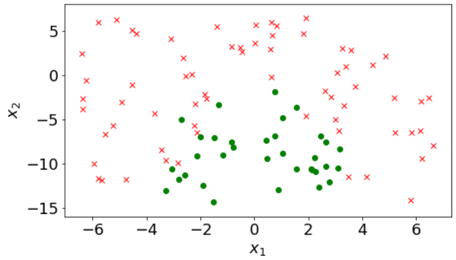
1. Machine learning contains a large number of matrix multiplications. Now we need to calculate the product of three matrices A, B, C. Suppose the dimensions of A, B, C are $m \times n, n \times p, p \times q$ respectively, where $m < n < p < q$. Which of the following calculations is correct and also the most efficient.
   A. $(AB)C$
   B. $(AC)B$
   C. $A(BC)$
   D. $A(CB)$
2. In Gradient Descent, what will happen if the learning rate is too small?
   A. The model may not converge
   B. The model may converge fast
   C. The model may converge slowly
   D. The model may converge properly
3. In Gradient Descent, what will happen if the learning rate is too large?
   A. The model may not converge
   B. The model may converge fast
   C. The model may converge slowly
   D. The model may converge properly
4. Given a linear function $\boldsymbol{y} = \boldsymbol{wx} + \boldsymbol{b}$ and an activation function $\boldsymbol{y} = \sigma(\boldsymbol{x})$, what is a neuron?
   A. $\boldsymbol{w}\sigma(\boldsymbol{x}) + \boldsymbol{b}$
   B. $\sigma(\boldsymbol{wx} + \boldsymbol{b})$
   C. $\boldsymbol{wx} + \boldsymbol{b} + \sigma(\boldsymbol{x})$
   D. $\sigma(\boldsymbol{wx}) + \boldsymbol{b}$
5. Suppose $\boldsymbol{z} \in \mathbb{R}^D$ is the output of a model. Which of the following functions should you use to get the prediction $\hat{\boldsymbol{y}}$ if you want to do linear regression?
   A. $\hat{\boldsymbol{y}} = \boldsymbol{z}$, where $D = 1$
   B. $\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z})$, where $\sigma(.)$ is the Sigmoid function and $D = 1$
   C. $\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z})$, where $\sigma(.)$ is the Softmax function and $D = 3$
   D. $\hat{\boldsymbol{y}} = \boldsymbol{wz} + \boldsymbol{b}$ and $D = 3$
6. Suppose $\boldsymbol{z} \in \mathbb{R}^D$ is the output of a model. Which of the following functions should you use to get the prediction $\hat{\boldsymbol{y}}$ if you want to do 2-class classification?
   A. $\hat{\boldsymbol{y}} = \boldsymbol{z}$, where $D = 1$
   B. $\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z})$, where $\sigma(.)$ is the Sigmoid function and $D = 1$
   C. $\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z})$, where $\sigma(.)$ is the Softmax function and $D = 3$
   D. $\hat{\boldsymbol{y}} = \boldsymbol{wz} + \boldsymbol{b}$ and $D = 3$
7. Suppose $\boldsymbol{z} \in \mathbb{R}^D$ is the output of a model. Which of the following functions should you use to get the prediction $\hat{\boldsymbol{y}}$ if you want to do N-class classification?
   A. $\hat{\boldsymbol{y}} = \boldsymbol{z}$, where $D = 1$
   B. $\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z})$, where $\sigma(.)$ is the Sigmoid function and $D = 1$
   C. $\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z})$, where $\sigma(.)$ is the Softmax function and $D = N$
   D. $\hat{\boldsymbol{y}} = \boldsymbol{wz} + \boldsymbol{b}$ and $D = N$
8. In neural networks, which of the following parameters is not the hyperparameter you need to set.
   A. The number of layers L
   B. The number of neurons in the l layer $D^{[l]}$
   C. The bias in the in the l layer $\boldsymbol{b}^{[l]}$

