Linear Algebra + Principal Component Analysis

31 March 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Recap and Today

- k-Means/GMM for Image
- Performance Metrics
- Linear Algebra (to prepare for Principal Component Analysis)

Classification Metrics

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Classification Metrics

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Is this enough??

Consider the scenario where 1000 people go for X-ray to detect Covid. They also undergo RT-PCR test for the same. Assuming RT-PCR gives the result with 100% accuracy, this is what was observed, what is the accuracy??

	X-ray	Covid +ve	Covid - ve
RT-PCR			
Covid +ve		50	20
Covid -ve		80	850

Classification Metrics

Consider another scenario where 1000 people go to a magician/future teller/optimistic person to detect Covid. They also undergo RT-PCR test for the same. Assuming RT-PCR gives the result with 100% accuracy, this is what was observed, what is the accuracy now??

	X-ray	Covid +ve	Covid - ve
RT-PCR			
Covid +ve		0	70
Covid -ve		0	930

Can we conclude that the magician is better than X-ray?

Classification Metrics

	X-ray	Covid +ve	Covid - ve
RT-PCR			
Covid +ve		True Positive(TP)	False Negative (FN)
Covid -ve		False Positive(FP)	True Negative (TN)

$\text{Precision(P)} = \frac{TP}{TP + FP}$ $\text{Recall(R)} = \frac{TP}{TP + FN}$

F-1 score is the harmonic mean of Precision and Recall

$\text{F-1 score} = \frac{2PR}{P+R}$

Classification Metrics

	X-ray	Covid +ve	Covid - ve
RT-PCR			
Covid +ve		True Positive(TP)	False Negative (FN)
Covid -ve		False Positive(FP)	True Negative (TN)

$\text{Precision(P)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ $\text{Recall(R)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$

F-1 score is the harmonic mean of Precision and Recall

F-1 score $= \frac{2PR}{P+R}$

What is meant by positive and negative when we have more than 2 classes?

PCA Motivation

- To reduce the number of dimensions in the data
- To visualize the data
- To avoid over fitting

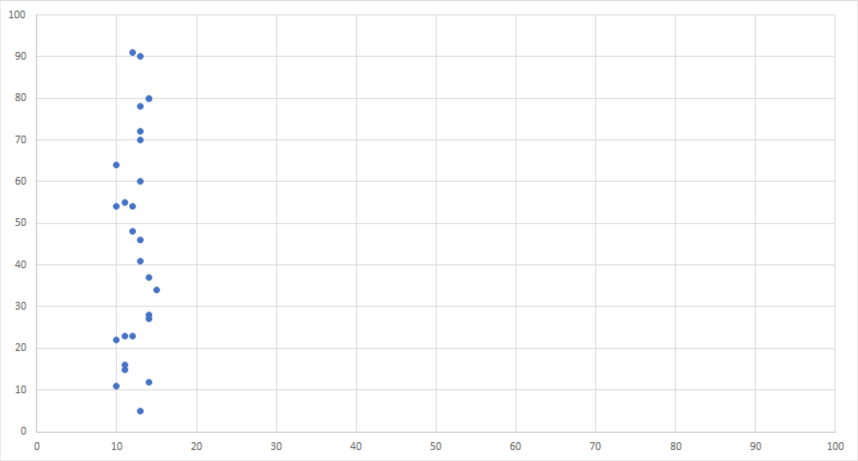
PCA Motivation

- To reduce the number of dimensions in the data
- To visualize the data
- To avoid over fitting

	Location	**City**	**Society**	**Ambience**	**Airport**
House 1	8	4	9	1	5
House 2	9	6	9	5	5
House 3	10	8	9	7	5
House 4	10	5	9	6	5
House 5	5	4	9	2	5
House 6	2	7	9	9	5
House 7	7	5	9	8	6
House 8	3	4	9	8	5
House 9	4	2	9	7	5
House 10	1	4	9	10	5

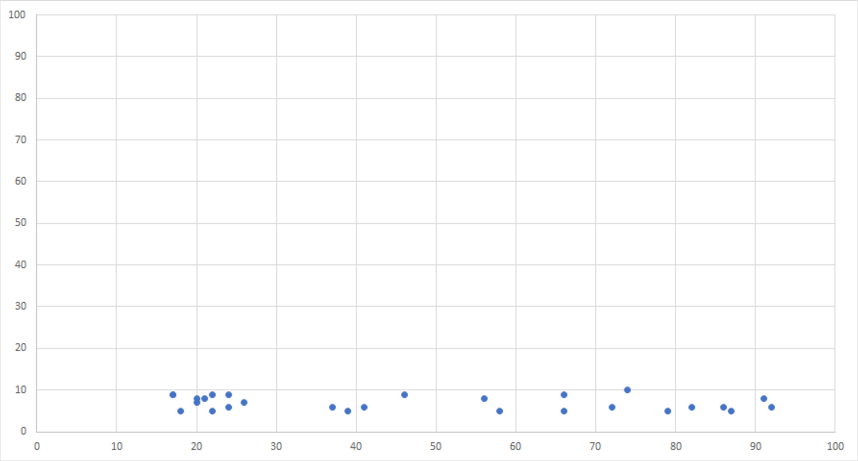
PCA Motivation

Which direction to skip?



