Practice

This document has 13 questions.

Question-1

Statement

An image is a collection of pixels. A pixel is stored as a float value and typically occupies 4 bytes of memory. Consider a dataset of 1000 images, where each image has dimensions 100×100. Approximately, how much memory does the entire dataset occupy?

Options

(a)

4 KB

(b)

4 MB

(c)

40 MB

(d)

4 GB

Answer

(c)

Solution

We require $100 \times 100 = 10000$ float values to represent one image. Since each float value occupies 4 bytes of memory, a single image occupies 40000 bytes of memory. Roughly, this corresponds to 40 KB. The entire dataset would occupy 1000×40 KB or 40 MB of memory. Here, we have used the following facts:

- 1 KB ≈ 1000 bytes
- 1 MB ≈ 1000 KB

Question-2

Statement

Consider a dataset that has 100 points that belong to $\mathbb{R}^3$. All of them are found to lie on a line that passes through the origin. We use a unit vector along the line as a representative and the coefficients with respect to it to represent the individual data-points. Compute the percentage decrease in the size of the dataset if we move to this new representation. Assume that it takes one unit of space to store one feature. Enter your answer correct to two decimal places; it should be in the range $[0, 100]$.

Answer

65.66

Range: $[65, 66]$

Solution

The size of the dataset in its original form is:

$$S_1 = 100 \times 3 = 300$$

The size of the dataset after moving to the new representation:

$$S_2 = 3 + 100 = 103$$

The percentage decrease in the size of the dataset is therefore:

$$\frac{S_2 - S_1}{S_1} \times 100 = \frac{300 - 103}{300} \times 100 \approx 65.66\%$$

Common Data for questions (3) and (4)

Statement

Consider the following dataset that has four points, all of which lie on a line:

$$S = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \begin{bmatrix} 6 \\ 8 \end{bmatrix}, \quad \begin{bmatrix} 1/5 \\ 4/15 \end{bmatrix} \right\}$$

Answer the questions that follow:

Question-3

Statement

Among the vectors given below, choose a representative that has unit length.

Options

(a)

$$\begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1/15 \\ 4/15 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Answer

(c)

Solution

The length of a vector $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ is given by:

$$||\mathbf{w}|| = \sqrt{\mathbf{w}^T\mathbf{w}} = \sqrt{w_1^2 + w_2^2}$$

We need to find that vector which has $||\mathbf{w}|| = 1$. From the options, we see that the required vector is:

$$\begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$$

Question-4

Statement

With respect to the representative in the previous question, compute the coefficients for these four points. The i^{th} element from the left in each option is the coefficient for the i^{th} element from the left in the set S.

Options

(a)

$\{0, \quad 5, \quad 10, \quad 1/3\}$

(b)

$\{0, \quad 1, \quad 2, \quad 1/3\}$

(c)

$\{0, \quad 5, \quad 10, \quad 3\}$

(d)

$\{0, \quad 1, \quad 10, \quad 1/3\}$

Answer

(a)

Solution

The representative and the dataset are given below:

$$\mathbf{w} = \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}, \quad S = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \begin{bmatrix} 6 \\ 8 \end{bmatrix}, \quad \begin{bmatrix} 1/5 \\ 4/15 \end{bmatrix} \right\}$$

The coefficient of a point $\mathbf{x}$ with respect to $\mathbf{w}$ is:

