CITY UNIVERSITY OF HONG KONG (DONGGUAN)

Course code & title:		SDSC 5001 Statistical Machine Learning I
Session	:	Semester A 2025/26
Exam	:	Midterm
Time allowed	:	2 hours

This paper has 6 pages (including this cover page).

---

1. Please fill in your information below:

   Student Name: _____ ID Number:_____

2. Answer all questions in the four sections on the exam paper.

3. Total points = 100

This is a **closed-book** examination.

Students are allowed to use the following materials/aids:

- One notes sheet (A4 size, double-sided)
- Approved calculator

Materials/aids other than those stated above are not permitted. Students will be subject to disciplinary action if any unauthorized materials or aids are found on them.

Section A

(15 questions, each 3 points, total 45 points)

Please read each statement and determine if it is true or false (**circle one**).

1	True	False	An event is a subset of outcomes in sample space of a population.
2	True	False	Ridge regression is a linear regression with L-1 norm.
3	True	False	In statistical machine learning, model overfitting can be identified via checking training error curve and validation error curve.
4	True	False	Roll a fair die repeatedly. The average number of rolls required until we see an even number is 6.
5	True	False	A weather dataset provides rainfall amount measured during the 1-hour period ending at the observation time. It is a nominal variable.
6	True	False	We should always remove outliers in a training dataset to improve prediction performance.
7	True	False	A pair of fair dice is rolled. The probability that the second die lands on a higher value than the first is 5/12.
8	True	False	R^2 is not a good measure for model comparison in linear regression as it always increases with more predictors in the model.
9	True	False	Given a dataset of *P* attributes and *N* number of records, feature selection is applied to realize a simpler model by reducing *N*.
10	True	False	The sample mean is an unbiased estimator of the population mean because the expected value of the sample mean is equal to the population mean.
11	True	False	Boxplot can be applied as a categorical descriptive statistic of a random variable.
12	True	False	The mean of a uniform distribution, Unif(a, b), is (a + b)/2.
13	True	False	The k nearest neighbors algorithm can do both classification and regression.
14	True	False	In multiple linear regression with p predictors and n observations, the dimension of the matrix $\mathbf{X}$ is $n \times (p + 1)$.
15	True	False	A statistic is a random variable.

Section B

(15 questions, each 3 points, total 45 points)

Each multiple-choice question is followed by several suggested answers. Select the **single best** one for the question. **Put all your answers in the table below.**

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15

1. Which of the following is *not* a data quality issue?
A. Sampling bias
B. Outlier
C. Happy data
D. Missing data

2. Given a random sample $X_1, X_2, \dots, X_n$ of a population, which of the following is the assumption of Central limit theorem?
A. $X_1, X_2, \dots, X_n$ are i.i.d. with mean μ and variance σ^2
B. $X_1, X_2, \dots, X_n$ are i.i.d., following normal distribution with mean μ and variance σ^2
C. $X_1, X_2, \dots, X_n$ are independent with mean μ_i and variance σ_i^2, $i = 1, \dots, n$
D. $X_1, X_2, \dots, X_n$ are normally distributed with mean μ_i and variance σ_i^2, $i = 1, \dots, n$

3. Which of the following is *not* a summary statistic for location?
A. Percentile
B. Mean
C. Median
D. Skewness

4. Which of the following statements about point estimation in statistical inference is *not* true?
A. The minimum variance unbiased estimator (MVUE) may exist in some cases, but in general it is difficult to find.
B. The method of moments (MM) estimator is obtained by equating population moments to the corresponding sample moments.
C. The maximum likelihood estimator (MLE) is the value of the parameter that maximizes the likelihood function.
D. For an i.i.d. sample from a normal distribution, the minimum variance unbiased estimator (MVUE) of the population mean does not exist.

5. In the following pairs of terms used in different fields, which one is different from others?
A. covariates vs. features

B. responses vs. outputs
C. gradient vs. learner
D. instance vs. data point

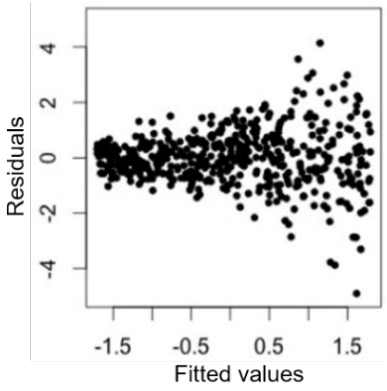
6. Given a dataset with a binary response and n observations, the complexity (i.e., effective number of parameters) of the KNN classifier is determined by:
A. n
B. n/k
C. k
D. k/n

7. Which of the following statements about the prediction performance of a flexible learning method vs. an inflexible learning method is *not* true?
A. When the variance of the error term is extremely high, the flexible method tends to perform worse than the inflexible method.
B. When the sample size is extremely large and the number of predictors is small, the flexible method tends to perform better than the inflexible method.
C. When the relationship between the predictors and response is highly non-linear, the flexible method tends to perform worse than the inflexible method.
D. When the number of predictors is extremely large and the number of observations is small, the flexible method tends to perform worse than the inflexible method.