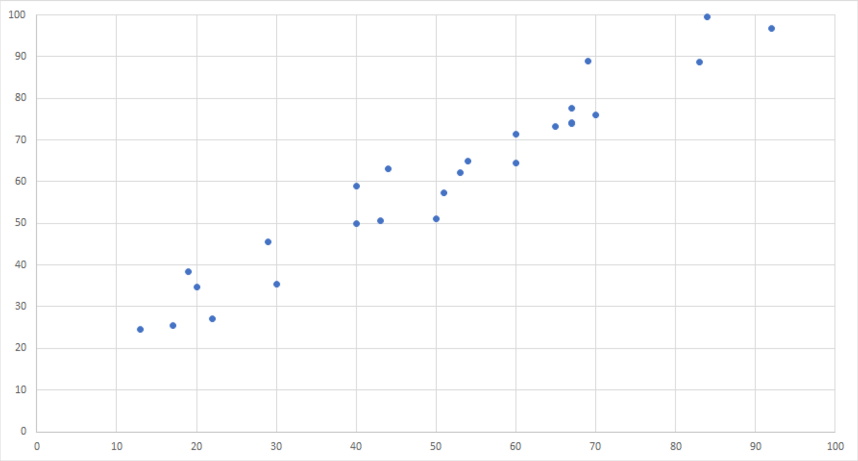
PCA Motivation

Which direction to skip?



PCA Motivation

Which direction to skip?



PCA Motivation

Looks like we are trying to capture as much variance as possible while reducing the number of dimensions.

Linear Algebra Revisit

What is a vector?

- a mathematical object that encodes a length and direction
- A vector is often represented as a 1-dimensional array of numbers, referred to as components and is displayed either in column form or row form
- Represented geometrically, vectors typically represent coordinates within a n-dimensional space

- Vector Space - Closed under addition and multiplication
- Vectors on line $y = 2 * x + 1$ form a vector space??
- Addition/Subtraction (Graphical Representation)
- Multiplication by a scalar
- Dot Product (Inner Product)
- Length/Magnitude
- Angle between two vectors - $\cos\theta = \frac{\vec{x}.\vec{y}}{||\vec{x}||.||\vec{y}||}$
- Dot product of perpendicular vectors = ??

- concept of zero vector

Consider m vectors in $\mathbb{R}^n$,
If $\alpha_1\vec{v_1} + \alpha_2\vec{v_2} + \cdots + \alpha_n\vec{v_n}$ implies
$\alpha_1 = \alpha_2 = \cdots = \alpha_m = \vec{0}$, then
the vectors are said to be linearly independent.

Matrices

In mathematics, a matrix (plural matrices) is a rectangular array or table of numbers, symbols, or expressions, arranged in rows and columns, which is used to represent a mathematical object or a property of such an object. - Wikipedia

- Matrix Size
- Representing any element of a matrix
- (1,n) and (m,1) matrices
- Square Matrix

Operations on Matrices

- Addition/Subtraction
- Multiplication of Matrix by a scalar
- Transpose of a Matrix
- Multiplication of two Matrices

Matrix as Linear Transformation

Types of Matrices

- Diagonal
- Identity
- Upper Triangular
- Symmetric
- Singular
- Mirror Matrix

Matrices contd.

- Inverse of a Matrix
- Trace of a Matrix
- Relationship of Eigenvalues with trace and Determinant
- Symmetric Matrix
- Orthogonal Matrix

Eigenvalues and Eigenvectors

Let M be a nxn matrix. A non-zero vector $\vec{X}$ is said to be an eigenvector of M corresponding to eigenvalue λ if

Spectral Decomposition

A n*n matrix can be written in terms of its eigenvalues and eigenvectors as follows -

$$M = \lambda_1 \vec{v_1} \vec{v_1}^T + \lambda_2 \vec{v_2} \vec{v_2}^T + ... + \lambda_n \vec{v_n} \vec{v_n}^T$$

$$\vec{x}^T M \vec{x} \geq 0$$

$$\vec{x}_{(n\times 1)}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

All the eigenvalues are greater than zero.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{pmatrix} 2x_1 + x_2 \\ x_1 + 2x_2 \end{pmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 + 2x_1x_2 + 2x_2^2$$

$$= x_1^2 + x_2^2 + (x_1 + x_2)^2$$

PCA - Objectives

We have a Data Matrix(D), whose dimension is $n * f$.

