Machine Learning

Dimensionality Reduction using PCA

Satishkumar L. Varma

Department of Information Technology
SVKM's Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai.
ORCID | Scopus | Google Scholar | Google Site | Website

Outline

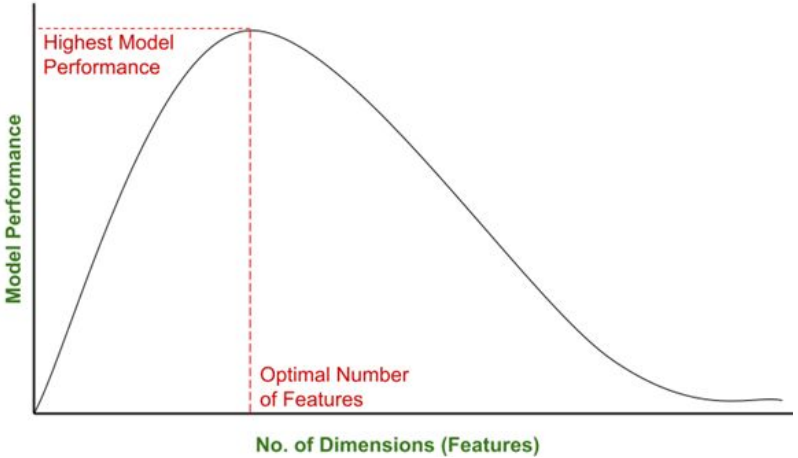
- Dimensionality Reduction
  - Dimensionality Reduction Techniques
    - Principal Components Analysis (Eigenvalues, Eigen vectors, Orthogonality)
    - Independent Component Analysis
    - Single Value Decomposition

Dimensionality Reduction

- Introduction
- Data Science and ML helps to solve several complex regression and classification problems.
- However, the performance of all these models depends on the input dataset.
- So, It is important to provide optimal dataset to ML models.
- Large dataset leads to increased computational demands
- Large dataset leads to overfitting.
- If we provide large dataset (with a large no. of features/columns) to ML models
  - it gives rise to the problem of overfitting,
  - wherein the model starts getting influenced by outlier values and noise.
  - This is called the Curse of Dimensionality.

Dimensionality Reduction

- Dimensionality Reduction is a statistical/ML-based technique
- It helps to reduce the number of features in our dataset.
- It helps to obtain a dataset with an optimal number of dimensions.
- It is useful in Feature Extraction
  - By reduce the number of dimensions by mapping a higher dimensional feature space to a lower-dimensional feature space.
- Effect of the change in model performance with the increase in the number of dimensions of the dataset.
- The model performance is best only at an option dimension, beyond which it starts decreasing.
- Technique of Dimensionality Reduction (Feature Extraction)
  - Principal Component Analysis (PCA)
  - Independent Component Analysis (ICA)
  - Singular Value Decomposition (SVD)
  - Linear Discriminant Analysis (LDA)



Dimensionality Reduction

- Application of DR wrt the type of datasets
  - Image Data: Compressing high-resolution images for efficient storage and retrieval.
  - Audio Data: Simplifying speech recordings for speaker recognition.
  - Audio Data: Reducing music audio data for genre classification.
  - Video Data: Compressing video for faster processing in surveillance systems.
  - Video Data: Reducing video sequences to key frames for action recognition.
  - Time Series Data: Reducing the complexity of financial time series data to reveal major trends.
  - Spatial Data: Reducing GPS data to visualization or summarizing geographic locations.
  - Graph Data: Reducing social network data to identify key communities.
  - Text Data: Efficient sentiment analysis.
  - Numerical Data: Reducing the number of variables in a dataset to predict housing prices.
  - Categorical and Ordinal Data: Reducing customer preferences for easier interpretation.

Principal Components Analysis

- PCA is unsupervised algorithms
- Used for data analysis, data compression, de-noising, reducing the dimension of data.
- It helps to reduce or eliminate similar data.
- PCA analysis reduces dimensionality without any data loss.
- It is a method of factor analysis.
- PCA is also called a dimensionality-reduction method.
- It helps you find out the most common dimensions of our dataset.
- It makes result analysis faster, easier and accurate.
- Dataset contains significant variables and dimensions.
- However, not all these variables will be critical.
- Some variables are primary key variables, whereas others are not.
- So, it helps to eliminate a few extra less important variables, without any data loss.
- Reduced data and dimensions help to achieve results faster and easy visualization.
- PCA helps to analyze all the dimensions and reducing them by maintaining the exact information.

