Mock Test 2 - Data Science - PPT - PW Skills

Section 1: MCQs
Section 2: Coding Questions

* Indicates required question

Email *

Your email

Name *

Your answer

Phone Number *

Your answer

1. Which analysis examines a single variable and describes its characteristics? * 2 points

- a) Univariate analysis
- b) Bivariate analysis
- c) Multivariate analysis
- d) Inferential statistics

2. What is the measure of central tendency that represents the middle value in a dataset? * 2 points

- a) Mean
- b) Mode
- c) Median
- d) Variance

3. Which method measures the spread or variability of a dataset? * 2 points

- a) Central tendency
- b) Methods of dispersion
- c) Bassel's correction
- d) Probability theory

4. What does Bassel's correction account for in statistics? * 2 points

- a) Skewness in the dataset
- b) Kurtosis in the dataset
- c) Sample size in relation to population size
- d) Outliers in the dataset

5. Which of the following is NOT part of the five-number summary? * 2 points

- a) Minimum
- b) Lower quartile (Q1)
- c) Median
- d) Outliers

6. Which branch of mathematics deals with the study of uncertainty and random events? * 2 points

- a) Probability theory
- b) Distribution of the data
- c) Types of sampling
- d) Inferential statistics

7. Which test is used to determine if a sample comes from a particular probability distribution? * 2 points

- a) Hypothesis testing
- b) Z-test
- c) Chi-square test
- d) Anova test (F-ratio)

8. What is the significance level in hypothesis testing? * 2 points

- a) The probability of making a Type I error
- b) The probability of making a Type II error
- c) The level of confidence in the test result
- d) The level of significance of the null hypothesis

9. Which statistical test is used to compare the means of two independent groups? * 2 points

- a) Z-test
- b) T-test
- c) Chi-square test
- d) Anova test (F-ratio)

10. What does correlation measure? * 2 points

- a) The strength of the linear relationship between two variables
- b) The difference between two sample means
- c) The spread or variability within a dataset
- d) The association between categorical variables

11. What is the purpose of inferential statistics? * 2 points

- a) To summarize and describe data
- b) To make predictions based on data
- c) To analyze relationships between variables
- d) To visualize data using graphs and charts

12. What is the p-value in hypothesis testing? * 2 points

- a) The probability of observing the given sample data, assuming the null hypothesis is true
- b) The probability of rejecting the null hypothesis when it is actually true
- c) The probability of accepting the null hypothesis when it is actually false
- d) The probability of a Type I error

13. What is a confidence interval? * 2 points

- a) A range of values within which the true population parameter is likely to fall
- b) The margin of error in a hypothesis test
- c) The level of significance chosen for a statistical test
- d) The probability of rejecting the null hypothesis

14. What does the term "standard error" represent? * 2 points

- a) The average value in a dataset
- b) The measure of spread in a dataset
- c) The standard deviation of the population
- d) The standard deviation of the sampling distribution

15. What is the degree of freedom in statistics? * 2 points

- a) The number of data points in a dataset
- b) The number of parameters estimated in a statistical model
- c) The number of variables in a dataset
- d) The number of categories in a categorical variable

16. Which statistical test is used to compare two population proportions? * 2 points

- a) Z-test
- b) T-test
- c) Chi-square test
- d) Anova test (F-ratio)

17. What is the purpose of A/B testing? * 2 points

- a) To compare two different populations using a statistical test
- b) To determine the correlation between two variables
- c) To measure the effectiveness of different strategies or interventions
- d) To calculate the mean and standard deviation of a dataset

18. What does covariance measure? * 2 points

- a) The strength of the linear relationship between two variables
- b) The difference between two sample means
- c) The spread or variability within a dataset
- d) The joint variability of two variables

19. How can missing values be handled in a dataset? * 2 points

- a) Removing the entire row with missing values
- b) Replacing missing values with the mean or median of the variable
- c) Using statistical imputation methods to estimate missing values
- d) All of the above

20. Which type of sampling technique ensures that each member of the population has an equal chance of being included in the sample? * 2 points

- a) Simple random sampling
- b) Stratified sampling
- c) Cluster sampling
- d) Systematic sampling