$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1f} \\ d_{21} & d_{22} & \dots & d_{2f} \\ \vdots & \vdots & \dots & \vdots \\ d_{(n-1)1} & d_{(n-1)2} & \dots & d_{(n-1)f} \\ d_{n1} & d_{n2} & \dots & d_{nf} \end{bmatrix}$$

We want to reduce the number of features to f_s.

PCA - Objectives

We have a Data Matrix(D), whose dimension is $n * f$.

$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1f} \\ d_{21} & d_{22} & \dots & d_{2f} \\ \vdots & \vdots & \dots & \vdots \\ d_{(n-1)1} & d_{(n-1)2} & \dots & d_{(n-1)f} \\ d_{n1} & d_{n2} & \dots & d_{nf} \end{bmatrix}$$

We want to reduce the number of features to f_s.

Let us solve a simpler problem first. Let us say we want to reduce the number of dimensions/features to just 1. Which is the best way to go about it?

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, ..., v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, ..., v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

Variance($D\vec{v}$) can be written as $\vec{v}^T \Sigma \vec{v}$, where Σ is the variance covariance matrix of D. (Try to prove this or atleast verify this)

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, ..., v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

Variance$(D\vec{v})$ can be written as $\vec{v}^T \Sigma \vec{v}$, where Σ is the variance covariance matrix of D. (Try to prove this or atleast verify this)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1f} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2f} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{(f-1)1} & \sigma_{(f-1)2} & \cdots & \sigma_{(f-1)f} \\ \sigma_{1f} & \sigma_{f2} & \cdots & \sigma_f^2 \end{bmatrix}$$

PCA Optimization problem

$\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_f$

Maximize - $\vec{v}^T \Sigma \vec{v}$ $\quad \vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_f$

such that $||\vec{v}|| = 1$

PCA Optimization problem

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $||\vec{v}|| = 1$

Σ can be shown to be positive semi-definite. Thus, all the eigenvalues are greater than zero, and all eigenvectors are orthogonal.

PCA Optimization problem

$\Sigma_{f \times f}$

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $||\vec{v}|| = 1$

Σ can be shown to be positive semi-definite. Thus, all the eigenvalues are greater than zero, and all eigenvectors are orthogonal.

$$\Sigma = \lambda_1 \vec{v_1} \vec{v_1}^T + \lambda_2 \vec{v_2} \vec{v_2}^T + ... + \lambda_f \vec{v_f} \vec{v_f}^T$$

$$\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_f$$

PCA Optimization problem

$\vec{v}_i \perp \vec{v}_j$ $\qquad$ $||\vec{v}_i|| = 1$

$(\alpha_1\vec{v}_1 + \alpha_2\vec{v}_2 + \dots + \alpha_f\vec{v}_f)^t$

Maximize - $\vec{v}^T(\lambda_1\vec{v_1}\vec{v_1}^T + \lambda_2\vec{v_2}\vec{v_2}^T + ... + \lambda_{f}\vec{v_{f}}\vec{v_{f}}^T)\vec{v}$

such that $||\vec{v}|| = 1$ $\qquad$ $\alpha_1^2 + \alpha_2^2 + \dots + \alpha_f^2 = 1$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$$

$$(1,2,3) = 1\cdot\begin{pmatrix}1\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\\0\end{pmatrix} + 3\begin{pmatrix}0\\0\\1\end{pmatrix}$$

$$\lambda_1 \vec{v}^T \vec{v_1} \vec{v_1}^T \vec{v} = \lambda_1 (\vec{v}^T \cdot \vec{v_1})^2$$

$$\lambda_1 \alpha_1^2 + \lambda_2 \alpha_2^2 + \ldots + \lambda_f \alpha_f^2$$

Maximize - $\vec{v}^T(\lambda_1 \vec{v_1}\vec{v_1}^T + \lambda_2 \vec{v_2}\vec{v_2}^T + ... + \lambda_n \vec{v_n}\vec{v_n}^T)\vec{v}$

such that $||\vec{v}|| = 1$ $\quad \alpha_1^2 + \alpha_2^2 + \_\_ + \alpha_f^2 = 1$

All $\vec{v_i}'s$ are unit vectors and orthogonal to each other, so the best choice for $\vec{v}$ is ??

$$\lambda_1 \geq \lambda_2 \ldots \geq \lambda_f$$

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

Fortunately, this happens if we keep going in the same sequence of the principal components.

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

Fortunately, this happens if we keep going in the same sequence of the principal components.

We can now try to maximize the sum of the variance terms in the 2 projected directions.
So, it looks like that the most logical thing to do is to pick up the eigenvector corresponding to the second largest eigenvector. This is what we do in PCA.

Reconstructing the data

One could start with the objective that we want to minimize the reconstruction error, and we would get the same result as what we have described above.

Thank you for your attention