Unit 13 - Week 10 - Structural Risk Minimization and Bias-Variance Dilemma

Course outline

Assignment 10

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2019-10-09, 23:59 IST.

Instructions:

1. **Attempt all questions.**
2. **Submission deadline**: 9th October 2019 23:59 IST
3. **Solutions to be posted**: 10th October 2019
4. Older browsers might show **unnecessary vertical bars** at the end of math equations.

1) Consider a regression data generated by a polynomial of degree 3. Then, the estimates of the linear regression on the data with respect to the true model will have *2 points*

- Low bias and high variance.
- Low bias and low variance.
- High bias and high variance.
- High bias and low variance.

No, the answer is incorrect.
Score: 0
Accepted Answers:
High bias and low variance.

2) In continuation with question 1, estimates of the polynomial of degree 3 on the data with respect to the true model will have *2 points*

- Low bias and high variance.
- Low bias and low variance.
- High bias and high variance.
- High bias and low variance.

No, the answer is incorrect.
Score: 0
Accepted Answers:
Low bias and low variance.

3) In continuation with question 1, estimates of the polynomial of degree 10 on the data with respect to the true model will have *2 points*

- Low bias and high variance.
- Low bias and low variance.
- High bias and high variance.
- High bias and low variance.

No, the answer is incorrect.
Score: 0
Accepted Answers:
Low bias and high variance.

4) Let $Y = f(x) + \epsilon$, where ϵ has a zero mean and variance σ^2. In K-nearest neighbor (KNN) regression, the prediction of Y at point x_0 is given by the average of the values Y at the k neighbors closest to x_0. Denote the l-nearest neighbor to x_0 by x_l and its corresponding Y value by y_l. Then the prediction $\hat{f}(x_0)$ of the KNN regression for x_0 in terms of y_l, $1 \le l \le k$ is given by $\hat{f}(x_0) = \frac{1}{k}\sum_{l=1}^{k} y_l$. Then, what is the behavior of the bias as k increases? *2 points*

- Increase.
- Decrease.
- Oscillates.
- Converges to a finite value.

No, the answer is incorrect.
Score: 0
Accepted Answers:
Increase.

5) In continuation with question 4, what is the behavior of the variance as k increases? *2 points*

- Increase.
- Decrease.
- Oscillates.
- Converges to a finite value.

No, the answer is incorrect.
Score: 0
Accepted Answers:
Decrease.

6) Given the training set $\{x_i, d_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^n$. Consider a regularization network with a hidden layer of m basis functions i.e., $\phi(x, t_j) = \mathrm{G}\left(x, t_j\right)$ where $j = 1, 2 \ldots m$. Then the approximated function $\hat{F}(x) = \sum_{i=1}^{m} w_i G\left(x, t_i\right)$. The weight vector $\mathbf{w}$ is determined by approximating the cost functional defined as *1 point*

$$\mathrm{E}(\hat{F}) = \|\mathbf{d} - \mathbf{G}\mathbf{w}\|^2 + \lambda\|\mathbf{D}\hat{F}\|^2$$

(True/False) Then $\mathbf{G}$ is a positive semi-definite matrix.

- True
- False

No, the answer is incorrect.
Score: 0
Accepted Answers:
False

7) In continuation with question 6, consider the following statements: *3 points*

a) The regularization network is a universal approximator, that can approximate any multivariate continuous function.
b) As the regularization parameter λ approaches to zero, the optimized weight vector converges to the pseudoinverse solution.
c) The term $\|\mathbf{D}\hat{F}\|^2$ denotes the smoothness functional.
d) The regularization parameter λ is estimated by either minimizing the averaged squared error or using generalized cross-validation.
Which of the above statements are correct?

- Only a and c
- Only b and d
- Only a,b and c
- a, b, c and d

No, the answer is incorrect.
Score: 0
Accepted Answers:
a, b, c and d

8) In continuation with question 6, by using the properties of adjoint of the differential operator $\mathbf{D}$ and Green's function, we get $\|\mathbf{D}\hat{F}\|^2 = \mathbf{w}^{\mathrm{T}}\mathbf{G}_0\mathbf{w}$. Choose the correct options from below. *1.5 points*

- $\mathbf{G}$ is a rectangular matrix of size $N \times m$.
- $\mathbf{G}_0$ is a square matrix of size $N \times N$.
- $\mathbf{G}_0$ is a symmetric matrix of size $m \times m$.
- $\mathbf{G}_0$ is a symmetric matrix of size $N \times N$.

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\mathbf{G}_0$ is a symmetric matrix of size $m \times m$.

9) In continuation with question 6, assume that $\lambda = 1$. Then the weight vector $\mathbf{w}$ determined by approximating the cost functional is given as *1.5 points*

- $\mathbf{w} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G} + \mathbf{G}\right)^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{d}$
- $\mathbf{w} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G} + \mathbf{G}_0\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}$
- $\mathbf{w} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G}_0\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}$
- $\mathbf{w} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G}_0 + \mathbf{G}_0\right)^{\mathrm{T}}\mathbf{G}_0^{\mathrm{T}}\mathbf{d}$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\mathbf{w} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G} + \mathbf{G}_0\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}$

10) Consider a nonlinear regression problem, described by a model as *2 points*

$$d_i = f(x_i) + \epsilon_i \quad i = 1, 2 \ldots N$$

Assume that $\sigma^2 = 1$ and $\mathrm{tr}\left[\mathbf{A}^2(\lambda)\right] = \mathrm{tr}\left[(\mathbf{I} - \mathbf{A}(\lambda))^2\right]$, where A is the influence matrix. Then the estimate for average squared error $\hat{R}(\lambda) = \mathbb{E}\left[R(\lambda)\right]$ is given as

- $\hat{R}(\lambda) = \frac{1}{N}\|(\mathbf{I} - \mathbf{A}(\lambda)\mathbf{d})\|^2 - \frac{1}{N}\mathrm{tr}\left[(\mathbf{I} - \mathbf{A}(\lambda))^2\right]$
- $\hat{R}(\lambda) = \frac{1}{N}\|(\mathbf{I} - \mathbf{A}(\lambda)\mathbf{d})\|^2$
- $\hat{R}(\lambda) = \frac{1}{N}\|(\mathbf{I} - \mathbf{A}(\lambda)\mathbf{d})\|^2 + \frac{2}{N}\mathrm{tr}\left[\mathbf{A}^2(\lambda)\right]$
- $\hat{R}(\lambda) = \frac{1}{2N}\|(\mathbf{I} - \mathbf{A}(\lambda)\mathbf{d})\|^2$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\hat{R}(\lambda) = \frac{1}{N}\|(\mathbf{I} - \mathbf{A}(\lambda)\mathbf{d})\|^2$

11) (True/False) In continuation with question 10, the minimizer of the estimate $\hat{R}(\lambda)$ can be taken as a good choice for the regularization parameter λ. *1 point*

- True
- False

No, the answer is incorrect.
Score: 0
Accepted Answers:
True