Question 1

A data scientist works with a dataset that has non-linear relationships between features. They want to apply a dimensionality reduction technique that generalizes PCA to handle nonlinearity. Which method should they use?

Options:

A) Standard PCA with L2 regularization

B) Kernel PCA using a radial basis function (RBF) kernel

C) Linear Discriminant Analysis (LDA)

D) t-SNE with PCA initialization

Answer: B

Explanation: Kernel PCA extends traditional PCA by mapping data to a higher-dimensional feature space using a kernel function (e.g., RBF), enabling it to capture nonlinear relationships. Standard PCA (A) works only for linear correlations. LDA (C) is a supervised method for classification, and t-SNE (D) is primarily for visualization rather than feature extraction.

Question 2

A data scientist at a property tech startup is building a regression model to predict housing prices. The dataset includes 50 features (e.g., square footage, proximity to schools, number of fireplaces). To prevent overfitting and improve interpretability, they want to apply regularization during gradient descent. However, they are unsure whether to use L1 (Lasso) or L2 (Ridge) penalties for this task. What is the primary practical difference between using L1 and L2 regularization in this scenario?

Options:

A) L1 penalty shrinks coefficients toward the mean, while L2 sets some exactly to zero

B) L1 penalty introduces bias, while L2 doesn't affect the model at all

C) L1 penalty tends to set some coefficients exactly to zero, while L2 shrinks all coefficients

D) Both penalties have the same effect but L1 is more computationally efficient

Answer: C

Explanation: The L1 (Lasso) penalty tends to produce sparse models by setting some coefficients exactly to zero, making it useful for feature selection. The L2 (Ridge) penalty shrinks all coefficients toward zero but rarely sets any exactly to zero.

Question 3 (MCQ)

A product team at a search engine company is simulating how users navigate the internet to improve their web ranking algorithm. They want to model a user who randomly clicks hyperlinks on different web pages without any specific goal or search strategy. This helps them analyze how page importance can be determined based on link structure alone.

In this context, what is a “random surfer”?

Options:

A) A user who browses websites without any specific goal

B) A theoretical model that randomly clicks hyperlinks to move between web pages

C) A bot designed to test website loading times

D) A person who uses search engines incorrectly

Answer: B

Explanation: The “random surfer” is a theoretical model introduced in algorithms like PageRank. It represents a user who randomly navigates through hyperlinks across web pages, modeling the likelihood of visiting a page. This model helps rank pages based on how often they would be visited in a large web graph.

Question 4 (MCQ)

A computer science student is analyzing a simple website structure with three pages: A, B, and C. Each page contains hyperlinks to one of the others, forming a circular connection (A links to B, B links to C, and C links back to A). The student is trying to identify potential issues with this kind of structure during user navigation or while designing search algorithms.

What issue is a common issue that can occur in this simple three-page (A, B, C) web structure?

Options:

A) The pages had too many hyperlinks between them

B) One page had no incoming links

C) The structure could create an infinite loop for users navigating through hyperlinks

D) The structure had broken links that led nowhere

Answer: C

Explanation: The circular reference structure (A→B→C→A→...) can cause a loop where users (or algorithms like the random surfer model) continuously cycle through the same pages. This loop prevents discovery of new content and can hinder effective indexing or ranking.

Question 5 (MCQ)

Your friend is studying Page-Rank Algorithm and he came across a term "in-degree". What does the "in-degree" of a page represent?

Options:

A) The number of pages that the page links to

B) The number of pages that link to the page

C) The ranking of the page in search results

D) The number of visitors to the page

Answer: B

Explanation: The in-degree of a page represents the number of other pages that have hyperlinks pointing to that page. In the example, page B had an in-degree of 2 because both pages A and C linked to it.

Question 6 (MCQ)

Let there are 3 pages, page A has links pointing to page C, page B has links pointing to page A, and page C has links pointing to both page A and page B. In this three-page example (A, B, C), what is the out-degree of page C?

Options:

A) 0

B) 1

C) 2

D) 3

Answer: C

Explanation: Page C had links pointing to both page A and page B, giving it an out-degree of 2. The out-degree represents the number of hyperlinks on a page that point to other pages.

Question 7 (NAT)

A network administrator is designing a simple ring topology for a small office network. In this setup, each of the 6 computers (nodes) is directly connected to exactly 2 other computers, forming a closed loop where data can circulate in either direction.

Question: Based on this configuration, how many direct connections (edges) are there in total in the network?

Correct Answer: 6

Explanation: Each of the 6 computers is connected to 2 others, so the total degree sum is $6 \times 2 = 12$. Since each connection is shared between two computers, the total number of unique connections is $12/2 = 6$.

Question 8

Question: A digital marketing team at a news website wants to improve the site's visibility on search engines like Google. The team understands that Google's PageRank algorithm plays a key role in determining how prominently their articles appear in search results. They are analyzing several factors that might influence their PageRank and are considering different strategies to boost it. Which factors will directly influence the PageRank of their web pages? (Select all that apply.)

A) Number of inbound links

B) Number of outbound links from linking pages

C) Content length of the page

D) The damping factor

Correct Answer: A) Number of inbound links, B) Number of outbound links from linking pages, D) The damping factor
Explanation: A page's rank depends on the ranks of linking pages, their outbound links, and the damping factor. (Content length does not directly impact PageRank.)

