23MAT204 PCA Assignment

Adithya Nair

August 17, 2024

Contents

1 PCA IEEE Report

1.1 What is PCA?

In a nutshell, what Principal Component Analysis does is reduce the dimensionality of a set of data points by linearly projecting them onto a lower-dimensional space where the reconstruction eror is as minimal as possible.

1.2 Problem Statement

When working with data, there are times when we encounter a situation where we obtain data with a dimensionality that is too high to be visualized or analyzed(by dimensionality we mean the number of dimensions of the data points.)

Principal Component Analysis is a method by which data points with a higher dimensionality are projected onto a linear subspace, with as minimal of a reconstruction error and projection error as possible.

Focus on just the one point x and its projection $\mathrm{x}_{\backslash|}$. d is the distance of x from the origin, r is the distance of x from $x_{\|}$ in the subspace, and v is the distance of $x_{\|}$

1.3 Projection And Reconstruction Error

Principal Component Analysis works by projection, it takes a higher dimensional space and linearly projects them to a lower dimensional subspace, in a manner that makes sure that the reconstruction of this data gives as minimal of an error as possible.

1.3.1 Example Using A 2D-1D Example

Let's say we have a set of data points in 2-dimensional space, and our goal is to place these data points in a line while mirroring their relationship to every other data point as closely as possible

Let's represent these data points as a vector $\vec{x}$, We choose the unit vector $\hat{v}$(Which is the unit vector that provides the least reconstruction error upon projection for x). The rest of this report deals with how to find this unit vector but assume that we have found the unit vector $\hat{v}$ with the aforementioned property

Then the projection $x_{||}$ is done by,

$$x_{||} := (\hat{v} \cdot x)\hat{v}$$

Where $v^T x$ gives us the projection value and v gives us the direction, we can write $v^T x$ as y to get,

$$x_{||} = y\hat{v}$$

This projection is clearly not the same as the original data points, which means that there is an error to account for between the projected points and the original data poitns. To compute this error, we take the mean squared distance between the projected data points and the original data points(although we can take just the distance, the squared distance is convenient for computation)

$$\text{Reconstruction Error} = \Sigma_{i=1}^{n} \|\vec{x} - \vec{x_{||}}\|^2$$

Where n is the number of data points. Since we're dealing with vectors, we can write an expression for the individual components as well,

$$\text{Reconstruction Error} = \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} (x_j - x_{||_j})^2$$

Where m is the dimensionality of the data points.

1.4 Reconstruction Error And Variance

1.5 Covariance

Covariance is a statistical tool that can define the behavior of two random data points. If a pair of data points has a positive covariance, then the data points 'vary' with each other or change in the same direction. If they have a negative covariance, then the data points don't 'change' in the same direction.

1.6 Covariance Matrix

How do we determine the direction of maximal variance. We can determine the individual component variances. For a given vector

$$x = (x_1, x_2)^T$$

, then the variances of the first and second component can be written as

$$C_{11} := \langle x_1 x_1 \rangle$$

,

$$C_{22} := \langle x_2 x_2 \rangle$$

, (the angle brackets indicate 'averaging' over all data points)

A relatively large value of C_{11} then the entire set of data points varies closely to the $[1, 0]^T$ unit vector. Similarly, for C_{22}, the data points varies most along the $[0, 1]^T$ direction the most.

In other words, the covariance matrix contains the direction of maximum variance. This is what we need to make sure the reconstruction error is minimal as shown earlier.

1.7 PCA By Diagonalizing The Covariance Matrix

To find the direction with which the data points vary most in, we take the covariance matrix and find the axis along which it is most varying. The question might arise, how do we go about doing this for when the axis of variation isn't obvious? We diagonalize the covariance matrix which gives us diagonal elements. The element with the largest value corresponds to the direction that the data varies most in.

Diagonalization is done by finding the eigenvalues and eigenvectors of the matrix. The eigenvector corresponding to the largest eigenvalue is the direction which our data must be projected on to yield the least amount of reconstruction error.

2 References

- https://arxiv.org/pdf/2403.15112
- https://arxiv.org/pdf/2402.15527
- https://medium.com/@anabelenmanjavacas/dimensionality-reduction-and-pca-23dbd7d6f367
- https://ieeexplore.ieee.org/document/10511242

3 Questions

3.1 Why Covariance specifically?

4 References

StatQuest Lecture Notes On Principal Component Analysis