A Tutorial on Principal Component Analysis

1. Preface

PCA (Principal Component Analysis) is a common method of data analysis. The original data is linearly transformed into a set of representations of which each dimension is linear independent. This can be used to extract the main feature components of the data. PCA is often used to reduce the dimension of high-dimensional data.

The aim of this tutorial is to introduce the basic mathematical theories of PCA and help readers understand the working mechanism of PCA. Additionally, I do not intend to write this tutorial as a pure mathematical article, but I hope to describe the mathematical principles of PCA in an intuitive and easy-to-understand way. Therefore, the whole article will not introduce strict mathematical derivation.

2. Introduction of Dimension Reduction of Data

In general, data is represented as vectors in the fields of data mining and machine learning. For example, the traffic and transactions of an Amazon store in the whole year of 2018 can be regarded as a set of records, in which the data of each day is a record in the following format:

$$(\text{Date}, \text{Pageviews}, \text{Number of visitors}, \text{Orders}, \text{Number of transactions}, \text{Amount of transactions})$$

Among the data, Date is a record mark rather than a measure, while measurements are mostly concerned in data mining. Therefore, if we ignore the Date, we will get a set of records. Each record can be expressed as a five-dimensional vector, one of which looks like this:

$$(500, 240, 25, 13, 2312.15)^T$$

Note that transpose is used here, because column vectors are customarily used to represent a record. This criterion will be followed later in this tutorial.

It is clear that we can analyze these five-dimensional vectors. However, it is known that the complexity of many data mining algorithms is closely related to the dimension of the data, even exponentially related to the dimension. The five-dimensional data may not matter in this case, but it is not uncommon to deal with tens of thousands or even hundreds of thousands of dimensions in actual learning problems. In such a case, the resource consumption of data mining and machine learning is unacceptable, so we must reduce the dimensionality of the data.

Although dimension reduction means the loss of information, in view of the correlation of the actual data itself, we can still find ways to reduce the loss of information while doing dimension reduction.

For example, if there are two columns of *M* and *F* in one school roll data, where the value of *M* column is 1 for male and 0 for female, while *F* column is 1 for female and 0 for male. If we count all of the student status data, it is not difficult to find that for any record, when *M* is 1, *F* must be 0, and vice versa, when *M* is 0, *F* must be 1. In this case, we are able to remove *M* or *F* without losing any information in actual, because the removed column can be completely reconstructed from the other.

The above example is an extreme situation, which may not occur in reality, but similar situations are still common. For example, the data of the Amazon store above, we can know from experience that Pageviews and Number of visitors tend to have a strong correlation, so do Orders and Number of transactions.

This situation shows that if we delete one of the indicators of Pageviews and Number of visitors, it is expected that we will not lose too much information. Therefore, we can delete one of them to reduce the complexity of data mining algorithms.

The above is a simple description of dimension reduction, which can help to intuitively understand the motivation and feasibility of dimension reduction, but it does not have operational guiding significance. For example, which column should we delete so that we lose the least information? Or is it not simply to delete several columns, but to change the original data into fewer columns by some transformations and minimize the loss of the information? How can we quantitatively measure the loss of the information or determine the specific steps of dimension reduction based on the original data?

To answer the above questions, we need to mathematically and formally discuss the dimension reduction problems. PCA is a dimension reduction method which has strict mathematical basis and has been widely used. Now I will analyze problems step-by-step rather than directly describing PCA. Let's invent PCA again.

3. Vector Representation and Base Transform

Since the data we are dealing with is represented as a set of vectors, it is necessary to study some of the mathematical properties of vectors. These mathematical properties will be the theoretical basis for the derivation of PCA.