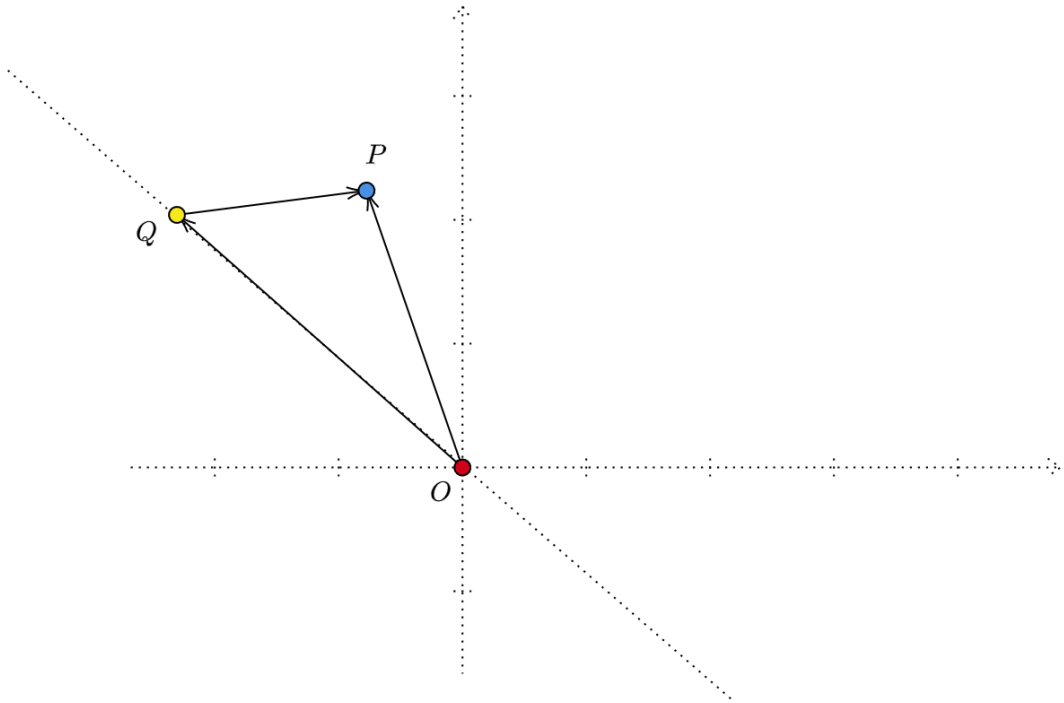
$$\mathbf{x}^T \mathbf{w}$$

Note that this equation holds only if $||\mathbf{w}|| = 1$. What will change if $||\mathbf{w}|| \neq 1$? Think about this.
The coefficients are therefore:

$$\{0, \quad 5, \quad 10, \quad 1/3\}$$

Common Data for questions (5) to (7)

Consider the following image. P is a point in 2D space. Q is a proxy for this point on a line passing through the origin. The image is drawn to scale.



Answer the questions that follow:

Question-5

Statement

Which of the following is the error vector?

Options

(a)

$\overrightarrow{OP}$

(b)

$\overrightarrow{OQ}$

(c)

$\overrightarrow{QP}$

Answer

(c)

Solution

Given these three vectors:

- a point
- its proxy
- the error vector

The following relationship holds:

$$\text{error-vector} = \text{point} - \text{proxy}$$

The error-vector is the difference between the original point and its proxy. Replacing the words with vector notation, we have:

$$\overrightarrow{QP} = \overrightarrow{OP} - \overrightarrow{OQ}$$

We have used the concept of vector addition which was covered in maths-2.

Question-6

Statement

Is Q the "best" representation of P on the line?

(a)

Yes

(b)

No

Answer

(b)

Solution

No, Q is not the best representation of P on the line. Geometrically, the best representation would be the one for which the error vector is perpendicular to the line. From the figure, we see that the line segment QP does not satisfy this property.

Question-7

Statement

If Q' is the "best" representation of P on the line, then which of the following statements are true? Notation: $|\overrightarrow{AB}|$ is the length of the vector $\overrightarrow{AB}$.

Options

(a)

$|\overrightarrow{QP}| < |\overrightarrow{Q'P}|$

(b)

$|\overrightarrow{Q'P}| < |\overrightarrow{QP}|$

(c)

$|\overrightarrow{QQ'}| = 0$

(d)

$|\overrightarrow{QQ'}| > 0$

Answer

(b), (d)

Solution

Computationally, the best representation would have the lowest reconstruction error. The smallest reconstruction error is achieved by the point Q', the tip of the projection of P onto the line. The reconstruction error is the square of the length of the error-vector. These are the terms $|\overrightarrow{Q'P}|^2$ and $|\overrightarrow{QP}|^2$. But we directly compare the lengths of the two error vectors. Think about why this is true.

We now have:

$$|\overrightarrow{Q'P}| < |\overrightarrow{QP}|$$

Since Q and Q' are two different points on the line, we have $\overrightarrow{QQ'} > 0$.

Question-8

Statement

Is the following statement true or false?

The projection of $\mathbf{x}$ onto $\mathbf{w}$ was derived to be $(\mathbf{x}^T\mathbf{w})\mathbf{w}$, where $\mathbf{w}$ is a unit vector. Since this derivation was done for the special case of 2D vectors, this formula is not applicable in the general case of d-dimensional vectors.

Options

(a)

True

(b)

False

Answer

(b)

Solution

This formula still holds for any two d-dimensional vectors. The geometry of 2D space is generalized to d-dimensional space. In 2D space, we can visually see what it means for the error-vector/residue to be perpendicular to the line. Though this visualization is not possible for higher dimensional spaces, the basic ideas still stand. For instance, the dot-product between two vectors in $\mathbb{R}^d$ is:

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

Likewise, the length of a vector is given by:

$$||\mathbf{x}|| = \sqrt{\mathbf{x}^T\mathbf{x}}$$

Question-9

Statement

Consider a mean-centered dataset of n points where each point belongs to $\mathbb{R}^d$. $\mathbf{w_1}, \cdots, \mathbf{w_k}$ are the first k principal components obtained by running PCA on the dataset, where $k < d$. The following relationship is observed:

$$\mathbf{x}_i - \left[\sum_{j=1}^{k} (\mathbf{x}_i^T \mathbf{w}_j)\mathbf{w}_j\right] = 0, \quad 1 \leq i \leq n$$

Which of the following statement about the dataset is true?

