Perplexity Mock exam

Linear Regression Mock Exam for First Year Bachelor Students

This comprehensive mock exam covers fundamental concepts of linear regression for first-year bachelor students in Machine Learning. The exam consists of 40 multiple-choice questions across five key areas: basic concepts, mathematical formulation, bias-variance tradeoff, regularization, and practical applications. Each question includes citations to academic and industry sources.

Basic Concepts and Definitions

Question 1

What is linear regression?

- a) A classification algorithm
- b) An algorithm that provides a linear relationship between variables to predict outcomes
- c) A clustering technique
- d) A dimensionality reduction method

Question 2

In linear regression, which of the following represents the dependent variable?

- a) The variable that remains unchanged
- b) The predictor variable
- c) The variable that changes with fluctuations in other variables
- d) The explanatory variable

Question 3

What does the "best fit line" in linear regression represent?

- a) The line with the highest slope
- b) The line that passes through the most data points
- c) The line that suitably fits all data points by minimizing error
- d) The line that connects the first and last data points

Question 4

Linear regression is classified as what type of machine learning?

- a) Unsupervised learning
- b) Supervised learning
- c) Reinforcement learning
- d) Semi-supervised learning

Question 5

What types of variables can linear regression predict?

- a) Categorical variables only
- b) Binary variables only
- c) Continuous or numeric variables
- d) Ordinal variables only

Question 6

What is the most basic form of the linear regression equation?

- a) $y = c + b*x$
- b) $y = a + bx + cx^2$
- c) $y = ax/b + c$
- d) $y = a*e^{(bx)}$

Question 7

In simple linear regression, how many independent variables are used?

- a) Zero
- b) One
- c) Two
- d) Many

Question 8

Multiple linear regression differs from simple linear regression in that it:

- a) Has multiple dependent variables
- b) Has multiple independent variables
- c) Uses non-linear equations
- d) Requires categorical variables

Question 9

Which of the following is NOT a common term for the dependent variable in regression?

- a) Outcome variable
- b) Criterion variable
- c) Predictor variable
- d) Regressand

Question 10

What graphical representation is commonly used for linear regression?

- a) Pie chart
- b) Bar graph
- c) Scatter plot with a straight line
- d) Box plot

Mathematical Formulation

Question 11

In the equation y = c + b*x for simple linear regression, what does 'c' represent?

- a) The correlation coefficient
- b) The y-intercept
- c) The slope of the line
- d) The coefficient of determination

Question 12

In linear regression, what method is commonly used to find the best fit line?

- a) Maximum likelihood estimation
- b) Gradient descent
- c) Ordinary least squares
- d) All of the above

Question 13

What is being minimized in the ordinary least squares (OLS) method?

- a) The sum of absolute errors

- b) The sum of squared errors
- c) The maximum error
- d) The correlation coefficient

Question 14

In the context of L2-regularized linear regression, what is added to the cost function?

- a) λ/2 times the L1 norm of θ
- b) λ/2 times the L2 norm squared of θ
- c) λ times the number of features
- d) λ times the maximum value of θ

Question 15

What does the regularization parameter λ control in L2-regularized linear regression?

- a) The learning rate
- b) The trade-off between fitting the data and keeping the model simple
- c) The number of iterations
- d) The number of features

Question 16

The closed-form solution for ordinary least squares linear regression can be written as:

- a) $\hat{\theta} = (X^T X)^{-1} X^T Y$
- b) $\hat{\theta} = X^T (X X^T)^{-1} Y$
- c) $\hat{\theta} = (X^T X) X^T Y$
- d) $\hat{\theta} = X^T Y$

Question 17

In the context of linear regression, what does the term "collinearity" refer to?

- a) When the dependent variable is linearly related to the independent variable
- b) When two or more independent variables are highly correlated
- c) When the residuals are normally distributed
- d) When the model is linear in parameters

Question 18

What is the mathematical definition of the mean squared error (MSE) in linear regression?