3.1. Inner Product and Projection

Let's start with a simple vector operation: inner product. The inner product of two vectors of the same dimension is defined as:

$$(a_1, a_2, \cdots, a_n) \cdot (b_1, b_2, \cdots, b_n)^T = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

The inner product maps two vectors into a real number. Its calculation is easy to understand, but its significance is not obvious. Next, we will analyse the geometric meaning of the inner product. Assume that A and B are two n-dimensional vectors. We know that n-dimensional vectors can be equivalent to a directed line segment emitted from the origin in n-dimensional space. For simplicity, we assume that A and B are two-dimensional vectors, then $A = (x_1, y_1)$, $B = (x_2, y_2)$. On the two-dimensional plane, A and B can be represented by two directed lines from the origin, as shown in the following figure:

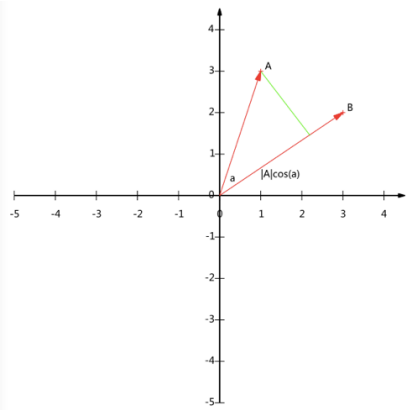

Figure 1. Geometric Representations of Vectors

Now let's draw a vertical line from point A to the straight line where point B is. We know that the intersection of the vertical line and B is called the projection of A on B. If the angle between A and B is α, then the vector length of the projection is $|A|cos(\alpha)$, where $|A| = \sqrt{x_1^2 + y_1^2}$ is called the modulus of vector A, which is the scalar length of line A.

Note that here we specifically distinguish between vector length and scalar length. Scalar length is always greater than or equal to 0, and the value is the length of the line segment; while vector length may be negative, its absolute value is the length of the line segment, and the symbol depends on its direction are the same as or opposite to the standard direction.

We still can't see what the inner product has to do with this, but if we express the inner product as another familiar form:

$$A \cdot B = |A||B|cos(\alpha)$$

Now it seems a little clearly that the inner product of A and B is equal to the projection length of A to B multiplied by the modulus of B. Furthermore, if we assume that the modulus of B is 1, that is, let $|B| = 1$, then it becomes:

$$A \cdot B = |A|cos(\alpha)$$

That is to say, if the modulus of vector B is 1, the inner product of A and B is equal to the vector length of the projection of A to B in the straight line. This is a geometric interpretation of the inner product, which is also the first important conclusion we get. This conclusion will be used repeatedly in the following derivation.

3.2. Base

Let's continue to discuss vectors in two-dimensional space. As mentioned above, a two-dimensional vector can correspond to a directed line segment starting from the origin in the two-dimensional Cartesian coordinate system. For example, the following vector:

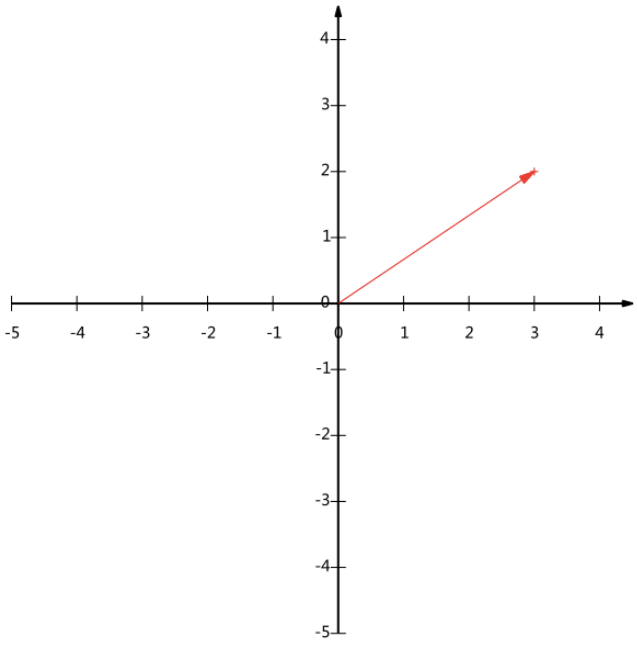

Figure 2. A Vector in the Cartesian Coordinate System

In algebraic representation, we often use point coordinates at the end of a line segment to represent vectors. For example, we are familiar with the above vector expressed as $(3, 2)$.

