축하합니다! 통과하셨습니다!

받은 학점 100% **최신 제출물 학점** 100% **통과 점수:** 80% 이상

다음 항목으로 이동

1. Which notation would you use to denote the 4th layer's activations when the input is the 7th example from the 3rd mini-batch? **1 / 1점**

- (●) $a^{[4]\{3\}(7)}$
- ( ) $a^{[3]\{7\}(4)}$
- ( ) $a^{[7]\{3\}(4)}$

더 보기

✓ 맞습니다
Yes. In general $a^{[l]\{t\}(k)}$ denotes the activation of the layer l when the input is the example k from the mini-batch t.

2. Suppose you don't face any memory-related problems. Which of the following make more use of vectorization. **1 / 1점**

- (●) Batch Gradient Descent
- ( ) Stochastic Gradient Descent, Batch Gradient Descent, and Mini-Batch Gradient Descent all make equal use of vectorization.
- ( ) Mini-Batch Gradient Descent with mini-batch size $m/2$.
- ( ) Stochastic Gradient Descent

더 보기

✓ 맞습니다
Yes. If no memory problem is faced, batch gradient descent processes all of the training set in one pass, maximizing the use of vectorization.

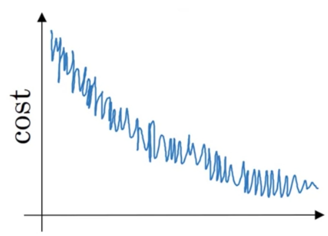
3. We usually choose a mini-batch size greater than 1 and less than m, because that way we make use of vectorization but not fall into the slower case of batch gradient descent. **1 / 1점**

- ( ) False
- (●) True

더 보기

✓ 맞습니다
Correct. Precisely by choosing a batch size greater than one we can use vectorization; but we choose a value less than m so we won't end up using batch gradient descent.

4. While using mini-batch gradient descent with a batch size larger than 1 but less than m, the plot of the cost function J looks like this: **1 / 1점**



You notice that the value of J is not always decreasing. Which of the following is the most likely reason for that?

- ( ) The algorithm is on a local minimum thus the noisy behavior.
- ( ) A bad implementation of the backpropagation process, we should use gradient check to debug our implementation.
- (●) In mini-batch gradient descent we calculate $J(\hat{y}^{\{t\}}, y^{\{t\}})$ thus with each batch we compute over a new set of data.
- ( ) You are not implementing the moving averages correctly. Using moving averages will smooth the graph.

더 보기

✓ 맞습니다
Yes. Since at each iteration we work with a different set of data or batch the loss function doesn't have to be decreasing at each iteration.

5. Suppose the temperature in Casablanca over the first two days of January are the same: **1 / 1점**

Jan 1st: $\theta_1 = 10^oC$

Jan 2nd: $\theta_2 = 10^oC$

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

Say you use an exponentially weighted average with $\beta = 0.5$ to track the temperature: $v_0 = 0, v_t = \beta v_{t-1} + (1-\beta)\theta_t$. If v_2 is the value computed after day 2 without bias correction, and $v_2^{corrected}$ 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.)

- ( ) $v_2 = 7.5$, $v_2^{corrected} = 7.5$
- ( ) $v_2 = 10$, $v_2^{corrected} = 10$
- ( ) $v_2 = 10$, $v_2^{corrected} = 7.5$
- (●) $v_2 = 7.5$, $v_2^{corrected} = 10$

더 보기

✓ 맞습니다

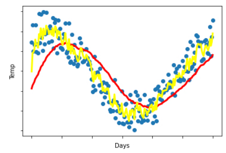
6. Which of the following is true about learning rate decay? **1 / 1점**

- ( ) The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take larger steps to accelerate the convergence.
- (•) The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take smaller steps to prevent large oscillations.
- ( ) It helps to reduce the variance of a model.
- ( ) We use it to increase the size of the steps taken in each mini-batch iteration.

더 보기

맞습니다
Correct. Reducing the learning rate with time reduces the oscillation around a minimum.

7. You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: $v_t = \beta v_{t-1} + (1-\beta)\theta_t$. The yellow and red lines were computed using values $beta_1$ and $beta_2$ respectively. Which of the following are true? 1 / 1점

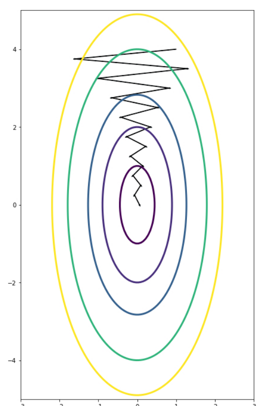


- ( ) $\beta_1 > \beta_2$.
- (•) $\beta_1 < \beta_2$.
- ( ) $\beta_1 = \beta_2$.
- ( ) $\beta_1 = 0, \beta_2 > 0$.

더 보기

맞습니다
Correct. $\beta_1 < \beta_2$ since the yellow curve is noisier.

8. Consider the figure: 1 / 1점



Suppose this plot was generated with gradient descent with momentum $\beta = 0.01$. What happens if we increase the value of β to 0.1?

- ( ) The gradient descent process moves more in the horizontal and the vertical axis.
- (•) The gradient descent process moves less in the horizontal direction and more in the vertical direction.
- ( ) The gradient descent process starts oscillating in the vertical direction.
- ( ) The gradient descent process starts moving more in the horizontal direction and less in the vertical.

더 보기

맞습니다
Yes. The use of a greater value of β causes a more efficient process thus reducing the oscillation in the horizontal direction and moving the steps more in the vertical direction.

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

- [x] Try better random initialization for the weights

  ✓ Correct

- [x] Try using Adam

  ✓ Correct

- [x] Try tuning the learning rate α

  ✓ Correct

- [x] Try mini-batch gradient descent

  ✓ Correct

- [ ] Try initializing all the weights to zero

더 보기

맞습니다
Great, you got all the right answers.

10. Which of the following are true about Adam? 1 / 1점

- (selected) Adam combines the advantages of RMSProp and momentum.
- Adam automatically tunes the hyperparameter α.
- Adam can only be used with batch gradient descent and not with mini-batch gradient descent.
- The most important hyperparameter on Adam is ϵ and should be carefully tuned.

더 보기

맞습니다
True. Precisely Adam combines the features of RMSProp and momentum that is why we use two-parameter β_1 and β_2, besides ϵ.