Options

(a)

The dataset lies in a d-dimensional subsapce of $\mathbb{R}^n$

(b)

The dataset lies in a k-dimensional subspace of $\mathbb{R}^d$

(c)

The dataset lies in a d-dimensional subspace of $\mathbb{R}^k$

(d)

The dataset lies in a k-dimensional subspace of $\mathbb{R}^n$

Answer

(b)

Solution

First, the dataset has d features and n examples. So, it doesn't make sense to talk about $\mathbb{R}^n$ as n is the number of examples. Secondly, we note that each principal component, $\mathbf{w}_i$, is a vector in $\mathbb{R}^d$. Thirdly, we know that the k principal components are orthogonal, and hence linearly independent. It follows that $S = \text{span}\left(\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}\right)$ is a k-dimensional subspace of $\mathbb{R}^d$. Finally, since each data-point in the dataset is a linear combination of these k principal components, we see that all of them should lie in S.

Question-10

Statement

In the context of PCA, given n data-points in $\mathbb{R}^d$ that are mean-centered, after estimating $\mathbf{w_1}$ in the first round, what is the mean of the residues?

Options

(a)

$\mathbf{w}_1$

(b)

$\mathbf{0}$

Answer

(b)

Solution

The mean of the residuals is the zero vector in $\mathbb{R}^d$:

$$\begin{aligned}\sum_{i=1}^{n} \mathbf{x}_i' &= \sum_{i=1}^{n} \mathbf{x}_i - (\mathbf{x}_i^T \mathbf{w_1})\mathbf{w_1} \\ &= \mathbf{0} - \left[\left(\sum_{i=1}^{n} \mathbf{x}_i\right)^T \mathbf{w_1}\right]\mathbf{w_1} \\ &= \mathbf{0}\end{aligned}$$

Here, we have used the fact that $\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$ as the data is mean-centered.

Question-11

Statement

Consider two ways of representing n datapoints that belong to $\mathbb{R}^d$ in the form of a matrix:

Approach-1: A matrix $\mathbf{X_1}$ of dimension $n \times d$

Approach-2: A matrix $\mathbf{X_2}$ of dimension $d \times n$

Assume that the dataset is mean-centered. Select all correct expressions for the covariance matrix.

Options

(a)

$\frac{1}{n}\mathbf{X}_1^T\mathbf{X}_1$

(b)

$\frac{1}{n}\mathbf{X}_2^T\mathbf{X}_2$

(c)

$\frac{1}{n}\mathbf{X}_1\mathbf{X}_1^T$

(d)

$\frac{1}{n}\mathbf{X}_2\mathbf{X}_2^T$

Answer

(a), (d)

Solution

Let the i^{th} data-point be $\mathbf{x}_i$. The expression for the covariance matrix is:

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$$

There are two ways to arrange the n data-points. We have a $d \times n$ matrix, where each column corresponds to one data-point. This form is particularly important as we will be using this extensively in the second week of the course:

$$\mathbf{X}_2 = \begin{bmatrix} \vdots & & \vdots \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ \vdots & & \vdots \end{bmatrix}$$

Then, we have:

$$\mathbf{X}_2\mathbf{X}_2^T = n \cdot \mathbf{C}$$

On the other hand, we have a $n \times d$ matrix, where each row corresponds to one data-point:

$$\mathbf{X}_1 = \begin{bmatrix} \cdots & \mathbf{x}_1 & \cdots \\ & \vdots & \\ \cdots & \mathbf{x}_n & \cdots \end{bmatrix}$$

Since $\mathbf{X}_1^T = \mathbf{X}_2$, we have:

$$\mathbf{X}_1^T \mathbf{X}_1 = n \cdot \mathbf{C}$$

Question-12

Statement

Consider a mean-centered dataset obtained from the banking domain that has 100 data-points, each of which is described by 7 features. The dataset is represented as a 100×7 matrix, $\mathbf{X}$. You run PCA on this dataset and observe that the residues vanish completely after k iterations.

A little later, a domain expert makes the following observations. If $\mathbf{c}_i$ represents the i^{th}column of $\mathbf{X}$, then:

- The set of vectors $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4\}$ are linearly independent.
- The following relations are satisfied:
  - $\mathbf{c}_5 = \mathbf{c}_1 + \mathbf{c}_3$
  - $\mathbf{c}_6 = 2\mathbf{c}_3 - 3\mathbf{c}_4$
  - $\mathbf{c}_7 = \mathbf{c}_2 + 3\mathbf{c}_4$

What is the value of k? Assume that the dataset is already mean-centered.