However, we often neglect that only one representation like $(3, 2)$ cannot express a vector accurately. It is not difficult to find that what 3 actually represents here is that the projection value of the vector on the x-axis is 3, and the projection value on the y-axis is 2. That is to say, we implicitly introduced a definition that based on the standard vectors whose lengths in the positive direction on the x-axis and y-axis are 1, a vector $(3, 2)$ actually means that the projection on the x-axis is 3 and the projection on the y-axis is 2. Notice that the projection is a vector, so it can be negative.

More formally, vectors (x, y) actually represent linear combinations:

$$x(1, 0)^T + y(0, 1)^T$$

It is not difficult to prove that all two-dimensional vectors can be represented as such a linear combination. Here $(1, 0)$ and $(0, 1)$ are called a set of bases in two-dimensional space.

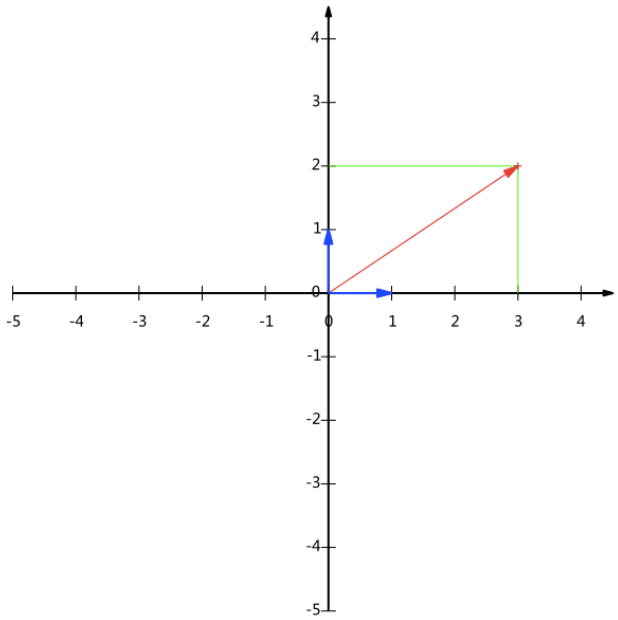

Figure 3. A Vector Represented as a Linear Combination of Bases

Therefore, in order to accurately describe vectors, we first need to determine a set of bases, and then give the projection values on each line on which the bases are located. But we often omit the first step, and the default is based on $(1, 0)$ and $(0, 1)$.

The reason why we choose $(1, 0)$ and $(0, 1)$) as the bases by default is that they are unit vectors in the positive direction of the x-axis and y-axis respectively, so that point coordinates and vectors on the two-dimensional plane correspond one by one. In fact, however, any two linearly independent two-dimensional vectors can be regarded as a set of bases. The linearly independent two-dimensional vectors can be visually regarded as two vectors that are not on a straight line in a two-dimensional plane.

For example, $(1, 1)$ and $(-1, 1)$ can also be a set of bases. In general, we want the modulus of the base to be 1, because we can see from the meaning of the inner product that if the modulus of the base is 1, then it is convenient to multiply the bases with the vector and directly obtain the coordinates on the new base system. In fact, for any vector, we can always find a vector whose modulus is 1 in the same direction, as

long as the two components are divided by the modulus. For example, the base above can be changed to $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ and $\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

Now, we want to obtain the coordinate of $(3, 2)$ on the new base, that is, the projection vector values in the two directions. Then according to the geometric meaning of the inner product, we only need to calculate the inner product of $(3, 2)$ with two bases separately. It is not difficult to get the new coordinate $\left(\frac{5}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$. The figure below shows the new base and $(3, 2)$ on the new base coordinate system:

