Name: _____ Enrolment no.: _____

Programme: _____

CSE548 Applied Data Science: End Semester Exam (Winter Sem.: 2024-25);
Marks: 80 Marks, Wt.: 40%
April 28, 2025, Monday, 11 AM – 1 PM

Instructions:

1 Fill up details about your enrolment no and name on the question paper.
2 Write answers in the question paper itself and submit it before the end time.
3 There is no negative marking. Each question carry one mark.
4 Only one option is correct in each question. Total 80 MCQs are there in the paper.

1. Which of the following dimensionality reduction techniques is best known for preserving local structures in high-dimensional data when visualizing it?

a) PCA b) t-SNE c) UMAP d) Isomap

Answer: _____

2. UMAP is commonly used for visualization because:

a) It is a linear method for dimensionality reduction
b) It captures both local and global structures of the data
c) It can only handle categorical data
d) It requires labeled data

Answer: _____

3. When using PCA for dimensionality reduction, which of the following is most likely to occur during the visualization process?

a) The data will be mapped into a grid of 3D space
b) The first few principal components will be used to visualize the maximum variance in the data
c) The data is clustered into distinct groups
d) Non-linear relationships between features are preserved

Answer: _____

4. Which of the following techniques is typically used in data visualization to reduce the dimensionality of data while preserving pairwise distances between neighboring data points?

a) PCA b) t-SNE c) Self-Organizing Maps (SOM) d) SVD

Answer: _____

5. Which technique would be most suitable for visualizing high-dimensional data that has a non-linear manifold structure?

a) PCA b) Isomap c) K-means clustering d) SVD

Answer: _____

6. What is the primary purpose of using Self-Organizing Maps (SOM) in data visualization?

a) To reduce dimensionality for large datasets
b) To visually represent high-dimensional data on a 2D grid
c) To find clusters in data before performing dimensionality reduction
d) To calculate principal components for data visualization

Answer: _____

7. In the context of data visualization, what is a key advantage of UMAP over t-SNE?

a) UMAP is more computationally expensive than t-SNE
b) UMAP can preserve both global and local structures while being faster on large datasets
c) UMAP requires more labeled data than t-SNE
d) UMAP only works with categorical data

Answer: _____

8. Which dimensionality reduction technique can be useful for visualizing data by preserving both the global and local geometrical structure of the data?

a) PCA b) MDS c) UMAP d) K-means clustering

Answer: _____

9. In data visualization, which dimensionality reduction technique is most likely to provide interpretable 2D or 3D scatter plots by transforming high-dimensional data while preserving variance?

a) t-SNE b) PCA c) LLE d) NMF

Answer: _____

10. Which of the following is an advantage of using t-SNE for visualizing high-dimensional data?

a) It can generate clear visualizations for non-linear data structures
b) It can visualize data without any pre-processing steps
c) It preserves global structures over local structures
d) It is computationally efficient on large datasets

Answer: _____

11. In data visualization, what is the primary goal when using Isomap?

a) To reduce the number of dimensions by linear projection
b) To preserve the manifold structure in a lower-dimensional representation
c) To cluster data into distinct visual groups
d) To enhance the visual appearance of the data

Answer: _____

12. Which of the following methods is often used for visualizing high-dimensional data by projecting it onto a 2D or 3D grid using a neural network?

a) Autoencoders b) t-SNE c) UMAP d) SOM (Self-Organizing Maps)

Answer: _____

13. Which dimensionality reduction technique would you likely choose if you want to visualize data while ensuring that distances between all points are preserved as much as possible?

a) PCA b) MDS c) t-SNE d) LLE

Answer: _____

14. Which of the following is a limitation of using PCA for data visualization?

a) It can only work on datasets with categorical features
b) It fails to capture the non-linear relationships between features
c) It requires the data to be labeled
d) It only works well with small datasets

Answer: _____

15. In the context of Generative Adversarial Networks (GANs), how can they be used in data visualization?

a) By creating synthetic data points to augment visual representations
b) By reducing the dimensionality of visual features
c) By organizing data into labeled clusters
d) By creating heatmaps of feature importance

Answer: _____

16. In TensorFlow, which function is used to define a model layer that performs a matrix multiplication followed by a bias addition?

a) tf.add() b) tf.matmul() c) tf.nn.bias_add() d) tf.keras.layers.Dense()

