1. In Module 3 assignment, there were 194 features (an intercept + one feature for each of the 193 important words). In this assignment, we will use stochastic gradient ascent to train the classifier using logistic regression. How does the changing the solver to stochastic gradient ascent affect the number of features? 1 point

- ○ Increases
- ○ Decreases
- ◉ Stays the same

2. Recall from the lecture and the earlier assignment, the log likelihood (without the averaging term) is given by 1 point

$$\ell\ell(\mathbf{w}) = \sum_{i=1}^{N} \left((\mathbf{1}[y_i = +1] - 1)\mathbf{w}^T h(\mathbf{x}_i) - \ln\left(1 + \exp(-\mathbf{w}^T h(\mathbf{x}_i))\right) \right)$$

whereas the average log likelihood is given by

$$\ell\ell_A(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left((\mathbf{1}[y_i = +1] - 1)\mathbf{w}^T h(\mathbf{x}_i) - \ln\left(1 + \exp(-\mathbf{w}^T h(\mathbf{x}_i))\right) \right)$$

How are the functions $\ell\ell(\mathbf{w})$ and $\ell\ell_A(\mathbf{w})$ related?

- ○ $\ell\ell_A(\mathbf{w}) = \ell\ell(\mathbf{w})$
- ◉ $\ell\ell_A(\mathbf{w}) = (1/N) \cdot \ell\ell(\mathbf{w})$
- ○ $\ell\ell_A(\mathbf{w}) = N \cdot \ell\ell(\mathbf{w})$
- ○ $\ell\ell_A(\mathbf{w}) = \ell\ell(\mathbf{w}) - \|\mathbf{w}\|$

3. Refer to the sub-section Computing the gradient for a single data point. 1 point

The code block above computed

$$\frac{\partial \ell_i(\mathbf{w})}{\partial w_j}$$

for j = 1 and i = 10. Is this quantity a scalar or a 194-dimensional vector?

- ◉ A scalar
- ○ A 194-dimensional vector

4. Refer to the sub-section Modifying the derivative for using a batch of data points. 1 point

The code block computed

$$\sum_{s=i}^{i+B} \frac{\partial \ell_s(\mathbf{w})}{\partial w_j}$$

for j = 10, i = 10, and B = 10. Is this a scalar or a 194-dimensional vector?

- ◉ A scalar

- [ ] A 194-dimensional vector

5. For what value of B is the term

$$\sum_{s=1}^{B} \frac{\partial \ell_s(\mathbf{w})}{\partial w_j}$$

the same as the full gradient

$$\frac{\partial \ell(\mathbf{w})}{\partial w_j}$$

? A numeric answer is expected for this question. Hint: consider the training set we are using now.

1 point

47780

6. For what value of batch size B above is the stochastic gradient ascent function logistic_regression_SG act as a standard gradient ascent algorithm? A numeric answer is expected for this question. Hint: consider the training set we are using now.

1 point

47780

7. When you set batch_size = 1, as each iteration passes, how does the average log likelihood in the batch change?

1 point

- [ ] Increases
- [ ] Decreases
- [x] Fluctuates

8. When you set batch_size = len(feature_matrix_train), as each iteration passes, how does the average log likelihood in the batch change?

1 point

- [x] Increases
- [ ] Decreases
- [ ] Fluctuates

9. Suppose that we run stochastic gradient ascent with a batch size of 100. How many gradient updates are performed at the end of two passes over a dataset consisting of 50000 data points?

1 point

1000

10. Refer to the section Stochastic gradient ascent vs gradient ascent. 1 point

In the first figure, how many passes does batch gradient ascent need to achieve a similar log likelihood as stochastic gradient ascent?

- [ ] It's always better
- [ ] 10 passes
- [ ] 20 passes
- [x] 150 passes or more

11. Questions 11 and 12 refer to the section Plotting the log likelihood as a function of passes for each step size. 1 point
Which of the following is the worst step size? Pick the step size that results in the lowest log likelihood in the end.

- [ ] 1e-2
- [ ] 1e-1
- [ ] 1e0
- [ ] 1e1
- [x] 1e2

12. Questions 11 and 12 refer to the section Plotting the log likelihood as a function of passes for each step size. 1 point
Which of the following is the best step size? Pick the step size that results in the highest log likelihood in the end.

- [ ] 1e-4
- [ ] 1e-2
- [x] 1e0
- [ ] 1e1
- [ ] 1e2