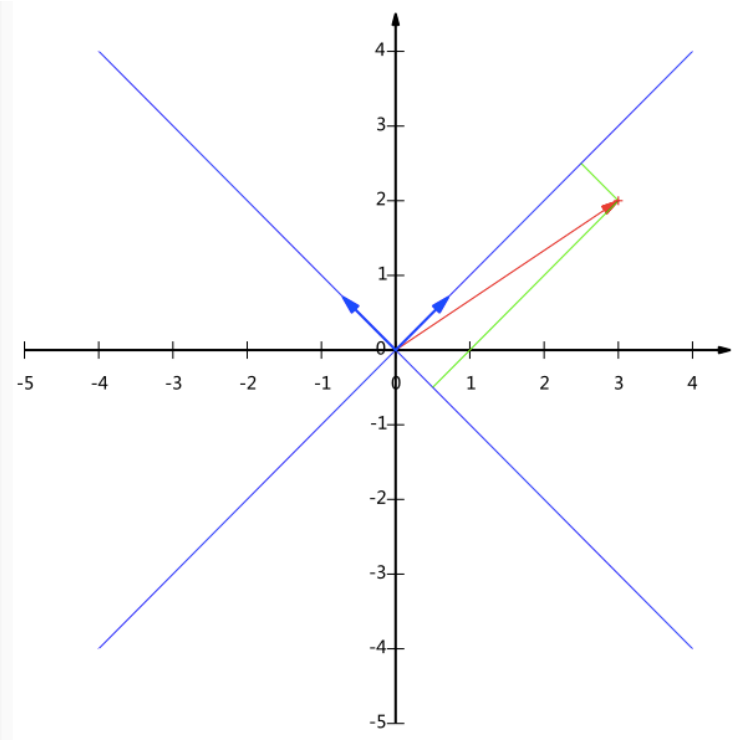

Figure 4. Linear Combinations of Vectors in the New Base Coordinate System

In addition, it should be noted that the examples here are orthogonal (That is the inner product is 0 or intuitively perpendicular to each other), but the only requirement that can be a set of bases is linearly independent. Therefore, the non-orthogonal ones can also be a set of bases. However, because of some properties of orthonormal basis, the bases are generally orthogonal.

3.3. Matrix Representation of Base Transformation

In this section, we will find an easy way to represent the base transform. Consider the same example as above. Converting $(3, 2)$ to the coordinates in the new base is using $(3, 2)$ to do the inner product with the first base to get the first new coordinate component, and then using $(3, 2)$ with the second base to get the second new coordinate component. In fact, we can simply represent this transformation in the form of matrix multiplication:

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

The two rows of the matrix are two bases respectively, multiplied by the original vector, and the result is just the coordinates in the new base. It can be slightly generalized that if we have *m* two-dimensional vectors, we can obtain the coordinates of all the vectors in the new base by arranging the two-dimensional

vectors into a two-row and *m*-column matrix and then multiplying the base matrix. For example, transform vectors $(1,1)$, $(2,2)$ and $(3,3)$ into the new base coordinate system, it can be represented as:

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 2/\sqrt{2} & 4/\sqrt{2} & 5/\sqrt{2} \\ 0 & 0 & 0 \end{pmatrix}$$

The base transform of a set of vectors is then cleanly represented as the multiplication of the matrices.

Generally, if we have *M N*-dimensional vectors and want to transform them into new spaces represented by *R N*-dimensional vectors, then we first form *R* bases into matrix *A* by row, and then group the vectors into matrix *B* by column. Then the product *AB* of the two matrices is the result of the transformation, wherein the *m*th column of AB is the result of the transformation of the *m*th column in *A*.

The mathematical expression is:

$$\begin{pmatrix} p_1 \\ \vdots \\ p_R \end{pmatrix}(a_1 \quad \cdots \quad a_M) = \begin{pmatrix} p_1 a_1 & \cdots & p_1 a_M \\ \vdots & \ddots & \vdots \\ p_R a_1 & \cdots & p_R a_M \end{pmatrix}$$

In the above equation, p_i is a row vector representing the *i*th base, and a_j is a column vector representing the *j*th raw data.

It is important to note that *R* can be less than *N*, and *R* determines the dimension of the transformed data. That is, we can transform an *N*-dimensional data into a lower-dimensional space, and the transformed dimension depends on the number of bases. Therefore, the representation of such matrix multiplication can also represent a dimension reduction transformation.

