**Principal component analysis (PCA)**

Machine Learning algorithms give their best performance when the training dataset is large and concise. Though huge amount of data is the backbone of any predictive model, using such a large data set has its own pitfalls – the biggest one being the curse of dimensionality.

In a large dimensional dataset, there may exist many redundant features or multiple inconsistencies in the features. This not only increases the computation time but also makes data processing. and exploratory analysis more convoluted. To overcome these issues, we perform dimensionality reduction techniques to filter only a limited set of significant features needed for training the model.

Principal components analysis is a dimensionality reduction technique that identifies correlations and patterns in a data set to transform it to a significantly lower dimension data set without losing any important information. It is usually performed to solve complex data-driven problems having very high number of dimensions. Follow the steps given below to perform dimensionality reduction using PCA:

Step 1: Standardize the data.

Step 2: Calculate the covariance matrix.

Step 3: Compute the eigenvectors and eigenvalues.

Step 4: Identify the Principal Components.

Step 5: Reduce the dimensions of the dataset.

Step 1: To standardize data, we need to scale the data in such a way that values of all the variables lie within a similar range. For example, if there are two variables – one having values ranging between 100 and 1000, and the other having values from just 10 to 100. In such a scenario, the output produced using these predictor variables will be highly biased. Variable with a larger range will impact the outcome in a bigger way. Therefore, data must be standardized into a comparable range. This is done by subtracting each value in the data from the mean and dividing it by the overall deviation in the dataset as given below.

A close up of words

Description automatically generated

Step 2: To identify the correlation and dependencies among the features in the data set. This

step is essential because strongly dependent variables contain biased and redundant information thereby affecting the overall performance of the model. Mathematically, a covariance matrix is a p × p matrix, where p is the number of dimensions in the data set.

Each entry in the matrix represents the covariance of the corresponding variables. We

can visualize the covariance matrix for variables a and b as,

A close up of text

Description automatically generated

Note that,

* Cov(a, a) is the covariance of a variable with itself.
* Cov(a, b) is the covariance of the variable ‘a’ with respect to the variable ‘b’.
* Cov(a, b) = Cov(b, a)
* If Cov(a,b) < 0, the respective variables are indirectly proportional to each other.
* If Cov(a,b) > 0, the respective variables are directly proportional to each other.

Step 3: Find the covariance matrix to determine the principal components of the dataset.

Remember that Principal Components are the new set of variables that are obtained from

the initial set of variables. These variables are highly significant and independent of each other. Principal components compress and possess most of the useful information that was scattered among the initial variables. For example, if the data set has 5 dimensions, then 5 principal components are computed in such a way that the first principal component stores the maximum possible information, the second stores the remaining maximum information, so on and so forth. Eigenvectors and eigenvalues are always computed as a pair. This means that for every eigenvector there is an eigenvalue. Number of eigenvectors depends on the number of dimensions in the data. For example, in a 2-dimensional dataset, two eigenvectors are computed. By looking at the Covariance matrix, we need to understand where in the data their maximum variance is. More variance gives more information about the data. Thus, eigenvectors and eigenvalues will then be used to compute the Principal Components of the dataset.

Step 4: After computing the Eigenvectors and eigenvalues, we need to arrange these in the

descending order because the eigenvector with the highest eigenvalue is the most significant and thus forms the first principal component. The principal components of lesser significances can be removed in order to reduce the dimensions of the data. Finally, a feature matrix containing all the significant data variables that possess maximum information about the data is created.

Step 5: The last step in the PCA technique is to reduce the dimensions of the data set. This is

done by re-arranging the original data with the final principal components having maximum and the most significant information of the data set. To replace the original data axis with the newly formed Principal Components, we need to multiply the transpose of the original data set by the transpose of the obtained feature vector.Technically, PCA can be explained as given below.

PCA reduces dimensions of data by projecting it onto lines drawn through data. Line that goes through the data in the direction of the greatest variance is drawn first. This is calculated by computing the eigenvectors of the covariance matrix.

To represent linear transformation through a matrix, we multiply some data points by the matrix to move around the graph.

The eigenvectors of a linear transformation are the lines where if a data point started on that line, they end on that line, and the eigenvalue says how far they have moved. Thus, eigenvectors talk about the direction of a linear transformation.

A covariance matrix can be used as a linear transformation, and the square root of the covariance matrix of some data will perform a linear transformation that moves Gaussian data points into the shape of our data having the same standard deviations and correlations as the data.

So, the eigenvectors of the square root of the covariance matrix tells us in what direction data

has been smeared away from perfectly normally distributed and the eigenvalues tell us how far. The direction of the smearing is the same as the principal components, and how far our data was smeared tells us which direction has the greatest variance.