Quiz Submissions - Quiz 7

Mike Gao (username: zenghao.gao@mail.mcgill.ca)

Attempt 1

Written: Nov 2, 2020 7:41 AM - Nov 2, 2020 4:25 PM

Submission View

Released: Nov 4, 2020 12:30 AM

Gradient Descent

Question 1 — 1 / 1 point

Given two convex functions f and g, select all of the statements that are true.

✓☐ Second derivatives of either f or g could be negative.

✓☑ **f + g is convex.**

✓☐ **min(f, g) is convex.**

✓☑ **If g is monotonically increasing then g(f(x)) is convex.**

Question 2 — 1 / 1 point

Choose all of the True statements regarding SGD.

✓☑ **Using a small learning rate could cause the optimizer to converge more slowly.**

✓☐ While optimizing a convex function using SGD, it is guaranteed that given enough time, the optimizer will always converge to the global minimum regardless of the learning rate.

✓☑

1. Using the following sequence of learning rate will always result in converging to a local minimum.

$$\alpha = \frac{1}{t+1}$$

✓☐

The main idea behind SGD is that each step is always in the right direction when doing the approximation.

Question 3 **0 / 1 point**

The larger batch size in SGD converges faster because they increase the variance in the gradient estimation of SGD. True/False

✗ ● True

➡ ○ False

▼ Hide Feedback

Larger batch size <u>reduces</u> the variance in gradient estimation of SGD thus leading to faster convergence.

Question 4 **1 / 1 point**

Suppose hypothesis has form

$$\hat{y} = \sum_{i=1}^{n} w_{1,i} sin(w_{2,i} + x_i)$$

Here w1 and w2 are the weight parameters. We are looking at minimizing the mean squared error. Is it possible that the gradient descent solution would be able to lead to a unique solution? Choose one of the following:

○ Yes

✓ ● No

▼ Hide Feedback

$$J(w1,w2)=\sum_{i=1}^{m}(y_i-\sum_{j=1}^{n}w_{1,j}sin(w_{2,j}+x_j))^2$$

$$\nabla_{w_{1,i}}J(w1,w2)=-2\sum_{i=1}^{m}\Big(y_i-\sum_{j=1}^{n}w_{1,j}sin(w_{2,j}+x_j)\Big)\Big(sin(w_{2,j}+x_j)\Big)$$

$$\nabla_{w_{2,i}}J(w1,w2)=-2\sum_{i=1}^{m}\Big(y_i-\sum_{j=1}^{n}w_{1,j}sin(w_{2,j}+x_j)\Big)\Big(w_{1,j}cos(w_{2,j}+x_j)\Big)$$

The periodic function of the hypothesis would result in the same result from many different values of the parameter. Therefore, not a unique solution.

Question 5 **1 / 1 point**

Similar to previous question with a hypothesis class

$$\hat{y}=\sum_{j=1}^{n}log(e^{w_j x_j})$$

With “w” as the parameter. Will the gradient descent be able to find weight which will converge to a unique solution? Choose on the following?

✓ ● Yes

○ No

▼ Hide Feedback

$$\hat{y}=\sum_{j=1}^{n}w_j x_j$$

The hypothesis is linear function and with mean squared loss function, gradient descent would lead to global optimum given the learning rates are following the conditions.

Regularization

Question 6 **0 / 1 point**

In a linear regression problem suppose a regularization penalty is added. Some of the coefficients of “w” parameter are zeroed. Choose ALL the different penalties which could have been chosen.

➡ ✓ ☑ L0 norm

➡ ✓ ☑ L1 norm

✗ ☑ L2 norm

Question 7 **1 / 1 point**

Select all correct statements:

✓ ☐ L2 penalty in a ridge regression is assuming a Laplace prior on the weights.

✓ ☑ A classifier trained on less data is more likely to overfit.

✓ ☑ When you develop a model with a 50-50 train-test splits based on m data points, the difference between training error and test error decreases as m increases in general.

✓ ☐ Ridge regression often sets several of the weights to zero.

Question 8 **1 / 1 point**

Adding a Gaussian prior on least square linear regression is equivalent to:

✓ ☑ L2 regularization

✓ ☐ L1 regularization

✓ ☐ Incorporating Laplace penalty term

Attempt Score: 6 / 8 - 75 %

Overall Grade (highest attempt): 6 / 8 - 75 %

Done