Finally, the above analysis also finds a physical explanation for matrix multiplication: the meaning of multiplication of two matrices is to transform each column of the vector in the right matrix into the space represented by each row vector in the left matrix. More abstractly, a matrix can represent a linear transformation.

4. Covariance Matrix and Optimization

We have discussed above that choosing different bases can give different representations to the same set of data, and if the number of bases is less than the dimension of the vector itself, the dimension reduction effect can be achieved. But we have not answered one of the most critical questions yet: how to choose the best base. In other words, if we have a set of *N*-dimensional vectors, and now we want to reduce them to *K*-dimensional where *K* is less than *N*, then how should we choose the *K* bases to retain the most of the original information?

To fully mathematicalize this problem is complicated, here we use an informal intuitive method to see this problem.

In order to avoid too abstract discussions, we still use a concrete example. Suppose our data consists of five records and represents them as a matrix:

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

Each column is a data record, while each row is an observer. For the convenience of following processing, we first subtract the row means from all the values in each row, resulting in the means of each row is 0. The reason and benefits of doing this will be seen later.

The first row of the above data has a mean of 2 and the second row has a mean of 3, so after the transformation:

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

We can see how the five pieces of data look like in the Cartesian coordinate system:

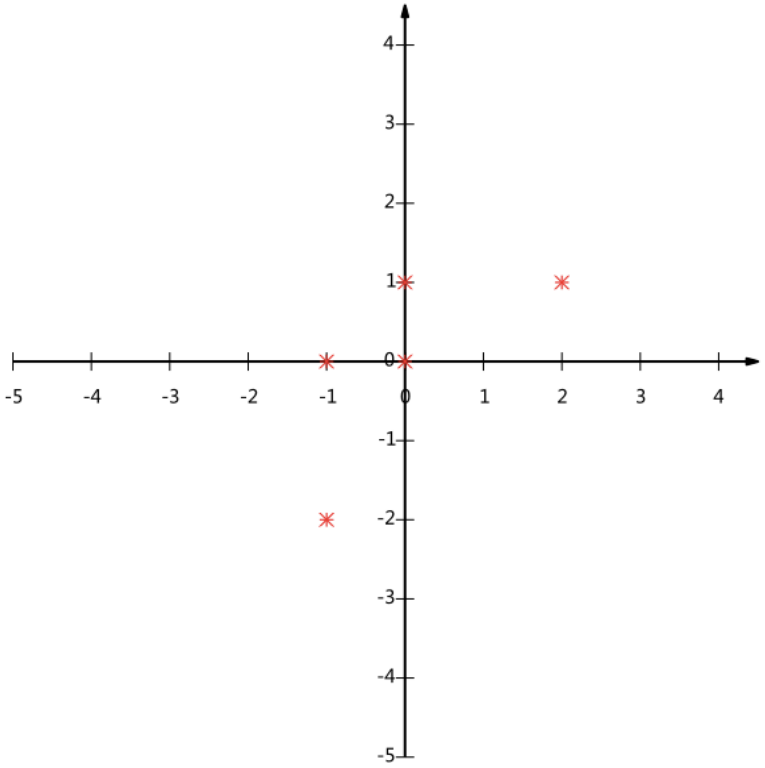

Figure 5. Five Pieces of Data in the Cartesian Coordinate System

If we have to use one dimension to represent these data and want to keep the original information as much as possible, how do you choose?

Through the discussion of the base transform in the previous section, we know that this problem is actually to select a direction in the two-dimensional plane to project all the data onto the line in this direction, and use the projection value to represent the original record. This is an actual two-to-one dimension reduction problem.

So how do we choose this direction or base to retain as much of the original information as possible? An intuitive view is that we expected the projected values after projection to be as scattered as possible.

As shown in the above figure, it can be seen that if we project to the x-axis, the two leftmost points and the two middle points will overlap, so that only two different values left where there used to be four different two-dimensional points. This is considered as a serious loss of information. Similarly, the two uppermost points and the two on the x-axis will overlap if they are projected to the y-axis. So it seems that the x-axis and y-axis are not the best projection choices. We are able to visually observe that the five

points can be distinguished after projection if they are projected obliquely through the first and third quadrants.

