MACHINE LEARNING

In Q1 to Q8, only one option is correct, choose the correct option:

1. The computation complexity of linear regression is:
   A) $O(n^{2.4})$ B) $O(n)$
   C) $O(n^2)$ D) $O(n^3)$

Answer: (B) = $O(n)$

2. Which of the following can be used to fit non-linear data?
   A) Lasso Regression B) Logistic Regression
   C) Polynomial Regression D) Ridge Regression

Answer: (C) = Polynomial Regression.

3. Which of the following can be used to optimize the cost function of Linear Regression?
   A) Entropy B) Gradient Descent
   C) Pasting D) None of the above

Answer: (B) = Gradient Descent

4. Which of the following method does not have closed form solution for its coefficients?
   A) Extrapolation B) Ridge
   C) Lasso D) Elastic Nets

Answer: (c) = Lasso

5. Which gradient descent algorithm always gives optimal solution?
   A) Stochastic Gradient Descent B) Mini-Batch Gradient Descent
   C) Batch Gradient Descent D) All of the above

Answer: (C) = Batch Gradient Descent.

6. Generalization error measures how well a model performs on training data.
   A) True B) False

Answer: (B) = False

7. The cost function of linear regression can be given as $J(w_0, w_1) = \frac{1}{2m}\sum_{i=1}^{m} \left(w_0 + w_1 x^{(i)} - y^{(i)}\right)^2$. The half term at start is due to:

A) Scaling cost function by half makes gradient descent converge faster.
B) Presence of half makes it easy to grid search.
C) It does not matter whether half is there or not.
D) None of the above.

Answer: (A) = Scaling cost function by half makes gradient descent converge faster.

8. Which of the following will have symmetric relation between dependent variable and independent variable?

A) Regression
B) Correlation
C) Both of Them
D) None of these

Answer: (C) = Both of them.

In Q9 to Q11, more than one options are correct, choose all the correct options:

9. Which of the following is true about Normal Equation is used to compute the coefficient Of the Linear Regression?

A) We don't have to choose the learning rate.
B) It becomes slow when number of features are very large.
C) We need to iterate.
D) It does not make use of dependent variable.

Answer: (A) = We don't have to choose the learning rate.
(B) = It becomes slow when number of features are very large.

10. Which of the following statement/s are true if we generated data with the help of polynomial features with 5 degrees of freedom which perfectly fits the data?

A) Linear Regression will have high bias and low variance.
B) Linear Regression will have low bias and high variance.
C) Polynomial with degree 5 will have low bias and high variance.
D) Polynomial with degree 5 will have high bias and low variance.

Answer: (A) & (D)
(A) Linear Regression will have high bias and low variance.
(D) Polynomial with degree 5 will have high bias and low variance.

11. Which of the following sentence is false regarding regression?
    A) It relates inputs to outputs.
    B) It is used for prediction.
    C) It discovers causal relationships.
    D) No inference can be made from regression line.

Answer: (C) & (D)
(C) = It discovers causal relationships.
(D) = No inference can be made from regression line.

Q12 & Q13 are subjective answer types question, answer them briefly:

12. Which Linear Regression training algorithm can we use if we have a training set with millions of features?

Answer:

Training set with millions of features is too large to handle for any local device and also it is computationally complex with this much of data so **The Normal Equation Method** won't be a good choice because it is computationally inefficient. For this type of data we could use following Linear Regression training algorithm:

1. Batch Gradient Descent.
2. Stochastic Gradient Descent.
3. Mini-Batch Gradient Descent.

1. If we have enough memory to load the data in that case Batch Gradient Descent could be used because Batch Gradient Descent use the whole training data per epoch.

2. But Stochastic Gradient Descent and Mini-Batch Gradient neither of them use the whole training data per epoch, Stochastic Gradient Descent use only single training example per epoch whereas.

3. Mini-Batch Gradient Descent lies between Batch Gradient Descent & Stochastic Gradient Descent it uses a small portion of training data per epoch, which will be good for training data with millions of features.

13. Which algorithms will not suffer or might suffer, if the features in training sets have very different scales?

Answer:

If the features in our training set have very different scale, then the Gradient Descent is the model which will suffer from features of different scales, because with different scale features training set Gradient Descent will take a very long time to converge because the cost function of Gradient Descent will have an Elongated Shape, which will lead to longer time to reach the global maximum. There is one simple and very famous technique to solve this problem is that before training our model we need to scale our feature data.

The Normal Equation will work just fine with different scales of features. We don't even need to scale the data in Normal Equation.