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: Application of Eigenvalues, Eigenvectors, and PCA in AI and Data Science

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

PCA is an important technique in AI and Data Science.

- Explain the significance of eigenvalues and eigenvectors in the computation of the covariance matrix and their role in PCA.
- Discuss how PCA can be applied to enhance machine learning models in AI, specifically focusing on a dataset with many features, such as heath records data data or large-scale sensor data from IoT devices.

Possible Answer 1:

Principal Component Analysis (PCA) is a method used to simplify complex data sets in AI and Data Science by reducing the number of features (or dimensions) in the data while retaining the most essential information.

Eigenvalues and **eigenvectors** play a crucial role in this process:

- **Eigenvalues** quantify how much variance (or spread) there is in the directions of the data. A higher eigenvalue indicates more variance, suggesting a feature's importance.
- **Eigenvectors** indicate the directions of these variances. They act like arrows showing where the bulk of the data is concentrated and which directions are most spread out.

In PCA, the covariance matrix is computed to determine how much each pair of features in the dataset varies together. Through the eigenvalues and eigenvectors, PCA identifies the principal directions where the data varies the most and reorients the dataset along these new axes.

The benefits of applying PCA in AI projects with high-dimensional data such as images or extensive sensor data include:

- **Reduces Overfitting:** By keeping only the most significant features, PCA helps simplify machine learning models and prevents them from merely memorizing the training data.
- **Improves Efficiency:** Reducing the data's complexity can speed up computing times and enhance the performance of models.

- **Uncovers Patterns:** PCA can reveal underlying patterns in the data that are not immediately obvious, useful for tasks like anomaly detection in sensor data or feature extraction in image processing.

Overall, PCA serves as a powerful tool for managing complex data, enhancing model performance, and discovering hidden patterns in AI and Data Science.

Possible Answer 2:

Eigenvalues and eigenvectors form the backbone of PCA, which is a powerful tool for dimensionality reduction in AI and Data Science. They are crucial for decomposing the covariance matrix, which helps in identifying the directions of maximum variance in the data, known as principal components.

Eigenvalues indicate the amount of variance carried in each principal component, while **eigenvectors** define the direction of these components. This decomposition is vital in PCA, allowing for the reduction of complex datasets into simpler, more manageable forms without significant loss of information.

Applying PCA in AI, particularly in handling high-dimensional datasets like image or sensor data, significantly improves the efficiency of machine learning models. It reduces overfitting by eliminating noisy or less informative features and enhances computational performance by lowering the dimensionality.

This technique is especially useful in scenarios where underlying patterns in the data need to be uncovered before further analysis, such as in anomaly detection in sensor networks or feature extraction in image processing.

Question 2: Application of Dot Product, Cosine Similarity, and Normalization in NLP and Semantic Search

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

- Explain how the dot product is used in calculating cosine similarity for semantic search.
- Discuss the importance of vector normalization in the context of cosine similarity and its implications for NLP applications.
- Describe how cosine similarity enables effective semantic search in NLP applications.

Answer:

In Natural Language Processing (NLP), understanding the relationship and similarity between text documents is crucial for tasks such as semantic search, where the goal is to find documents that are contextually similar to a query.

Dot Product and **Cosine Similarity:**

- The **dot product** is a fundamental mathematical operation used to calculate the cosine similarity between two vectors. In NLP, these vectors often represent

text documents or words converted into vectors using models like TF-IDF or word embeddings.

- **Cosine similarity** measures the cosine of the angle between two vectors. It is calculated using the dot product divided by the product of the magnitudes (norms) of the vectors. This metric indicates how similar two documents are, irrespective of their size, by focusing purely on the direction of the vectors.

Normalization is critical in the context of cosine similarity because:

- Normalizing a vector to unit length adjusts all vectors to a consistent scale and simplifies the calculation of cosine similarity. It ensures that the similarity measurement focuses on the orientation of the vectors, not their magnitude, which is particularly important in comparing text documents of different lengths.
- Normalization helps stabilize computations by avoiding issues with very large or small scale values in vector components, facilitating faster and more stable convergence in algorithms that rely on iterative methods.
- By normalizing vectors, the focus shifts more to the deviation of angles rather than scale, enhancing the accuracy of similarity measures in semantic analysis.

Cosine similarity, facilitated by the dot product and normalization, allows semantic search systems to effectively retrieve documents that are contextually related to a search query, enhancing the user's search experience by focusing on the content's meaning rather than just keyword matches.

These techniques form the backbone of advanced search algorithms in NLP and are essential for systems that require a deep understanding of textual content, such as recommendation systems and automated question-answering systems.

2 Critical-Thinking Questions

Time: 10 minutes **Score:** 4 points **Mapped CLOs:** CLO 1, CLO 3

For this critical thinking question, provide the correct answer along with a brief, one-sentence explanation.

Question 1: Understanding SVD in Dimensionality Reduction

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

Singular Value Decomposition (SVD) is extensively used in data science, particularly for dimensionality reduction in large datasets. What is the **PRIMARY** process through which SVD achieves dimensionality reduction? **It is important to clearly explain your answer.**

A) By directly reducing the computational complexity of data processing.

B) By optimizing the storage requirements of large datasets.

C) By eliminating less significant features based purely on their variance contribution.

D) By identifying and decomposing the dataset into components of maximum variance and significance.

Answer: D
Explanation: SVD achieves dimensionality reduction primarily by decomposing a matrix into its singular values and corresponding vectors, focusing on the components that capture the most variance and significance.

Details:

A) While reducing computational complexity is a benefit of using SVD, it is not the primary mechanism of dimensionality reduction. SVD's main role is in data transformation and feature extraction, not direct computational optimization.

B) Optimizing storage is a secondary effect of reducing the dimensionality of datasets via SVD; however, it does not describe how SVD functions to reduce dimensions.

C) This option is partially correct as SVD does eliminate less significant features, but it is not solely based on variance. SVD considers both the magnitude (significance) and direction (variance) of data features, providing a more holistic view than simply dropping features based on variance alone.

D) Correct. SVD decomposes the dataset into a series of singular values and vectors, where the largest singular values and their corresponding vectors represent the most significant data features with the highest variance. Retaining these components and

discarding the rest effectively reduces the dimensionality while preserving essential data characteristics.

Question 2: Vector Representation in Data Science

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO 4 **Word Count:** 50

In data science, raw data such as text or images often need to be transformed into feature vectors. What is the primary reason for representing this data in vector form? Explain your answer.

A) To increase the size of the dataset for better processing.

B) To enable the application of matrix-based operations that are essential to build machine learning models.

C) To enhance the visual representation of data when presenting to stakeholders.

D) To ensure data security and compliance with data protection regulations.

Answer: B
Explanation: Representing data in vector form allows the application of mathematical and algebraic operations necessary for machine learning algorithms to process and learn from the data effectively.

Details: Data such as text and images are inherently unstructured and high-dimensional. By converting text into vectors using techniques like TF-IDF or word embeddings, and images into pixel intensity vectors or more sophisticated feature vectors using techniques like CNNs, we can apply linear algebra operations, optimize computational efficiency, and implement machine learning algorithms effectively. This transformation enables models to identify patterns, make predictions, and provide insights that are not feasible with raw data.