In the following sections, we use mathematical methods to express this problem.

4.1. Variance

As mentioned above, we want the projection values to be as scattered as possible, and the degree of dispersion can be expressed by variance. Here, the variance of a row can be thought of as the mean of the sum of the squares of the differences between each element and the row mean, that is:

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m}(a_i - \mu)^2$$

Since we have averaged the mean of each row to 0 above, the variance can be directly represented by the sum of the squares of each element divided by the number of elements:

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m}a_i^2$$

The above problem is then formalized as finding a one-dimensional base to maximize the variance after all data is transformed into the coordinate representation in this base.

4.2. Covariance

For the two-to-one dimension reduction problem above, it is to find the direction to maximize the variance. However, for higher dimensions, there are still problems to be solved. Consider the three-to-two dimension reduction problem. Similarly, first we want to find a direction that maximizes the variance after projection, thus completing the first direction selection, and then we choose the second projection direction.

If we still only choose the direction with the largest variance, it is obvious that this direction and the first direction should be almost coincident. Obviously, such a dimension is useless, so there should be other constraints. Intuitively, in order to let the two rows represent more original information as much as possible, we do not want linear correlation between them, because correlation means that the two rows are not completely independent and there must be repeated information.

Mathematically, the covariance of two rows can be used to indicate their correlation. Since each row has been modified to have a mean of 0, then:

$$Cov(a) = \frac{1}{m}\sum_{i=1}^{m}a_i b_i$$

It can be seen that in the case where the row mean is 0, the covariance of two rows is simply expressed as the inner product of them divided by the number of elements m.

When the covariance is 0, it means that the two rows are completely independent. In order to let the covariance be 0, we select the second base only in the direction orthogonal to the first base. Therefore the two directions eventually must be orthogonal.

So far, we have obtained the optimization of the dimension reduction problem: reducing a set of *N*-dimensional vectors to *K*-dimensional (*K* is greater than 0 and less than *N*), the goal is to select *K* unit (modulus is 1) orthogonal basis, so that the covariance between any two rows is 0 after the original data is transformed into this set of bases, and the variances of rows are as large as possible (under the orthogonal constraint, the largest *K* variances are chosen).

4.3. Covariance Matrix

We have exported the optimization goal above, but it does not seem to be directly considered as an operational guide or algorithm, because it only says what it is rather than saying how to do it. So we have to continue to study the calculation scheme in mathematics.

We can see that the ultimate goal is closely related to the variance within the rows and the covariance between the rows. Therefore, we hope that we might find a way to express the two together. We can observe that both can be expressed as the form of inner product, and the inner product is closely related to the matrix multiplication. Then we have the following inspiration.

Suppose we only have two rows *a* and *b*, then we group them by rows into a matrix *X*:

$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix}$$

Then we multiply *X* by the transpose of *X* and multiply the factor by $1/m$:

$$\frac{1}{m}XX^T = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{pmatrix}$$

The two elements on the diagonal of this matrix are the variance of the two rows, while the other elements are the covariance of *a* and *b*. Both are unified into a single matrix.

According to the principle of matrix multiplication, this conclusion can easily be generalized to the general case.

Suppose we have m n-dimensional data, and arrange them in a matrix X of n by m. Set $C = \frac{1}{m}XX^T$, then C is a symmetric matrix whose diagonals are the variances of the respective rows. The element of *i*th row *j*th column are the same as that of *j*th row *i*th column, representing the covariance of the two rows *i* and *j*.

4.4. Covariance Matrix Diagonalization

Based on the above derivation, we find that to achieve optimization, it is equivalent to diagonalizing the covariance matrix. That is, the elements other than the diagonal are 0 and the elements on the diagonal are sorted from maximum to minimum. In this way, we have achieved optimization. This might not be enough clear. Let us focus further at the relationship between the original matrix and the matrix covariance matrix after the base transformation.