Answer: _____

17. Which of the following methods is used in TensorFlow to convert a Numpy array into a tensor?

a) tf.convert_to_tensor() b) tf.Tensor() c) tf.array() d) tf.cast()

Answer: _____

18. Which TensorFlow function is used to create a dataset from a Python generator?

a) tf.data.from_tensor_slices()
b) tf.data.Dataset.from_generator()
c) tf.data.create_dataset()
d) tf.data.generator()

Answer: _____

19. In TensorFlow, which of the following is used to optimize a neural network model during training?

a) tf.losses() b) tf.train.Optimizer() c) tf.Variable() d) tf.Session()

Answer: _____

20. Which of the following layers in TensorFlow is commonly used for image data and is part of Convolutional Neural Networks (CNNs)?

a) tf.keras.layers.Conv2D
b) tf.keras.layers.Dense
c) tf.keras.layers.Flatten
d) tf.keras.layers.LSTM

Answer: _____

21. Which TensorFlow function is used to perform gradient descent optimization?

a) tf.keras.optimizers.Adam()
b) tf.train.AdamOptimizer()
c) tf.optimizers.SGD()
d) All of the above

Answer: _____

22. In TensorFlow, which of the following is used to compile a model for training?

a) model.fit() b) model.compile() c) model.train() d) model.summary()

Answer: _____

23. Which of the following TensorFlow classes is primarily used to build models in a high-level API?

a) tf.Model b) tf.keras.Model c) tf.keras.layers.Layer d) tf.keras.Sequential

Answer: _____

24. Which TensorFlow function allows you to efficiently transform datasets by applying a function in parallel across all elements?

a) tf.data.Dataset.map() b) tf.data.parallel_map() c) tf.data.prefetch() d) tf.data.batch()

Answer: _____

25. In TensorFlow, what is the primary difference between `tf.keras.Model.fit()` and `tf.keras.Model.evaluate()`?

a) `fit()` performs model training, and `evaluate()` performs validation
b) `fit()` uses a dataset for training, while `evaluate()` uses a dataset for testing
c) `fit()` performs training for one epoch, while `evaluate()` handles multiple epochs
d) `fit()` computes the loss function, while `evaluate()` computes the accuracy only

Answer: _____

26. What is the purpose of `tf.function()` in TensorFlow?

a) To enable automatic differentiation during backpropagation
b) To convert a Python function into a TensorFlow graph for optimization
c) To create a custom layer in a model
d) To perform non-linear activation functions

Answer: _____

27. What is the default behavior of `tf.keras.layers.Dropout` during inference (model evaluation)?

a) It has no effect; it drops random units
b) It applies the same dropout rate as during training
c) It disables dropout entirely, passing all units through
d) It randomly drops a fixed number of units during evaluation

Answer: _____

28. Which of the following is a common reason to use `tf.data.Dataset.prefetch()` in TensorFlow data pipelines?

a) To shuffle the dataset before training b) To optimize data input by loading data asynchronously
c) To apply a data augmentation function d) To normalize the data before input into the model

Answer: _____

29. When using `tf.keras.layers.Conv2D`, which of the following is true about the padding mode 'same'?

a) It will only pad the input if the dimensions are odd
b) It pads the input such that the output size is the same as the input size
c) It performs padding only in one dimension (height or width)
d) It always results in zero-padding

Answer: _____

30. When training a model in TensorFlow using `tf.keras.Model.fit()`, which of the following is NOT automatically handled by the `fit()` method?

a) Loss computation
b) Gradient computation
c) Weight updates
d) Model evaluation on a validation dataset

Answer: _____

31. Which of the following is a characteristic of temporal data?

A) Data collected over multiple regions at a specific time
B) Data points are dependent on a time sequence
C) Data is completely independent
D) Data has no correlation with time

Answer: _____

32. What does the additive model in time series analysis assume?

A) The trend, seasonality, and residual components multiply together
B) The components of the series are added together
C) The trend is exponential
D) The series is a combination of both stationary and non-stationary parts

Answer: _____

33. In a multiplicative model, how are the components of the time series combined?

A) They are subtracted
B) They are added together
C) They are multiplied
D) They are divided

Answer: _____

34. Which of the following is *NOT* a component of time series decomposition?

A) Trend
B) Seasonal
C) Exponential
D) Residual

Answer: _____

35. In time series decomposition, what does the trend component represent?

A) The short-term fluctuations around a mean
B) The cyclical patterns over long periods
C) The overall direction of the data over time
D) The random noise in the data