8. Which of the following statements about LOOCV and K-fold CV is true?
A. LOOCV has larger bias but smaller variance than K-fold CV.
B. LOOCV has smaller bias but larger variance than K-fold CV.
C. LOOCV has smaller bias and smaller variance than K-fold CV.
D. LOOCV has larger bias and larger variance than K-fold CV.

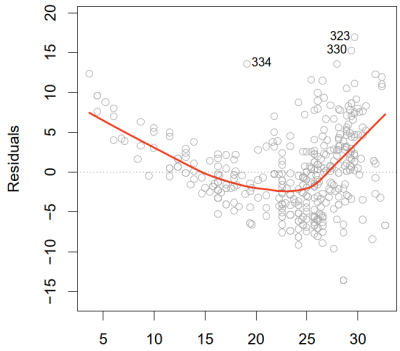
9. Which of the following is *not* a concern in a supervised learning study for inference?
A. The distribution of each predictor and the response
B. The relationship between the response and each predictor
C. How the response changes when a predictor changes.
D. The predictors that have significant effect on the response

10. The residual plot produced in a linear regression study is shown below



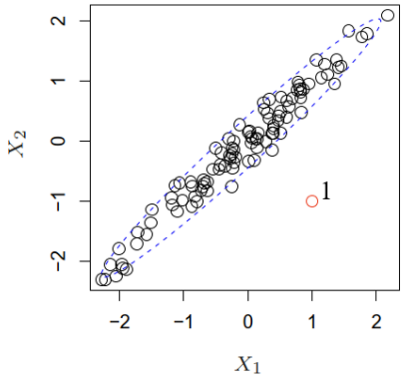
Which problem does the data have based on the residual plot?
A. Outliers
B. Nonnormality
C. Nonlinearity
D. Heteroscedasticity

11. The residual plot produced in a linear regression study is shown below



Which problem does the data have based on the residual plot?
A. Autocorrelation
B. Homogeneity
C. Nonlinearity
D. Heterogeneity

12. Which of the following statements about Point 1 in the figure is *not* true?



A. It is an outlying x observation.
B. It is not an outlying observation since it does not have unusual x or y value.
C. It has a considerably higher leverage value than other observations.
D. Its residual may be small and like the residuals of other observations.

13. Among the following statements about the parameter estimate $\hat{\beta}_1$ in simple linear regression, which one is *not* true?
A. $\hat{\beta}_1$ is a random variable.
B. $\hat{\beta}_1$ is the BLUE estimator of β_1.
C. The accuracy of $\hat{\beta}_1$ is affected by random error variance σ^2. When σ^2 is larger, $\hat{\beta}_1$ is more accurate.

D. The distribution of $\hat{\beta}_1$ refers to its different values resulted from repeated sampling when predictor values are held constant from sample to sample.

14. In multiple linear regression with p predictors and n observations, the dimension of the matrix $\mathbf{X}$ and the least squares estimate of coefficients are:
A. $\mathbf{X}$ is $n \times p$, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y}$
B. $\mathbf{X}$ is $n \times (p+1)$, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y}$
C. $\mathbf{X}$ is $n \times (p+1)$, $\hat{\beta} = (\mathbf{X}\mathbf{X}^T)^{-1}\,\mathbf{X}^T\mathbf{y}$
D. $\mathbf{X}$ is $n \times (p+1)$, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T$

15. Assume three variables X_1, X_2, and Y have the following relationship:

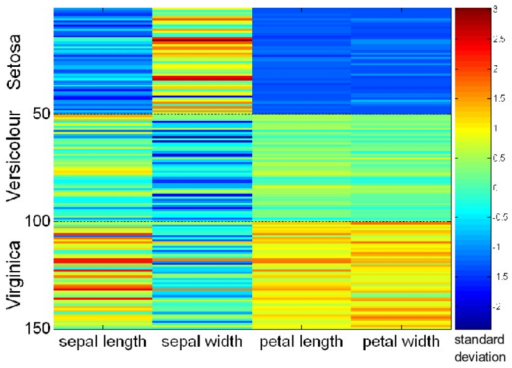
$$X_1 \sim N(0,1)$$
$$X_2 = 0.5X_1 + \epsilon/10 \quad \text{where } \epsilon \sim N(0,1)$$
$$Y = 2 + 2X_1 + 0.5X_2 + e \quad \text{where } e \sim N(0,1)$$

A student simulates 100 observations of the variables, and using the data, fits a linear regression model to predict Y using X_1 and X_2. Which of the following is the most possible results of p-values of X_1 and X_2 (from t tests) in the output?
A. p-value(X_1) < 0.0001, p-value(X_2) = 0.364
B. p-value(X_1) = 0.245, p-value(X_2) = 0.001
C. p-value(X_1) = 0.631, p-value(X_2) = 0.565
D. p-value(X_1) < 0.0001, p-value(X_2) < 0.0001

Section C

(2 questions, 10 points)

Please make drawing in the given figures following the requirements in each question.
1. Given the figure below, please describe the name of the visualization technique _____ (5 points).



2. In linear regression with a large number of predictors, should we use the adjusted R^2 to assess model fitting performance? Yes __ or No __ (5 points)

-END-