PCA

Abbavaram Gowtham Reddy

March 21, 2020

Problem statement:

Given a set of N data points that are in R^D(i.e., N data points in D dimensional space, each dimension represents a feature), project those points onto lower dimenstion(possibly 2 or 3) for better visualisation. Data is represented by a NxD matrix corresponding to N data points and D features for each data points

Mathematical Background

Projections:

Projecting a vector x that belongs to R^n on to a subspace V of dimension m(R^m) where $n \geq m$ is given by

$$P_V^x = A(A^TA)^{-1}A^Tx$$

where A is a vector whose columns are the basis vectors of subspace V. For more details how to prove the above equation, visit this link.

If columns of A are orthonormal, then number of columns of A are less than or equal to number of rows of A(to match the rank property, rank(A) $\leq$ min(number of columns, number of rows)). Then A^TA is identity, then projection simplifies to

$$P_V^x = AA^Tx \tag{1}$$

Mean of a dataset:

In given problem setting, we need to find the mean of the data set for great mathematical simplification during PCA evaluation. we have NxD data matrix. Mean of the dataset results in a vector of R^D, that is mean is calculate across D features. We subtract this mean from each data point to move the entire dataset centered at zero with out loss of generality.

Orthogonality, Orthonormality:

Two vectors x, y are orthogonal to each other if $x^Ty = 0$

Two vectors x, y are orthonormal to each other if $x^Ty = 0$ and $||x|| = 1$ and $||y|| = 1$

A square matrix A is an orthogonal matrix if $A^T = A^{-1}$. That is columns/rows of A are orthonormal.

Linear Combination, Linear Independence:

Let $X = x_1, x_2, x_3, \ldots, x_n$ are n vectors, then $\sum a_i x_i$ for $i = 1, 2, ...n$ where a_i are some scalars, is called linear combination of vectors of X.

If the only solution to the equation $a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0$ is $a_1 = a_2 = \cdots = a_n = 0$ then $x_1, x_2, \ldots, x_n$ are linearly independent vectors.

Eigen Values and Eigen Vectors

For a square matrix A, if there exists a vector x such that $Ax = \lambda x$ then we call λ an eigen value of A and we call x an eigen vector of A corresponding to the eigen value λ

Mathematics of PCA

Let $X = x_1, x_2, x_3, \ldots, x_N$ are N vectors(data points), and $x_i \in R^D$(D features).

Now consider some orthonormal basis of R^D with vectors $b_1, b_2, b_3, \ldots, b_D$. Then any vector x_n can be represented as a linear combination of these basis vectors as follows

$$x_n = \sum_{i=1}^{D} \beta_{in} b_i \tag{2}$$

for some scalars β_{ij}. Since $b_i^{'s}$ are orthonormal, we can write β_{in} as follows

$$\beta_{in} = x_n^T b_i$$

Now let B be the matrix formed by placing our orthonormal basis vectors as columns. i.e., $B = [b_1 b_2 b_3, \ldots, b_D]$ Then from equation 1, the projection of any vector x_n onto the subspace spanned by column of B is

$$\tilde{x_n} = BB^T x_n$$

Here it is easy to see that the matrix-vector product $B^T x_n$ is a vector consisting of coordinate of $\tilde{x_n}$ with respect to the basis $b_1, b_2, b_3, \ldots, b_D$.

Now let us rewrite $\tilde{x_n}$ as follows

$$\tilde{x_n} = \sum_{i=1}^{M} \beta_{in} b_i + \sum_{i=M+1}^{D} \beta_{in} b_i \in R^D$$

$$\tilde{x_n} \approx \sum_{i=1}^{M} \beta_{in} b_i \tag{3}$$

In the above equation, we are interested in M such that first term in the right hand side has as much variability/information as possible. we call such M-dimensional subspace of D-dimensional subspace, "Principle subspace". That is $b_1, b_2, \ldots, b_M$ spans principle subspace. So we ignore the second term in the right hand side.

Now the problems becomes finding β_{in} and b_i in the above equation such that average reconstruction error between x_n and $\tilde{x_n}$ is minimized. Mathematically, we need to minimize the following term.

$$J = \frac{1}{N} \sum_{n=1}^{N} ||x_n - \tilde{x_n}||^2 \tag{4}$$

By minimizing the above term, the parameters of interest β_{in} and b_i can be found by taking partial derivatives with respect to parameters and equating to zero.

$$\frac{\partial J}{\partial \beta_{in}} = \frac{\partial J}{\partial \tilde{x_n}} \frac{\partial \tilde{x_n}}{\partial \beta_{in}}$$

$$\frac{\partial J}{\partial \tilde{x_n}} = \frac{-2}{N} (x_n - \tilde{x_n})^T$$

Since data is centered at zero, $E[X] = 0$ and $b_1, b_2, \ldots, b_M$ form an ONB.

