choice poll

Survey (9/12)

0 1 5

Suppose your learning algorithm's cost J, plotted as a function of the number of iterations, looks like this:
(1/2)

Whether you're using batch gradient descent or mini-batch gradient descent, something is wrong.

7 %

If you're using mini-batch gradient descent, this looks acceptable. But if you're using batch gradient descent, something is wrong.

80 %

Whether you're using batch gradient descent or mini-batch gradient descent, this looks acceptable.

7 %

Multiple-choice poll

Survey (9/12)

0 1 5

Suppose your learning algorithm’s cost J, plotted as a function of the number of iterations, looks like this:
(2/2)

If you’re using mini-batch gradient descent, something is wrong. But if you’re using batch gradient descent, this looks acceptable.

7 %

Multiple-choice poll

Survey (10/12)

0 1 6

Consider this figure: these plots were generated with gradient descent; with gradient descent with momentum (β = 0.5) and gradient descent with momentum (β = 0.9). Which curve corresponds to which algorithm?
(1/2)

(1) is gradient descent with momentum (small β), (2) is gradient descent with momentum (small β), (3) is gradient descent.

19 %

(1) is gradient descent. (2) is gradient descent with momentum (large β) . (3) is gradient descent with momentum (small β)

6 %

(1) is gradient descent with momentum (small β). (2) is gradient descent. (3) is gradient descent with momentum (large β)

6 %

Multiple-choice poll

Survey (10/12)

0 1 6

Consider this figure: these plots were generated with gradient descent; with gradient descent with momentum (β = 0.5) and gradient descent with momentum (β = 0.9). Which curve corresponds to which algorithm?

(2/2)

(1) is gradient descent. (2) is gradient descent with momentum (small β). (3) is gradient descent with momentum (large β)

69 %

Multiple-choice poll (Multiple answers)

Survey (11/12)

0 1 6

Suppose batch gradient descent in a deep network is taking excessively long to find a value of the parameters that achieves a small value for the cost function J. Which of the following techniques could help:
(1/2)

Try mini-batch gradient descent. ✓

75 %

Try initializing all the weights to zero.

13 %

Try better random initialization for the weights. ✓

44 %

Try tuning the learning rate α (alpha). ✓

81 %

Multiple-choice poll (Multiple answers)

Survey (11/12)

0 1 6

Suppose batch gradient descent in a deep network is taking excessively long to find a value of the parameters that achieves a small value for the cost function J. Which of the following techniques could help:
(2/2)

Try using Adam. ✓

56 %

Multiple-choice poll

Survey (12/12)

0 1 8

Which of the following statements about Adam is False?

Adam combines the advantages of RMSProp and momentum.

6 %

We usually use “default” values for the hyperparameters β1, β2 and ε in Adam (β1 = 0.9, β2 = 0.999, 10^{-8})

6 %

The learning rate hyperparameter α in Adam usually needs to be tuned.

28 %

Adam should be used with batch gradient computations, not with mini-batches.

61 %