21. Which statistical model is used to analyze the relationship between a dependent variable and one or more independent variables? * 2 points

- a) General Linear Model
- b) Support Vector Machine (SVM)
- c) Decision Tree (DT)
- d) Ensemble Techniques

22. What is the purpose of regression analysis? * 2 points

- a) To classify data into different groups
- b) To measure the strength of association between variables
- c) To reduce the dimensionality of the dataset
- d) To predict a continuous dependent variable based on independent variables

23. What does the loss function measure in machine learning? * 2 points

- a) The accuracy of the model's predictions
- b) The amount of error between the model's predictions and the actual values
- c) The complexity of the model
- d) The regularization term in the model

24. Which optimization algorithm is commonly used in machine learning to minimize the loss function? * 2 points

- a) Gradient Descent (GD)
- b) Stochastic Gradient Descent (SGD)
- c) AdaBoost
- d) Random Forest

25. What is regularization in machine learning? * 2 points

- a) The process of reducing bias in a model
- b) The process of reducing variance in a model
- c) The process of adding penalties to the model to prevent overfitting
- d) The process of removing outliers from the dataset

26. Which algorithm is primarily used for classification and regression tasks, based on finding an optimal hyperplane that separates the data into different classes or predicts continuous values? * 2 points

- a) General Linear Model
- b) Support Vector Machine (SVM)
- c) Decision Tree (DT)
- d) Ensemble Techniques

27. What does DT stand for in machine learning? * 2 points

- a) Decision Theory
- b) Decision Task
- c) Decision Table
- d) Decision Tree

28. What is the main idea behind ensemble techniques in machine learning? * 2 points

- a) Combining multiple weak models to create a strong model
- b) Creating an ensemble of identical models for increased accuracy
- c) Using bagging to reduce overfitting in the model
- d) Boosting the performance of a single model through feature selection

29. Which term represents the equation that describes the relationship between the independent variables and the dependent variable in a linear model? * 2 points

- a) Loss function
- b) Optimization function
- c) Regression equation
- d) Regularization equation

30. What is the purpose of the loss function in a machine learning model? * 2 points

- a) To measure the accuracy of the model's predictions
- b) To quantify the amount of regularization applied to the model
- c) To define the objective that the model seeks to minimize during training
- d) To evaluate the impact of outliers on the model's performance

31. Which optimization algorithm is an extension of gradient descent that uses a randomly selected subset of training examples in each iteration? * 2 points

- a) Gradient Descent (GD)
- b) Stochastic Gradient Descent (SGD)
- c) Adaptive Moment Estimation (Adam)
- d) Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)

32. What is the purpose of regularization in machine learning? * 2 points

- a) To prevent underfitting by increasing the complexity of the model
- b) To reduce overfitting by adding a penalty term to the loss function
- c) To improve the interpretability of the model's coefficients
- d) To speed up the training process by reducing the number of iterations

33. Which machine learning algorithm can be used for both classification and regression tasks and is based on finding the optimal splitting points in a hierarchical structure? * 2 points

- a) General Linear Model
- b) Support Vector Machine (SVM)
- c) Decision Tree (DT)
- d) Ensemble Techniques

34. What is the main idea behind ensemble techniques in machine learning? * 2 points

- a) Combining multiple weak models to create a strong model
- b) Creating an ensemble of identical models for increased accuracy
- c) Using bagging to reduce overfitting in the model
- d) Boosting the performance of a single model through feature selection

35. Which ensemble technique creates multiple subsets of the original dataset through random sampling with replacement and trains separate models on each subset? * 2 points

- a) Bagging
- b) Boosting
- c) Stacking
- d) Voting

36. Which ensemble technique combines predictions from multiple models by assigning weights to each model's prediction based on their performance? * 2 points

- a) Bagging
- b) Boosting
- c) Stacking
- d) Voting

37. Which ensemble technique builds a sequence of models, where each subsequent model focuses on improving the mistakes made by the previous models? * 2 points

