CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

Quiz 04

Name (Print):_____ Section:_____

1. Which of the following statements best describes **feature extraction**?
   (a) It selects a subset of the original features based on certain criteria.
   (b) It transforms the original features into a new set of features, often combining them to capture the underlying structure of the data.
   (c) It removes outlier data points to clean the dataset.
   (d) It increases the dimensionality of the data by creating additional features.

2. Which of the following statements best describes **feature selection**?
   (a) It combines features to form a new feature space.
   (b) It creates new features by transforming the original ones.
   (c) It chooses a subset of the original features based on measures of relevance or importance.
   (d) It standardizes features to have zero mean and unit variance.

3. The LOSS method is an example of a feature extraction technique
   (a) True
   (b) False

4. PCA is an example of a feature extraction technique
   (a) True
   (b) False

5. What is the main goal of dimensionality reduction in data mining?
   (a) To increase the number of features available.
   (b) To remove redundant or irrelevant features while preserving important structure.
   (c) To convert categorical variables into numerical values.
   (d) To maximize the number of dimensions for visualization.

6. Dimensionality reduction can help alleviate issues related to overfitting in machine learning models.
   (a) True
   (b) False

7. Which technique is typically used for supervised dimensionality reduction, where class labels are available?

   (a) PCA
   (b) LDA
   (c) Independent Component Analysis (ICA)
   (d) Hierarchical Clustering

8. Select the correct statement:

   (a) PCA is a supervised and LDA is an unsupervised technique
   (b) PCA is an unsupervised and LDA is a supervised technique

9. In PCA, a **principal component** is defined as:

   (a) A linear combination of the original features that captures the maximum variance.
   (b) A nonlinear transformation of the data.
   (c) The original feature with the highest variance.
   (d) A randomly selected subset of features.

10. PCA is sensitive to the scaling of data, so it is often necessary to standardize features before applying it

    (a) True
    (b) False

11. LDA seeks to maximize which of the following?

    (a) Within-class variance
    (b) Between-class variance
    (c) Total variance
    (d) Noise ratio

12. In content-based recommendation systems, the recommendations are primarily based on:

    (a) User ratings from similar users.
    (b) Item attributes and content features.
    (c) Random selection of items.
    (d) Global popularity of items.

13. Collaborative filtering can suffer from the cold-start problem

    (a) True
    (b) False

14. Matrix factorization in recommendation systems is primarily used to decompose which matrix?

(a) The feature-by-feature covariance matrix.

(b) The user-item rating matrix.

(c) The item similarity matrix.

(d) The user-user similarity matrix.

15. What is one of the key benefits of using matrix factorization techniques in recommendation systems?

(a) They guarantee a perfect prediction for every user.

(b) They uncover latent factors that explain user-item interactions.

(c) They remove the need for any user data.

(d) They are not affected by the cold-start problem.

16. A **scree plot (elbow plot)** in PCA is often used as part of the process to select the number of principal components. Which of the following is a common rule of thumb for this selection?

(a) Choose the number of components that account for at least 80–90% of the total variance.

(b) Select the number that minimizes the reconstruction error exactly.

(c) Use all components regardless of the variance explained.

(d) Choose the number that equals the number of original features.

17. **(Scenario)**: You are given a high-dimensional dataset of genomic data with hundreds of features, many of which are highly correlated. Your goal is to reduce the dimensionality for downstream clustering analysis while addressing multicollinearity. Which dimensionality reduction technique is most appropriate in this context?

(a) Principal Component Analysis (PCA)

(b) Linear Discriminant Analysis (LDA)

(c) Matrix Factorization

(d) KNN

18. **(Scenario)** Imagine you are developing a recommendation system for a new online streaming service. Since the service is new, there is very little user rating data available, but extensive metadata exists for each movie (such as genre, director, and cast). Which recommendation approach is most appropriate for this scenario?

(a) Pure collaborative filtering

(b) Content-based filtering

(c) A hybrid system that combines collaborative and content-based methods

(d) Model-based collaborative filtering only