CS316: Introduction to AI and Data Science
Major Exam

Anis Koubaa
akoubaa@psu.edu.sa
Prince Sultan University,
College of Computer and Information Sciences

Date: Monday, October 28, 2024
Duration: 60 Minutes

Instructions

- This exam is a closed book and closed notes.
- Electronic devices are not allowed except for a simple calculator.
- Read each question carefully and answer to the best of your ability.
- Write your name and student ID in the space provided below.

Student Information

Student ID:	Student Name:

1 Interview Questions

Time: 10 minutes **Score:** 4 points **Mapped CLOs:** CLO 2

Question 1: Utilizing Pandas DataFrames in AI and Data Science

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO1, CLO 2 **Word Count:** 150

Pandas is a crucial library for data manipulation and analysis in Python, essential for AI and Data Science.

- Describe how Pandas DataFrames can be used to preprocess large datasets effectively, including operations such as handling missing data and data normalization.
- Discuss the benefits of using Pandas in preparing datasets for machine learning models, particularly when dealing with complex datasets like financial records or social media data.

Answer:

Pandas DataFrames provide a powerful and flexible structure for manipulating large datasets, crucial in various stages of data preprocessing which is a fundamental step in building machine learning models. By utilizing Pandas, data scientists can efficiently handle missing data through functions like `fillna()`, `dropna()` which help in either filling the missing values with statistical measures such as mean, median or removing rows and columns with missing data entirely, depending on the scenario and necessity of data integrity.

Additionally, Pandas offers robust tools for data normalization, which is vital for many machine learning algorithms that are sensitive to the scale of data. Methods like `apply()` can be used to normalize or standardize data, ensuring that each feature contributes proportionally to the final analysis and improving the performance of algorithms like gradient descent by speeding up the convergence.

The use of Pandas becomes particularly advantageous when dealing with complex datasets such as financial records or social media data. For financial datasets, Pandas can handle time series analysis efficiently, allowing for operations like date range generation, frequency conversion, and moving window statistics, which are essential for predictive modeling. For social media data, Pandas' ability to handle large volumes of data and its string manipulation functions make it excellent for text preprocessing, enabling tasks such as tokenization, removal of stop words, and vectorization necessary for NLP applications.

Overall, Pandas is indispensable in the data scientist's toolkit, offering a rich set of functions that simplify the preparation of datasets for machine learning, enhancing both the efficiency and effectiveness of the models developed.

Question 2: Understanding Projection in PCA and Its Applications

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO 1, CLO 2 **Word Count:** 150

- Explain the role of projection in Principal Component Analysis (PCA) and how it is used to capture the most significant features in a dataset.
- Discuss the mathematical process of projecting high-dimensional data onto a lower-dimensional subspace using PCA.
- Describe scenarios in various fields where projection via PCA is used to enhance the analysis and interpretation of complex datasets.

Answer:

Projection in Principal Component Analysis (PCA) plays a pivotal role in identifying and capturing the most informative features of a dataset. By projecting the original data onto a lower-dimensional subspace, PCA seeks to retain the variance of the dataset, which reflects its intrinsic structure.

Mathematically, projection in PCA involves calculating the eigenvalues and eigenvectors of the covariance matrix of the data. The eigenvectors represent the directions of the new feature space, and the eigenvalues indicate the importance of these directions in terms of explained variance. The data is then projected onto the top k eigenvectors, corresponding to the k largest eigenvalues, effectively reducing the dimensionality while preserving as much of the data's variation as possible.

In practice, projection via PCA is extensively used across various fields. In image processing, PCA helps in noise reduction and feature extraction. In genomics, it aids in identifying patterns and structures within genetic data, which are crucial for further biological analysis and interpretation. By reducing the complexity of data while maintaining essential information, PCA enables more efficient processing and deeper insights in these and other applications.

Question 2: Role of Determinants and Eigenvalues in Linear Algebra

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO 1, CLO 2 **Word Count:** 50

Determinants and eigenvalues are fundamental concepts in linear algebra, crucial for understanding system stability and transformations. What is the **PRIMARY** role of determinants and eigenvalues in the analysis of matrices in linear algebra? **A clear explanation of your answer is essential.**

A) By providing a method for matrix inversion and solving linear systems.

B) By determining the solvability of linear systems and their behavior under transformations.

C) By optimizing computational algorithms for faster matrix calculations.

D) By facilitating the simplification of matrix operations and reducing their complexity.

Answer:

Answer: B) By determining the solvability of linear systems and their behavior under transformations.

Explanation: Determinants help in assessing the solvability of linear systems, indicating whether a unique solution exists (non-zero determinant) or not (zero determinant). Eigenvalues are critical for understanding the behavior of a matrix under linear transformations, including the stability, rotation, and scaling of systems, which are essential in many applications of linear algebra.

Question 4: Interpreting Box Plot Results in Data Analysis

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO 3 **Word Count:** 50

Box plots are a graphical representation commonly used in statistical analysis to depict groups of numerical data through their quartiles. What key aspects should be considered when interpreting the results of a box plot? It is important to clearly explain your answer.

Answer:

Answer: Box plots, also known as whisker plots, provide a concise visual summary of key statistical data points, which are essential for analyzing and interpreting data distributions. The central feature of a box plot is the box itself, which delineates the first and third quartiles of the dataset, encapsulating the middle 50

The whiskers of a box plot extend to show the range of the data, typically up to 1.5 times the IQR from the quartiles, though this can vary depending on the specific data or field conventions. Data points that fall outside this range are considered outliers and are often plotted as individual points. These outliers are crucial for identifying anomalies or errors in the data, as well as highlighting data variability.

Furthermore, the symmetry of the box and whiskers around the median can indicate the skewness of the data distribution. A symmetric box plot suggests a more uniform distribution, while an asymmetric one can indicate skewed data. Thus, interpreting a box plot can provide valuable insights into the spread, central value, and overall shape of the data distribution, which are vital for statistical analysis and decision-making processes.