Principal Components Analysis

- **Applications of PCA Analysis**
  - PCA in ML is used to visualize multidimensional data.
  - Healthcare data: To explore the factors important in increasing the risk of any chronic disease.
  - You can also use PCA to analyze patterns when you are dealing with high-dimensional data sets.
  - PCA helps to resize an image.
  - PCA is used to analyze stock data and forecasting data.
- **Domain of PCA Applications:**
  - Data cleaning;
  - Data preprocessing;
  - Denoise the information
  - Analyzing different dimensions.
  - Visualize multidimensional data.
  - Compress the information and transmit the same without any loss in quality.
  - Applied for face recognition, image identification, pattern identification.
  - PCA in ML helps in simplifying complex business algorithms.

Principal Components Analysis

- **When to use PCA**
  - To reduce the number of dimensions in the dataset.
  - To decide the critical variables.
  - To categorize the dependent and independent variables in the dataset.
  - To eliminate the noise components in our dimension analysis.
- Principle components with greater amount of variance are grouped under one category.
- Principle components with smaller variance are grouped under the second category.
- Note:
  - Vectors calculated in the Principal Component Method of factor analysis are not calculated at random.
  - All the calculated components can be combined as linear components and
  - so a single straight vector of each component helps identify difference in features much easier than ever.

Principal Components Analysis

- **Advantages of Principal Component Analysis**
  - Easy to calculate and compute.
  - Prevents predictive algorithms from data overfitting issues.
  - Speeds up machine learning computing processes and algorithms.
  - PIncreases performance of ML algorithms by eliminating unnecessary correlated variables.
  - Principal Component Analysis results in high variance and increases visualization.
  - Helps reduce noise that cannot be ignored automatically.
- **Disadvantages of Principal Component Analysis**
  - Sometimes, PCA is difficult to interpret.
    - Sometimes difficult to identify the most important features even after computing the PCs.
  - You may face some difficulties in calculating the covariances and covariance matrices.
  - Sometimes, the PCAs can be more difficult to read rather than the original set of components.

Principal Components Analysis

- Example: Steps to perform PCA
- 1. Standardize the range of continuous initial variables
- 2. Compute the covariance matrix to identify correlations
- 3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the Principal Components
- 4. Create a feature vector to decide the Principal Components
- 5. Recast the data along the Principal Components axes

Principal Components Analysis

- Example 1: Find the eigenvalues and eigenvectors of the matrix $A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$

The eigenvalues and eigenvectors are found by solving the eigenvalue probelm

$$AX = \lambda X \qquad X = \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{i.e.} \quad (A - \lambda I)X = 0.$$

Non-trivial solutions will exist if $\det(A - \lambda I) = 0$

that is, $\det\left\{\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right\} = 0, \quad \therefore \quad \begin{vmatrix} 1-\lambda & 0 \\ 1 & 2-\lambda \end{vmatrix} = 0,$

expanding this determinant: $(1-\lambda)(2-\lambda) = 0$. Hence the solutions for λ are: $\lambda = 1$ and $\lambda = 2$.

So we have found two values of λ for this 2×2 matrix A. Since these are unequal they are said to be **distinct** eigenvalues.

To each value of λ there corresponds an eigenvector. We now proceed to find the eigenvectors.

Principal Components Analysis

- Example 2: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$$

$\lambda = 1$ (smaller eigenvalue). Then our original eigenvalue problem becomes: $AX = X$.

$$\begin{aligned} x &= x \\ x + 2y &= y \end{aligned} \qquad \text{Simplifying} \qquad \begin{aligned} x &= x \\ x + y &= 0 \end{aligned}$$

All we can deduce here is that $x = -y$ $\quad \therefore \quad X = \begin{bmatrix} x \\ -x \end{bmatrix}$ for any $x \neq 0$

(We specify $x \neq 0$ as, otherwise, we would have the trivial solution.)

So the eigenvectors corresponding to eigenvalue $\lambda = 1$ are all proportional to $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, e.g. $\begin{bmatrix} 2 \\ -2 \end{bmatrix}$, $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ etc.

Sometimes we write the eigenvector in **normalised** form that is, with modulus or magnitude 1. Here, the normalised form of X is $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ which **is unique**.

Principal Components Analysis

- Example 1: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$$

Now we consider the larger eigenvalue $\lambda = 2$. Our original eigenvalue problem $AX = \lambda X$ becomes $AX = 2X$ which gives the following equations:

$$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 2\begin{bmatrix} x \\ y \end{bmatrix} \quad \text{i.e.} \quad \begin{aligned} x &= 2x \\ x + 2y &= 2y \end{aligned}$$

These equations imply that $x = 0$ whilst the variable y may take any value whatsoever (except zero as this gives the trivial solution).