Answer

4

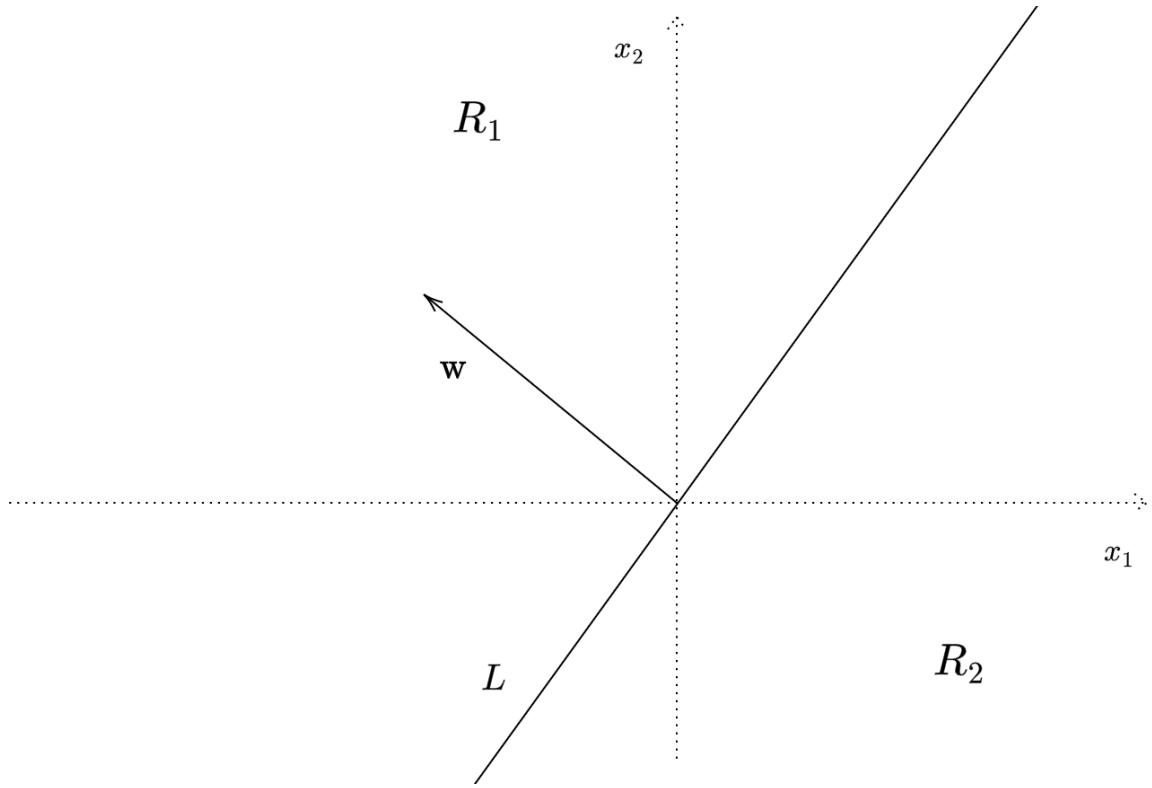
Solution

Since the last three columns are linear combinations of the first four, and since the first four columns are linearly independent, the rank of the matrix $\mathbf{X}$ is 4. This means that the rows of the matrix (the data-points) belong to a four dimensional subspace of $\mathbb{R}^7$. Intuitively, we see that PCA should terminate after four iterations and the principal components will form a basis of this subspace. For now, we shall skip the proof of this statement.

Question-13

Statement

Consider the following image:



Here, $\mathbf{w}$ is a vector and L is a line perpendicular to $\mathbf{w}$ that passes through the origin. R_1 and R_2 are two regions on either side of the line L. If $\mathbf{x}$ is an arbitrary vector in the plane, select all correct statements.

Options

(a)

$$R_1 : \mathbf{w}^T\mathbf{x} > 0$$

(b)

$$R_2 : \mathbf{w}^T\mathbf{x} > 0$$

(c)

$$R_1 : \mathbf{w}^T\mathbf{x} < 0$$

(d)

$$R_2 : \mathbf{w}^T\mathbf{x} < 0$$

(e)

$$L : \mathbf{w}^T\mathbf{x} = 0$$

Answer

(a), (d), (e)

Solution

All points (vectors) in the region R_1 make an acute angle with $\mathbf{w}$. Hence, $\mathbf{w}^T\mathbf{x} > 0$ for these points. All points (vectors) in the region R_2 make an obtuse angle with $\mathbf{w}$. Hence, $\mathbf{w}^T\mathbf{x} < 0$ for these points. All points on the line L make a right angle with $\mathbf{w}$. Hence, $\mathbf{w}^T\mathbf{x} = 0$ for these points.