- a) Bagging
- b) Boosting
- c) Stacking
- d) Voting

38. What is the primary advantage of support vector machines (SVM) over other machine learning algorithms? * 2 points

- a) Ability to handle missing values in the dataset
- b) Flexibility in handling large-scale datasets
- c) High interpretability of the model's coefficients
- d) Ability to handle nonlinear relationships between variables

39. You have a dataset containing the exam scores of students and their corresponding hours of study. You want to build a linear regression model to predict a student's score based on their study hours. Given the dataset as a built-in dataset called 'load_exam_data' in scikit-learn, complete the code snippet below to fit a linear regression model using the General Linear Model in scikit-learn and make a prediction for a student who studied for 5 hours. * 2 points

```
from sklearn.datasets import load_exam_data

from sklearn.linear_model import LinearRegression


# Load the dataset

data = load_exam_data()


# Separate the features (study hours) and target variable (scores)

X = data.data.reshape(-1, 1)

y = data.target


# Initialize a linear regression model

model = LinearRegression()


# Fit the model to the data

# Your code goes here


# Make a prediction for a student who studied for 5 hours

prediction = model.predict([[5]])

print("Predicted score for 5 hours of study:", prediction[0])
```

- ○ a) model.fit(X, y)

- b) model.predict(X, y)
- c) model.transform(X, y)
- d) model.score(X, y)

40. You have a dataset containing information about housing prices, including features such as the number of bedrooms, square footage, and location. You want to build a regression model using the Gradient Descent optimizer to predict the price of a house based on its features. Given the dataset as a built-in dataset called 'load_boston' in scikit-learn, complete the code snippet below to fit a linear regression model using Gradient Descent in scikit-learn. * 2 points

```
from sklearn.datasets import load_boston

from sklearn.linear_model import SGDRegressor


# Load the dataset

data = load_boston()


# Separate the features and target variable

X = data.data

y = data.target


# Initialize a linear regression model with Gradient Descent optimizer

model = SGDRegressor(loss='squared_loss', max_iter=1000,
random_state=42)


# Fit the model to the data

# Your code goes here
```

- a) model.fit(X, y)
- b) model.predict(X, y)
- c) model.transform(X, y)

d) model.score(X, y)

41. You have a dataset containing information about various flowers, including sepal length, sepal width, petal length, and petal width. You want to use the Naive Bayes classifier to classify a flower's species based on these features. Given the dataset as a built-in dataset called 'load_iris' in scikit-learn, complete the code snippet below to train a Naive Bayes classifier and make a prediction for a flower with sepal length 5.0, sepal width 3.6, petal length 1.4, and petal width 0.2. * 2 points

```
from sklearn.datasets import load_iris

from sklearn.naive_bayes import GaussianNB


# Load the dataset

data = load_iris()


# Separate the features and target variable

X = data.data

y = data.target


# Initialize a Naive Bayes classifier

model = GaussianNB()


# Fit the model to the data

# Your code goes here


# Make a prediction for a flower with specific measurements

prediction = model.predict([[5.0, 3.6, 1.4, 0.2]])

print("Predicted species:", data.target_names[prediction[0]])
```

- ○ a) model.fit(X, y)
- ○

b) model.predict(X, y)

c) model.transform(X, y)

d) model.score(X, y)

42. You have a dataset containing customer information, including age, income, and spending score. You want to use the K-nearest neighbors (KNN) algorithm to classify customers into different groups based on their characteristics. Given the dataset as a built-in dataset called 'load_customers' in scikit-learn, complete the code snippet below to train a KNN classifier with k=5 and make a prediction for a new customer with age 30, income 50, and spending score 70. * 2 points

```
from sklearn.datasets import load_customers

from sklearn.neighbors import KNeighborsClassifier


# Load the dataset

data = load_customers()


# Separate the features and target variable

X = data.data

y = data.target


# Initialize a KNN classifier with k=5

model = KNeighborsClassifier(n_neighbors=5)


# Fit the model to the data

# Your code goes here


# Make a prediction for a new customer

prediction = model.predict([[30, 50, 70]])

print("Predicted group:", data.target_names[prediction[0]])
```

- ○ a) model.fit(X, y)
- ○

b) model.predict(X, y)

c) model.transform(X, y)

d) model.score(X, y)

Submit Clear form