Answer: _____

36. What does autocorrelation measure in a time series?

A) The relationship between the trend and seasonal components
B) The correlation of a time series with its own past values
C) The relationship between the residuals and the observed data
D) The variation of data around its trend

Answer: _____

37. Which condition must be met for a time series to be stationary?

A) The variance of the series should not change over time
B) The mean should be increasing over time
C) The series should have no autocorrelation
D) The series must have constant seasonal components

Answer: _____

38. In the exponential smoothing method, which factor controls the weight given to the most recent observation?

A) The smoothing parameter (alpha)
B) The number of past observations
C) The seasonal component
D) The trend component

nswer: _____

39. What is the primary advantage of exponential smoothing over traditional moving average methods?

A) It is computationally simpler
B) It places more weight on recent observations
C) It does not require seasonal decomposition
D) It can handle non-linear trends

Answer: _____

40. Which of the following is *NOT* an assumption of the ARIMA model?

A) The series is stationary
B) The series contains a trend
C) The series is independent of past values
D) The series may contain autocorrelation

Answer: _____

41. In ARIMA, what does the parameter 'p' represent?

A) The number of lag observations included in the model for the autoregressive term
B) The number of differencing required to make the series stationary
C) The number of periods to forecast ahead
D) The number of seasonal components

Answer: _____

42. Which of the following is a key feature of the ARIMA model?

A) It accounts for both autoregressive and moving average components
B) It only handles seasonal components
C) It requires no prior data processing
D) It is primarily used for classification tasks

Answer: _____

43. What does the 'I' stand for in ARIMA?

A) Independence
B) Interpolation
C) Integration (differencing)
D) Invariance

Answer: _____

44. Which of the following techniques would be most appropriate to handle a time series with both trend and seasonality?

A) Simple Moving Average
B) Exponential Smoothing with Trend and Seasonality
C) ARIMA without differencing
D) Seasonal Decomposition of Time Series (STL)

Answer: _____

45. What is a key difference between a stationary and a non-stationary time series?

A) Stationary time series have no seasonality
B) Non-stationary time series have constant variance and mean
C) Stationary time series have a constant mean and variance over time
D) Non-stationary time series do not require differencing

Answer: _____

46. Which of the following metrics is commonly used to evaluate the accuracy of a time series forecast by comparing the difference between the predicted and actual values?

A) Mean Absolute Percentage Error (MAPE)
B) R-Squared
C) F1-Score
D) Precision

Answer: _____

47. What does the Mean Squared Error (MSE) evaluate in time series forecasting?

A) The average absolute difference between predicted and actual values
B) The squared differences between predicted and actual values
C) The ratio of forecast error to actual values
D) The proportion of correct predictions in a dataset

Answer: _____

48. Which of the following metrics is preferred when comparing forecast errors where larger errors should be penalized more heavily?

A) Mean Absolute Error (MAE)
B) Root Mean Squared Error (RMSE)
C) Mean Absolute Percentage Error (MAPE)
D) R-Squared

Answer: _____

49. In which situation is the Mean Absolute Percentage Error (MAPE) metric NOT appropriate to use?

A) When the time series data contains large seasonal variations
B) When there are zero or near-zero values in the actual values
C) When dealing with positive-only data
D) When comparing forecast models with different time series

Answer: _____

50. What is the key characteristic of the Symmetric Mean Absolute Percentage Error (sMAPE) metric?

A) It penalizes forecast errors equally for over-predictions and under-predictions
B) It requires the forecast to always be larger than the actual value
C) It only works for continuous data
D) It ignores outliers in the forecast errors

Answer: _____

---

51. Which of the following evaluation metrics is commonly used for binary classification tasks to assess the performance of a model based on both precision and recall?

A) Mean Squared Error (MSE) B) F1-Score C) R-Squared D) Adjusted R-Squared

Answer: _____

52. What does the Confusion Matrix provide for evaluating the performance of a classifier?

A) A summary of the model's true positive, false positive, true negative, and false negative counts
B) The overall accuracy of the model
C) A plot showing the model's performance on training and validation sets
D) The coefficients of the linear regression model

Answer: _____

53. What is the purpose of using Cross-Validation in model evaluation?

A) To ensure the model has learned the underlying patterns in the data
B) To reduce overfitting by testing the model on different subsets of the dataset
C) To increase the complexity of the model
D) To speed up the training process

Answer: _____

54. Which of the following metrics is commonly used to evaluate the performance of a regression model by measuring the proportion of variance explained by the model?