- a) $MSE = 1/n * \Sigma(y_i - \hat{y}_i)$
- b) $MSE = 1/n * \Sigma|y_i - \hat{y}_i|$
- c) $MSE = 1/n * \Sigma(y_i - \hat{y}_i)^2$
- d) $MSE = \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$

Bias-Variance Tradeoff

Question 19

What is the bias-variance decomposition of the expected squared prediction error?

- a) $Error = Noise + Bias^2 + Variance$
- b) $Error = Noise + Bias + Variance$
- c) $Error = Bias^2 * Variance$
- d) $Error = (Bias + Variance)^2$

Question 20

High bias in a model typically leads to:

- a) Overfitting
- b) Underfitting
- c) Perfect fit
- d) Random predictions

Question 21

High variance in a model typically leads to:

- a) Overfitting
- b) Underfitting
- c) Constant predictions
- d) Negative correlation

Question 22

In the bias-variance tradeoff, what happens to the variance as model complexity increases?

- a) Variance decreases

- b) Variance increases
- c) Variance remains constant
- d) Variance becomes negative

Question 23

In the bias-variance tradeoff, what happens to the bias as model complexity increases?

- a) Bias decreases
- b) Bias increases
- c) Bias remains constant
- d) Bias becomes negative

Question 24

What is the irreducible error in the bias-variance decomposition?

- a) The error due to the bias of the model
- b) The error due to the variance of the model
- c) The error due to noise in the data that no model can eliminate
- d) The error due to inappropriate model selection

Question 25

In the context of linear regression with L2-regularization, increasing the regularization parameter λ will:

- a) Increase bias and increase variance
- b) Increase bias and decrease variance
- c) Decrease bias and increase variance
- d) Decrease bias and decrease variance

Question 26

How can you determine if your linear regression model has high bias?

- a) High training error
- b) High test error but low training error
- c) Low training and test error
- d) Equal training and test error

Regularization

Question 27

What is the primary purpose of regularization in linear regression?

- a) To increase model complexity
- b) To prevent overfitting
- c) To speed up computation
- d) To handle missing values

Question 28

L2 regularization (Ridge regression) adds which term to the cost function?

- a) $\lambda * ||\theta||_1$ (L1 norm)
- b) $\lambda * ||\theta||_2^2$ (squared L2 norm)
- c) $\lambda * \max(|\theta|)$
- d) $\lambda * \log(||\theta||)$

Question 29

What effect does L2 regularization have on the coefficients of a linear regression model?

- a) It forces many coefficients to be exactly zero
- b) It shrinks all coefficients toward zero
- c) It increases the magnitude of all coefficients
- d) It has no effect on the coefficients

Question 30

What is the optimal way to select the regularization parameter λ in practice?

- a) Always use $\lambda = 1$
- b) Set λ as small as possible
- c) Use cross-validation to find the λ that minimizes generalization error
- d) Set λ equal to the number of features

Question 31

What is the relationship between regularization and the bias-variance tradeoff?

- a) Regularization only affects bias, not variance
- b) Regularization only affects variance, not bias
- c) Regularization helps balance bias and variance
- d) Regularization increases both bias and variance

Question 32

In terms of the eigenvalues of X^T X, what does L2 regularization effectively do?

- a) It adds λ to each eigenvalue
- b) It subtracts λ from each eigenvalue
- c) It multiplies each eigenvalue by λ
- d) It divides each eigenvalue by λ

Applications and Interpretation

Question 33

Which of the following is a common application of linear regression?

- a) Image classification
- b) Sales forecasting
- c) Clustering customer segments
- d) Language translation

Question 34

In finance, linear regression can be used for:

- a) Facial recognition
- b) Asset valuation
- c) Speech recognition
- d) Image segmentation

Question 35

What does the coefficient of determination (R^2) measure in linear regression?