$$\frac{\partial \tilde{x_n}}{\partial \beta_{in}} = b_i \quad \forall i = 1, 2, \ldots, M$$

$$\frac{\partial J}{\partial \beta_{in}} = \frac{-2}{N}(x_n - \tilde{x_n})^T b_i$$

From equation 3,

$$\frac{\partial J}{\partial \beta_{in}} = \frac{-2}{N}(x_n - \sum_{i=1}^{M} \beta_{in} b_i)^T b_i$$

$$\frac{\partial J}{\partial \beta_{in}} = \frac{-2}{N}(x_n^T b_i - \beta_{in} b_i^T b_i)$$

equating to zero,

$$\frac{\partial J}{\partial \beta_{in}} = \frac{-2}{N}(x_n^T b_i - \beta_{in} b_i^T b_i) = 0$$

$$\beta_{in} = x_n^T b_i \quad \forall i = 1, 2, \ldots, M \tag{5}$$

The above β_{in} are coordinates of $\tilde{x_n}$ with respect to projected subspace of M-dimensions.

From equation 3,

$$\tilde{x_n} = \sum_{i=1}^{M} \beta_{in} b_i$$

by using optimal β_{in} from above,

$$\tilde{x_n} = \sum_{i=1}^{M} (x_n^T b_i) b_i$$

The term which is in the open brackets in the above equation is dot product, and a scaler, we can manipulate it to get the following.

$$\tilde{x_n} = \sum_{i=1}^{M} (b_i b_i^T) x_n \tag{6}$$

here $\sum_{i=1}^{M} (b_i b_i^T)$ is the projection matrix with respect to orthonormal basis $b_1, b_2, \ldots, b_M$ and also we can write x_n as,

$$x_n = \sum_{i=1}^{M} (b_i b_i^T) x_n + \sum_{i=M+1}^{D} (b_i b_i^T) x_n \tag{7}$$

from the above two equations we can observe that x_n and $\tilde{x_n}$ differ by the second additive term in equation 7. and

$$x_n - \tilde{x_n} = \sum_{i=M+1}^{D} (b_i b_i^T) x_n$$

$$x_n - \tilde{x_n} = \sum_{i=M+1}^{D} (b_i^T x_n) b_i \tag{8}$$

using equation 8 in equation 4 gives the following.

$$J = \frac{1}{N} \sum_{n=1}^{N} || \sum_{i=M+1}^{D} (b_i^T x_n) b_i ||^2$$

$$J = \frac{1}{N} \sum_{n=1}^{N} || \sum_{i=M+1}^{D} (b_i^T x_n) b_i ||^2$$

Using the properties of orthonormal basis, our expression simplifies to

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D}(b_i^T x_n)^2$$

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D}(b_i^T x_n x_n^T b_i)$$

Rearranging terms

$$J = \sum_{i=M+1}^{D} b_i^T (\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T) b_i$$

Now the term inside the open brackets is the data covariance matrix S! Now rewriting the loss term using S

$$J = \sum_{i=M+1}^{D} b_i^T S b_i$$

Rearranging terms

$$J = trace((\sum_{i=M+1}^{D} b_i b_i^T)S)$$

In the above equation we have a projection matrix. The above expression says that, loss is the projection of covariance matrix on the subspace that is the orthogonal complement of principle subspace. So we need to minimize the variance of the data that lies outside of the principle subspace. In other words, we need to maximize the variance of the data that lies on the principle subspace. From the following,

$$J = \sum_{i=M+1}^{D} b_i^T S b_i$$

We need to find basis of dimension M that minimize J subject to some constraints.

$$\underset{\forall b_i, ||b_i||_2=1}{\arg\min} \sum_{i=M+1}^{D} b_i^T S b_i$$

Using Lagrangian multipliers, there exists $\lambda_{M+1}, \lambda_{M+2}, \ldots, \lambda_D$ such that the solution to the above is given by

$$Minimize \quad \sum_{i=M+1}^{D} b_i^T S b_i + \sum_{i=M+1}^{D} \lambda_i ||b_i||_2^2$$

setting the derivative with respect to b_i to zero gives,

$$Sb_i = \lambda_i b_i$$

That is the solution is given by eigen vectors b_i and corresponding eigen values λ_i.

And the reconstruction error is given By

$$J = \sum_{i=M+1}^{D} b_i^T S b_i = \sum_{i=M+1}^{D} b_i^T \lambda_i b_i = \sum_{i=M+1}^{D} \lambda_i$$

Clearly to minimize reconstruction error, we need to discard the $D - M$ directions that have the smallest eigenvalue of Covariance matrix.