coursera

검색

- 설정
- 업데이트
- 성취도
- 도움말 센터
- 로그아웃
- Youngeun In

테스트: Optimization Algorithms
10개의 질문

- **Programming Assignment**
- **Heroes of Deep Learning (Optional)**

테스트테스트 • 20 min20 minutes

Optimization Algorithms

과제 제출
기한년 8월 30일 오후 3:59 KST년 8월 30일 오후 3:59 KST
시도하기8 hours당 3회
다시 시도해주십시오

성적 받기
통과 점수:80% 이상
성적
100%
피드백 보기
최고 점수가 유지됩니다.

탐색 확인

이 페이지에서 나가시겠습니까?
이 페이지에 머물기 이 페이지에서 나가기

Optimization Algorithms
성적 평가 퀴즈 • 20 min

만료 년 8월 30일 오후 3:59 KST

축하합니다! 통과하셨습니다!
통과 점수: 80% 이상
학습 계속하기
성적
100%

Optimization Algorithms

최신 제출물 성적
100%
1.
질문 1

Which notation would you use to denote the 3rd layer's activations when the input is the 7th example from the 8th minibatch?

1 / 1점

$a^{[3]\{8\}(7)}$a[3]{8}(7)

$a^{[8]\{7\}(3)}$a[8]{7}(3)

$a^{[8]\{3\}(7)}$a[8]{3}(7)

$a^{[3]\{7\}(8)}$a[3]{7}(8)

맞습니다
2.
질문 2

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

1 / 1점

○○

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).

◉○

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

○○

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

✓
맞습니다
3.
질문 3

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

1 / 1점
☑☐

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

✓
맞습니다
☑☐

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.

✓
맞습니다
☐☐

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

☐☐

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

4.
질문 4

Suppose your learning algorithm's cost JJ, plotted as a function of the number of iterations, looks like this:

Which of the following do you agree with?

1 / 1점
◉○

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

○○

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

○○

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

○○

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

✓
맞습니다
5.
질문 5

Suppose the temperature in Casablanca over the first two days of January are the same:

Jan 1st: $\theta_1 = 10^o C$θ1 = 10oC

Jan 2nd: $\theta_2 10^o C$θ210oC

(We used Fahrenheit in lecture, so will use Celsius here in honor of the metric world.)

Say you use an exponentially weighted average with $\beta = 0.5$β = 0.5 to track the temperature: $v_0 = 0$v0 = 0, $v_t = \beta v_{t-1} + (1-\beta)\theta_t$vt = βvt−1 + (1 − β)θt. If v_2v2 is the value computed after day 2 without bias correction, and $v_2^{corrected}$v2corrected is the value you compute with bias correction. What are these values? (You might be able to do this without a calculator, but you don't actually need one. Remember what bias correction is doing.)

1 / 1점
◉○

$v_2 = 7.5$v2 = 7.5, $v_2^{corrected} = 10$v2corrected = 10

○○

$v_2 = 10$v2 = 10, $v_2^{corrected} = 7.5$v2corrected = 7.5

○○

$v_2 = 10$v2 = 10, $v_2^{corrected} = 10$v2corrected = 10

○○

$v_2 = 7.5$v2 = 7.5, $v_2^{corrected} = 7.5$v2corrected = 7.5

✓
맞습니다
6.
질문 6

Which of these is NOT a good learning rate decay scheme? Here, t is the epoch number.

1 / 1점
◉○

$\alpha = e^t \alpha_0$α = etα0

○○

$\alpha = 0.95^t \alpha_0$α = 0.95tα0

○○

$\alpha = \frac{1}{1 + 2 * t}\alpha_0$α = 1+2∗t1α0

○○

$\alpha = \frac{1}{\sqrt{t}}\alpha_0$α = t
√
1α0

✓
맞습니다
7.
질문 7

You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: $v_t = \beta v_{t-1} + (1-\beta)\theta_t$vt = βvt−1 + (1 − β)θt. The red line below was computed using $\beta = 0.9$β = 0.9. What would happen to your red curve as you vary ββ? (Check the two that apply)

?
1 / 1점
☐☐

Decreasing ββ will shift the red line slightly to the right.

☑☐

Increasing ββ will shift the red line slightly to the right.

✓
맞습니다

True, remember that the red line corresponds to $\beta = 0.9$β = 0.9. In lecture we had a green line $$\beta = 0.98) that is slightly shifted to the right.

☑☐

Decreasing ββ will create more oscillation within the red line.

✓
맞습니다

True, remember that the red line corresponds to $\beta = 0.9$β = 0.9. In lecture we had a yellow line $$\beta = 0.98 that had a lot of oscillations.

☐☐

Increasing ββ will create more oscillations within the red line.

8.
질문 8

Consider this figure:

?

These plots were generated with gradient descent; with gradient descent with momentum (ββ = 0.5) and gradient descent with momentum (ββ = 0.9). Which curve corresponds to which algorithm?

1 / 1점
○○

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

◉○

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

○○

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

○○

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

✓
맞습니다
9.
질문 9

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 $J(W^{[1]}, b^{[1]}, \ldots, W^{[L]}, b^{[L]})$
J (W[1], b[1], ..., W[L], b[L]). Which of the following techniques could help find parameter values that attain a small value for$J$J ? (Check all that apply)

1 / 1점
☑☐

Try tuning the learning rate αα

✓
맞습니다
☐☐

Try initializing all the weights to zero

☑☐

Try using Adam

✓
맞습니다
☑☐

Try mini-batch gradient descent

✓
맞습니다
☑☐

Try better random initialization for the weights

✓
맞습니다
10.
질문 10

Which of the following statements about Adam is False?

1 / 1점

○○

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

○○

We usually use “default” values for the hyperparameters β_1, β_2β1, β2 and εε in Adam ($\beta_1 = 0.9$β1 = 0.9, $\beta_2 = 0.999$β2 = 0.999, $\varepsilon = 10^{-8}$ε = 10−8)

◉○

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

○○

Adam combines the advantages of RMSProp and momentum

✓
맞습니다