- a) The slope of the regression line
- b) The proportion of variance in the dependent variable explained by the independent variables

- c) The y-intercept of the regression line
- d) The average prediction error

Question 36

When interpreting a linear regression model, what does a positive coefficient for a predictor variable indicate?

- a) The dependent variable decreases as the predictor increases
- b) The dependent variable increases as the predictor increases
- c) The predictor is not statistically significant
- d) The predictor is correlated with other predictors

Question 37

What assumption about the residuals is typically made in linear regression?

- a) Residuals should be correlated with the predictors
- b) Residuals should have non-constant variance
- c) Residuals should be normally distributed with mean zero
- d) Residuals should be large in magnitude

Question 38

When would using multiple linear regression be more appropriate than simple linear regression?

- a) When we have a very small dataset
- b) When we want to analyze the effect of multiple variables on the outcome
- c) When computational resources are limited
- d) When the relationship is clearly non-linear

Question 39

What does linear regression NOT tell us about the relationship between variables?

- a) The strength of the relationship
- b) The direction of the relationship
- c) Causation between variables
- d) The mathematical form of the relationship

Question 40

In what case would linear regression be an inappropriate model choice?

- a) When the relationship between variables is linear
- b) When predicting continuous outcomes
- c) When the relationship between variables is highly non-linear
- d) When analyzing the impact of multiple predictors

Answer Key with Sources

1. b - An algorithm that provides a linear relationship between variables to predict outcomes17
2. c - The variable that changes with fluctuations in other variables17
3. c - The line that suitably fits all data points by minimizing error17
4. b - Supervised learning17
5. c - Continuous or numeric variables17
6. a - y = c + b*x18
7. b - One1820
8. b - Has multiple independent variables1820
9. c - Predictor variable18
10. c - Scatter plot with a straight line1720
11. b - The y-intercept1820
12. d - All of the above1620
13. b - The sum of squared errors16
14. b - λ/2 times the L2 norm squared of θ16
15. b - The trade-off between fitting the data and keeping the model simple1619
16. a - $\theta \hat{=} (X^T X)^{-1} X^T Y$1618
17. b - When two or more independent variables are highly correlated20
18. c - $MSE = 1/n * \Sigma(y_i - \hat{y}_i)^2$16
19. a - $Error = Noise + Bias^2 + Variance$1719
20. b - Underfitting161719
21. a - Overfitting161719

22. b - Variance increases[17][19]
23. a - Bias decreases[17][19]
24. c - The error due to noise in the data that no model can eliminate[16]
25. b - Increase bias and decrease variance[16]
26. a - High training error[16]
27. b - To prevent overfitting[16][19]
28. b - $\lambda * ||\theta||_2^2$ (squared L2 norm)[16]
29. b - It shrinks all coefficients toward zero[16][19]
30. c - Use cross-validation to find the λ that minimizes generalization error[16]
31. c - Regularization helps balance bias and variance[16][19]
32. a - It adds λ to each eigenvalue[16]
33. b - Sales forecasting[17][20]
34. b - Asset valuation[20]
35. b - The proportion of variance in the dependent variable explained by the independent variables[20]
36. b - The dependent variable increases as the predictor increases[17][18]
37. c - Residuals should be normally distributed with mean zero[16][20]
38. b - When we want to analyze the effect of multiple variables on the outcome[18][20]
39. c - Causation between variables[20]
40. c - When the relationship between variables is highly non-linear[17][20]

Conclusion

This mock exam covers the fundamental concepts of linear regression that first-year bachelor students should understand. By mastering these concepts, students will develop a solid foundation in one of the most important predictive modeling techniques in machine learning. The exam tests understanding of basic principles, mathematical formulations, the important bias-variance tradeoff, regularization techniques, and practical applications of linear regression.

For further study, students should practice implementing these concepts in programming environments and explore more advanced regression techniques that build upon these foundational principles.