Question 9

A technology company is developing a new web search engine. The product team wants to ensure that users get the most relevant and authoritative web pages at the top of their search results, even when those pages don't have the highest word count or most keywords. The team is evaluating different ranking algorithms. One approach under consideration uses the structure of hyperlinks between web pages to determine which pages are more important. Why would the PageRank algorithm be particularly useful for this search engine's ranking system?

A. It ranks pages solely based on their content.

B. It ranks pages based on both content and connectivity.

C. It ranks pages based on their connectivity.

D. It ensures all pages have equal ranks.

Answer: C
Explanation:

- The PageRank algorithm ranks pages based on their connectivity in the web graph.
- It uses the structure of hyperlinks to measure the relative importance of pages.

Question 10

A data analyst at a retail company is working with a dataset containing two features for each product: price and customer rating. The analyst wants to simplify the data for visualization and further analysis.

The analyst applies a linear transformation using the matrix

$$A = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$$

to different vectors representing product data.

Which of the following statements are correct?

A. The vector $[1, 0]$ changes its direction when transformed by A.

B. The vector $[2, 2]$ is an eigenvector of A with eigenvalue 5.

C. The vector $[\frac{1}{2}, -\frac{1}{2}]$ is an eigenvector of A with eigenvalue 3.

D. All eigenvectors of A have the same eigenvalue.

Solutions: A, B, C

Explanation: Finding the eigenvalues using $|A - \lambda I| = 0$ gives $\lambda = 5, 3$. Thus, there are 2 eigenvalues. Then, solve for v (eigenvector) using equation $Av = \lambda v$ which gives $[0.7071, -0.7071]$ and $[0.7071, 0.7071]$. Their multiples will also be eigenvectors. Can verify the same by substituting the vector in place of v in $Av = \lambda v$.

Question 11

A machine learning engineer is preparing a dataset for training a predictive model. The dataset contains information about four different products, each described by two features: sales and returns.

$$A = \begin{bmatrix} 5 & 2 \\ 3 & 1 \\ 4 & 3 \\ 2 & 2 \end{bmatrix}$$

What will be the centered matrix of this matrix?

Options:

A) $\begin{bmatrix} 1.5 & 0 \\ -0.5 & -1 \\ 0.5 & 1 \\ -1.5 & 0 \end{bmatrix}$

B) $\begin{bmatrix} 1.5 & -1.5 \\ 1 & -1 \\ 0.5 & -0.5 \\ 0 & 0 \end{bmatrix}$

C) $\begin{bmatrix} 1.5 & 0 \\ 0.5 & -1 \\ -0.5 & 1 \\ -1.5 & 0 \end{bmatrix}$

D) None of the Above

Answer: C

Explanation: Calculate column means (as datapoints are columns of the matrix):

Column 1 mean: $(5+3+4+2)/4 = 3.5$
Column 2 mean: $(2+1+3+2)/4 = 2$

Now, subtract from each datapoint that column's mean. Resultant matrix:

$$\begin{bmatrix} 1.5 & 0 \\ -0.5 & -1 \\ 0.5 & 1 \\ -1.5 & 0 \end{bmatrix}$$

Question 12

A business analyst is reviewing a dataset of three products, each with two features: units sold and customer complaints.

$$X = \begin{bmatrix} 5 & 2 \\ 3 & 1 \\ 4 & 3 \end{bmatrix}$$

What will be the centered matrix of this matrix?

A. $\begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{bmatrix}$

B. $\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$

C. $\begin{bmatrix} 0.5 & 1.0 \\ 1.0 & 0.5 \end{bmatrix}$

D. None of the above

Answer: A
Explanation:

- Center the data by subtracting the column means:

$$X_{\text{centered}} = \begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{bmatrix}$$

Question 13

After performing PCA on the covariance matrix of data from Q12, which eigenvector should be selected as the principal component?

A. The eigenvector corresponding to the smallest eigenvalue.

B. The eigenvector corresponding to the largest eigenvalue.

C. The eigenvector corresponding to the mean of eigenvalues.

D. None of the above.

Answer: B
Explanation: The eigenvector corresponding to the largest eigenvalue captures the most variance in the data and is chosen as the principal component.

Question 14

Continuing from the previous scenario, Provided covariance matrix: $\begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$ and eigenvalues: 0.5 and 1.5, the startup wants to quantify how much variance is captured by the second principal component after PCA (rounded to 2 decimal places)?
Answer: 0.25
Explanation: The variance captured by the second principal component is calculated by dividing 2nd PC by sum of PCs: $0.5/(1.5 + 0.5) = 0.25$

Question 15

A data science team is discussing the properties of PCA as they prepare a report on dimensionality reduction for management. Which of the following statements about PCA are correct?

A. The principal component maximizes the variance in the data projection.

B. PCA reduces dimensionality while preserving 100 percent of the original data information.

C. The first principal component is orthogonal to all other components.

D. Eigenvalues determine the amount of variance captured by their respective eigenvectors.

Answer: A, C, D
Explanation:

- PCA reduces dimensionality by maximizing the variance in data projection (A).
- The first principal component is orthogonal to all other components (C).
- Eigenvalues quantify the variance captured by their eigenvectors (D).
- However, PCA reduces dimensionality at the cost of losing some information, making B incorrect.