A) Precision B) Recall C) R-Squared (R^2) D) Area Under the ROC Curve (AUC)

Answer: _____

55. In a multi-class classification problem, what does the Macro-averaged F1-Score calculate?

A) The F1-Score for each class and then averages them, treating all classes equally
B) The weighted average F1-Score of the classes based on their frequency
C) The harmonic mean of Precision and Recall for the majority class only
D) The average of Precision and Recall for all classes

Answer: _____

56. In extreme class classification (ECC), what does the term "extreme" refer to?

A) Classes with very similar features
B) A large number of classes, often in the order of thousands or more
C) Classes with very high precision
D) The extreme importance of only a few classes

Answer: _____

57. Which of the following is a typical challenge in extreme class classification?

A) Having too few features to learn from
B) High computational complexity due to the large number of classes
C) A lack of available labeled data
D) Difficulty in distinguishing between classes with very few features

Answer: _____

---

58. Which of the following evaluation metrics is not typically used for extreme class classification?

A) Precision at k (P@k)
B) F1-Score
C) Area Under the ROC Curve (AUC)
D) Mean Squared Error (MSE)

Answer: _____

---

59. Which of the following is an example of structured data?

A) Image files B) CSV files C) Audio recordings D) Video clips

Answer: _____

60. Which of the following is not an example of semi-structured data?

A) JSON files B) XML files C) Relational databases D) CSV files with missing headers

Answer: _____

61. What is the primary characteristic of multimodal data?

A) Data derived from multiple structured sources
B) Data generated from multiple different modalities, such as text, images, and audio
C) Data processed through a single modality
D) Data that contains missing values

Answer: _____

62. Which of the following is a mutable data type in Python?

A) Tuple B) String C) List D) Integer

Answer: _____

63. What is the default data type of a variable in Python when it is assigned an integer value?

) String B) Integer C) Float D) List

Answer: _____

64. Which of the following Python data types is immutable?

A) Dictionary B) List C) Set D) String

Answer: _____

65. What does the `type()` function in Python return?

A) The length of an object
B) The value of an object
C) The data type of an object
D) The memory address of an object

Answer: _____

66. Which of the following is commonly used for data augmentation in image processing?

A) Normalization
B) Random rotations and flips
C) Feature scaling
D) One-hot encoding

Answer: _____

67. In text data augmentation, what is a common technique to create more diverse training data?

A) Replacing words with their synonyms
B) Normalizing the text
C) Removing stop words
D) Tokenization

Answer: _____

68. Which of the following is a popular method for generating synthetic data in machine learning?

A) K-fold cross-validation
B) Generative Adversarial Networks (GANs)
C) Principal Component Analysis (PCA)
D) Hierarchical clustering

Answer: _____

76 In Random Forest models, what is the primary reason for using multiple decision trees?

A) To overfit the training data
B) To reduce variance by averaging the predictions of multiple trees
C) To speed up the training process
D) To reduce the complexity of the decision-making process

Answer: _____

77. Which of the following is an example of an ensemble learning method?

A) Logistic Regression B) K-Nearest Neighbors (KNN)
C) Gradient Boosted Tree D) Naive Bayes

Answer: _____

78. What does the Gini impurity measure?

) The variance within the data B) The accuracy of the model
C) The homogeneity of a node (how pure the data in the node is) D) The maximum depth of the tree

Answer: _____

79. Which of the following is not a parameter that can be tuned in a decision tree algorithm?

A) `max_depth` B) `min_samples_split` C) `learning_rate` D) `criterion`

Answer: _____

80. Which of the following statements correctly describes the difference between Boosting and Bagging?

A) **Boosting** combines weak learners to create a strong learner by focusing on the mistakes made by previous models, while **Bagging** combines models trained on different subsets of data to reduce variance.
B) **Boosting** trains multiple models independently, while **Bagging** trains models sequentially, with each model improving on the previous one.
C) **Boosting** is used to increase model complexity, while **Bagging** is used to simplify models.
D) **Bagging** focuses on reducing bias, while **Boosting** focuses on reducing variance.

Answer: _____

*****