D. The learning rate λ

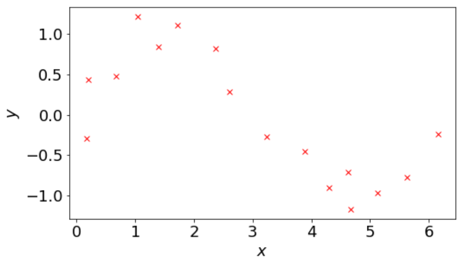
9. Which of the following loss functions is the loss function for linear regression.
   A. $(\hat{y} - y)^2/2$
   B. $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$
   C. $-\boldsymbol{y}^T \log \hat{\boldsymbol{y}}$
   D. Any one of the above functions
10. Which of the following loss functions is the loss function for logistic regression.
   A. $(\hat{y} - y)^2/2$
   B. $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$
   C. $-\boldsymbol{y}^T \log \hat{\boldsymbol{y}}$
   D. Any one of the above functions
11. Which of the following loss functions is the loss function for Softmax regression.
   A. $(\hat{y} - y)^2/2$
   B. $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$
   C. $-\boldsymbol{y}^T \log \hat{\boldsymbol{y}}$
   D. Any one of the above functions
12. Which of the following loss functions is the loss function for neural networks.
   A. $(\hat{y} - y)^2/2$
   B. $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$
   C. $-\boldsymbol{y}^T \log \hat{\boldsymbol{y}}$
   D. Any one of the above functions
13. What is a linear function in 2-dimensional space.
   A. A line
   B. A plane
   C. A hyperplane
   D. A point
14. What is a linear function in 3-dimensional space.
   A. A line
   B. A plane
   C. A hyperplane
   D. A point
15. What is a linear function in 4-dimensional space.
   A. A line
   B. A plane
   C. A hyperplane
   D. A point
16. In Perceptron, how to determine whether sample $\boldsymbol{x}$ is mis-classified.
   A. See if $y(\boldsymbol{w}^T\boldsymbol{x} + b) < 0$
   B. See if $(\boldsymbol{w}^T\boldsymbol{x} + b) < 0$
   C. See if $y < 0$
   D. See if $(\boldsymbol{w}^T\boldsymbol{x} + b) > 0$

17. Given the following training set, which of the following answers is correct to train a classification model $\hat{y} = \sigma(w_0 + w_1 x_1 + w_2 x_2)$, where $\sigma(.)$ is the Sigmoid function.



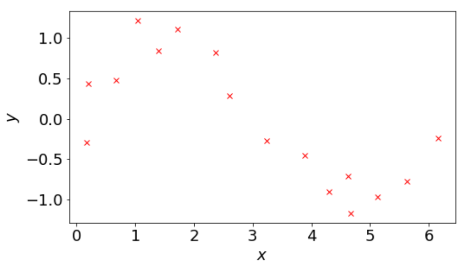
A. The model has a high bias and a low variance
B. The model will be overfitting
C. The model has a low bias and a low variance
D. The model has a low bias and a high variance

18. Given the following training set, which of the following answers is correct to train a regression model $\hat{y} = w_0 + w_1 x$.



A. The model has a high bias and a high variance
B. The model will be underfitting
C. The model has a low bias and a low variance
D. The model has a low bias and a high variance

19. Given the following training set, which of the following answers is correct to train a regression model $\hat{y} = w_0 + w_1 x + w_2 x + w_3 x + w_4 x + w_5 x + w_6 x + w_7 x + w_8 x + w_9 x$.



A. The model has a high bias and a high variance
B. The model will be underfitting
C. The model has a low bias and a low variance
D. The model has a low bias and a high variance

20. Given a cost function with L^2 regularization $\mathcal{J}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$, which of the following answers is correct when $\lambda = 0$?

A. It is equivalent to the cost function $\mathcal{J}(\boldsymbol{w})$ without any regularization
B. It leads to a result that every $w_i \approx 0$
C. It successfully solves the problem of overfitting
D. It brings overfitting to the model

21. Given a cost function with L^2 regularization $\mathcal{J}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$, which of the following answers is correct when $\lambda \to \infty$?
   A. It is equivalent to the cost function $\mathcal{J}(\boldsymbol{w})$ without any regularization
   B. It leads to a result that every $w_i \approx 0$
   C. It successfully solves the problem of underfitting
   D. It brings overfitting to the model
22. Suppose the dataset contains two positive samples $\boldsymbol{x}^{(1)} = [2,2]^T$ and $\boldsymbol{x}^{(2)} = [2,4]^T$, and two negative sample $\boldsymbol{x}^{(3)} = [0,0]^T$ and $\boldsymbol{x}^{(4)} = [-1,0]^T$. Please calculate the SVM decision hyperplane