Suppose the covariance matrix of the original data matrix X is C, and P is a matrix of bases arranged by rows. Let $Y = PX$, then Y is the data after X is transformed to P. Suppose the covariance matrix of Y is D. Let us derive the relationship between D and C.

$$\begin{aligned} D &= \frac{1}{m} YY^T \\ &= \frac{1}{m} (PX)(PX)^T \\ &= \frac{1}{m} PXX^T P^T \\ &= P\left(\frac{1}{m} XX^T\right) P^T \\ &= PCP^T \end{aligned}$$

The P we are looking for is the matrix that can diagonalize the original covariance matrix. In other words, the optimization becomes to find a matrix P, which satisfies that PCP^T is a diagonal matrix and the diagonal elements are arranged in order from maximum to minimum. Then the first K rows of P are the bases we want. Multiplying X by a matrix consisting of the first K rows of P, it is able to reduce X from N-dimension to K-dimension, which also satisfies the above optimization conditions.

At this point, there is only one step away from the invention of PCA. Now all the focus is on the diagonalization of the covariance matrix.

It is known from the above that the covariance matrix C is symmetric. In linear algebra, the real symmetric matrix has a series of great properties:

1) The eigenvectors corresponding to different eigenvalues of the real symmetric matrix must be orthogonal.

2) If the multiplicity of the eigenvector λ is r, then there must be r linearly independent eigenvectors corresponding to λ so that the r eigenvectors can be unitized and orthogonalized.

It can be seen from the above two that a real symmetric matrix of n by n can find n unit orthogonal eigenvectors. Suppose the n eigenvectors are $e_1, e_2, \cdots, e_n$, and we form a matrix by column:

$$E = (e_1 \quad e_2 \quad \cdots \quad e_n)$$

Then the covariance matrix C has the following conclusion:

$$E^T CE = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

Λ is a diagonal matrix and the diagonal elements are eigenvalues corresponding to each eigenvector.

Here, we find that we have found the matrix P we need:

$$P = E^T$$

P is a matrix arranged by rows from the unitized eigenvectors of the covariance matrix, where each row is an eigenvector of C. If P is arranged maximum to minimum according to the eigenvalues in Λ and eigenvectors are arranged from top to bottom, then by multiplying the matrix consisting of the first K rows of P with the original data matrix X, we will get the dimension reduction matrix Y.

So far, we have completed the discussion of the entire mathematical principles of PCA.

5. PCA Algorithm

Suppose there are m pieces of n-dimensional data.

1) Make the original data into an n-by-m matrix X;

2) Zero-average each row of X, that is to subtract the row mean from each row;

3) Find the covariance matrix $C = \frac{1}{m}XX^T$;

4) Find the eigenvalues of the covariance matrix and the corresponding eigenvectors;

5) Arrange the eigenvectors into a matrix from top to bottom according to the magnitude of the corresponding eigenvalues, and take the first k rows to form a matrix P;

6) $Y = PX$ is the data after dimension reduction from n-dimension to k-dimension.

6. Further Discussion

Based on the above explanation of the mathematical principles of PCA, we can understand some of the capabilities and limitations of it. PCA essentially takes the direction with the largest variance as the main feature, and dissociates the data in each orthogonal direction. That is to make them have no correlation in different orthogonal directions.

Therefore, PCA also has some limitations. For example, it can release the linear correlation, but there is no way for high-order correlation. In addition, PCA assumes that the main features of the data are distributed in the orthogonal direction. If there are several directions with large variances in the non-orthogonal direction, the effect of PCA is greatly reduced.

Finally, it should be noted that PCA is a parameterless technology. That is to say, everyone will obtain the same results without considering cleaning if processing the same data. PCA is easy to implement generally because of no intervention of subjective parameters, but it cannot personalized optimization.

References

[1] https://en.wikipedia.org/wiki/Principal_component_analysis

[2] https://blog.csdn.net/fngy123/article/details/45153163

[3] https://blog.csdn.net/u012421852/article/details/80458340

[4] https://blog.csdn.net/HLBoy_happy/article/details/77146012

[5] https://zhuanlan.zhihu.com/p/36546123

[6] http://www.voidcn.com/article/p-ocgoxvok-ck.html

[7] http://jermmy.xyz/2017/03/25/2017-3-25-understand-PCA/