Thus the eigenvector corresponding to eigenvalue $\lambda = 2$ has the form $\begin{bmatrix} 0 \\ y \end{bmatrix}$, e.g. $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$ etc.

The normalised eigenvector here is $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

In conclusion: the matrix $A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$ has two eigenvalues and two associated normalised eigenvectors:

$$\lambda_1 = 1, \qquad \lambda_2 = 2 \qquad X_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Principal Components Analysis

- Example 2: Obtain the eigenvalues and the eigenvectors of the symmetric 2 × 2 matrix

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

The characteristic equation for A is

$$(4-\lambda)(1-\lambda)+4=0 \qquad \text{or} \qquad \lambda^2 - 5\lambda = 0$$

giving $\lambda = 0$ and $\lambda = 5$, both of which are of course real and also unequal (i.e. distinct). For the larger eigenvalue $\lambda = 5$ the eigenvector $X = \begin{bmatrix} x \\ y \end{bmatrix}$ satisfy

$$\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5x \\ 5y \end{bmatrix} \quad \text{i.e.} \quad -x-2y=0, \qquad -2x-4y=0$$

Both equations tell us that $x = -2y$ so an eigenvector for $\lambda = 5$ is $X = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$ or any multiple of this. For $\lambda = 0$ the associated eigenvectors satisfy

$$4x - 2y = 0 \qquad -2x + y = 0$$

i.e. $y = 2x$ (from both equations) so an eigenvector is $Y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ or any multiple.

Principal Components Analysis

- Example 3: Prove that the X and Y matrix are orthogonal.

$$X^TY = [2, \quad -1] \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 2 \times 1 - 1 \times 2 = 2 - 2 = 0$$

$X^TY = 0$ means are X and Y are **orthogonal**.

Principal Components Analysis

- Example 4: Show that these three eigenvectors X, Y,Z are mutually orthogonal.

$$X = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \qquad Y = \begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \end{bmatrix} \qquad Z = \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix}$$

$$X^T Y = [1, \quad 0, \quad -1] \begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \end{bmatrix} = 1 - 1 = 0$$

$$Y^T Z = [1, \quad -\sqrt{2}, \quad 1] \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix} = 1 - 2 + 1 = 0$$

$$Z^T X = [1, \quad \sqrt{2}, \quad 1] \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = 1 - 1 = 0$$

Covariance Matrix

- It is a type of matrix that is used to represent the covariance values
  - between pairs of elements given in a random vector.
- It can also be referred to as the variance covariance matrix.
- This is because the variance of each element is represented along the main diagonal of the matrix.
- It is always a square matrix.
- It is positive semi-definite, and symmetric.
- This matrix is very useful in stochastic modeling and principal component analysis.
- It displays the variance exhibited by elements of datasets and covariance between a pair of datasets.
- Variance is a measure of dispersion and can be defined as the spread of data from the mean of the given dataset.
- Covariance is calculated between two variables and is used to measure how the two variables vary together.
- It is a square matrix where diagonal elements represent the variance and the off-diagonal elements represent the covariance.
- The covariance between two variables can be positive, negative, and zero.
- A positive covariance indicates that the two variables have a positive relationship whereas negative covariance shows that they have a negative relationship.
- If two elements do not vary together then they will display a zero covariance.

Covariance Matrix

- Example: Covariance Matrix 2 × 2

Covariance Matrix Example

$$\begin{bmatrix} Var(x_1) & \cdots\cdots & Cov(x_n,x_1) \\ \vdots & \cdot & \vdots \\ \vdots & \cdot & \vdots \\ Cov(x_n,x_1) & \cdots\cdots & Var(x_n) \end{bmatrix}$$

Population Var: $var(x) = \frac{\sum_1^n (x_i-\mu)^2}{n}$ Sample Var: $var(x) = \frac{\sum_1^n (x_i-\bar{x})^2}{n-1}$

Population Cov: $cov(x, y) = \frac{\sum_1^n (x_i-\mu_x)(y_i-\mu_y)}{n}$ Sample Cov: $cov(x, y) = \frac{\sum_1^n (x_i-\bar{x})(y_i-\bar{y})}{n-1}$

μ = mean of population data. n = number of observations in the dataset.

$\bar{x}$ = mean of sample data. x_i = observations in dataset x.

Dataset	X	Y
	3	6
	2	4
Sum (∑)	5	10
Mean ($\bar{x}$)	2.5	5

	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
	0.5	1	0.5	0.25	1
	-0.5	-1	0.5	0.25	1
Sum (∑)	0	0	1	0.5	2
Sample Var				0.5	2
Sample Cov			1		

Covariance Matrix 2 × 2

var(x)	cov(x, y)
cov(x, y)	var(x)

Covariance Matrix 2 × 2

0.5	1
1	2

Covariance Matrix

- Example:
- Covariance Matrix 3 × 3

$$\begin{bmatrix} Var(x_1) & \cdots\cdots & Cov(x_n,x_1) \\ \vdots & \ddots & \vdots \\ Cov(x_n,x_1) & \cdots\cdots & Var(x_n) \end{bmatrix}$$

Population Var: $var(x) = \frac{\sum_1^n (x_i-\mu)^2}{n}$ Sample Var: $var(x) = \frac{\sum_1^n (x_i-\bar{x})^2}{n-1}$

Population Cov: $cov(x, y) = \frac{\sum_1^n (x_i-\mu_x)(y_i-\mu_y)}{n}$ Sample Cov: $cov(x, y) = \frac{\sum_1^n (x_i-\bar{x})(y_i-\bar{y})}{n-1}$

μ = mean of population data. n = number of observations in the dataset.

$\bar{x}$ = mean of sample data. x_i = observations in dataset x.

Dataset	Math (X)	Science (Y)	English (Z)
1	70	80	50
2	65	30	40
3	90	70	60
Sum (∑)	**225**	**180**	**150**
Mean ($\bar{x}$)	**75**	**60**	**50**

	$x_i-\bar{x}$	$y_i-\bar{y}$	$z_i-\bar{z}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})(z_i-\bar{z})$	$(y_i-\bar{y})(z_i-\bar{z})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(z_i-\bar{z})^2$
	-5	20	0	-100	0	0	25	400	0
	-10	-30	-10	300	100	300	100	900	100
	15	10	10	150	150	100	225	100	100
Sum (∑)	**0**	**0**	**0**	**350**	**250**	**400**	**350**	**1400**	**200**
Sample Var							**175**	**700**	**100**
Sample Cov				**175**	**125**	**200**			

Covariance Matrix 3 × 3

var(x)	cov(x, y)	cov(x,z)
cov(x, y)	var(y)	cov(y,z)
cov(x,z)	cov(y,z)	var(z)

Covariance Matrix 3 × 3

175	175	125
175	700	200
125	200	100

Covariance Matrix

- **Properties of Covariance Matrix**
- It is a very important tool used by data scientists to understand and analyze multivariate data.
- It depicts the variance of datasets and covariance of a pair of datasets in matrix format.
- The diagonal elements represent the variance of a dataset and
  - the off-diagonal terms give the covariance between a pair of datasets.
- The variance covariance matrix is always square, symmetric, and positive semi-definite.
- Extremely useful properties of Covariance Matrix:
  - It is always a square matrix. That is # of rows of the matrix will be equal to the number of columns.
  - The matrix is symmetric. Suppose M is the cov.mat. then MT = M.
  - It is positive semi-definite.
    - Let u be a column vector, uT is the transpose of that vector and M be the cov.mat. then uTMu ≥ 0.
  - All eigenvalues of the variance covariance matrix are real and non-negative.

Covariance Matrix

- **Interpretation of the variance covariance matrix:**
- 1) The diagonal elements 500, 340 and 800 indicate the variance in data sets X, Y and Z respectively.
  - Y shows the lowest variance whereas Z displays the highest variance.
- 2) The covariance for X and Y is 320.
  - As this is a positive number it means that when X increases (or decreases) Y also increases (or decreases)
- 3) The covariance for X and Z is –40.
  - As it is a negative number it implies that when X increases Z decreases and vice – versa.
- 4) The covariance for Y and Z is 0.
  - This means that there is no predictable relationship between the two data sets.

Covariance Matrix 3 × 3

var(x)	cov(x, y)	cov(x,z)
cov(x, y)	var(y)	cov(y,z)
cov(x,z)	cov(y,z)	var(z)

Covariance Matrix 3 × 3

500	320	–40
320	340	0
–40	0	800

References

Text books:

1. Ethem Alpaydin, “Introduction to Machine Learning , 4th Edition, The MIT Press, 2020.
2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012."
3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.
4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1ed, O'reilly, 2016.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012."

Reference Books:

6. Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd Edition, Shroff/O'Reilly, 2019.
7. Witten Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal., “Data Mining: Practical machine learning tools and techniques”, 1st Edition, Morgan Kaufmann, 2016.
8. Han, Kamber, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.
9. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, “Foundations of Machine Learning”, 1ed, MIT Press, 2012.
10. H. Dunham, “Data Mining: Introductory and Advanced Topics”, 1st Edition, Pearson Education, 2006.

Thank You.