Chapter 3: Machine Learning Algorithms

Principle Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

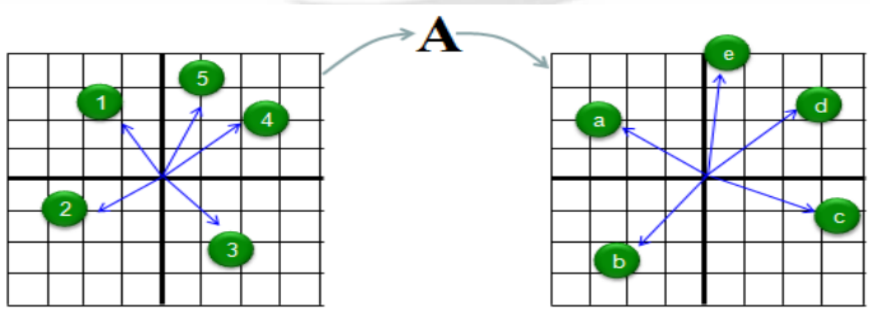
Eigenvector and Eigenvalue

If you think of a Matrix as a geometric transformer, the Matrix usually perform two types of transformational action. One is 'scaling (extend/shrink)' and the other one is 'rotation'. (There are some additional types of transformation like 'shear', 'reflection', but these would be described by special combination of scaling and rotation).

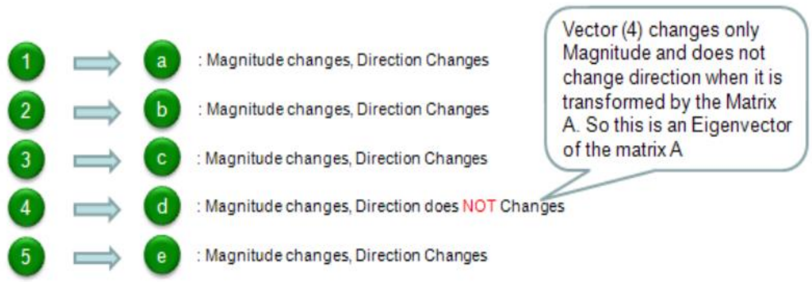
Eigenvector and Eigenvalue can give you the information on the scaling and rotational characteristics of a Matrix. Eigenvector would give you the rotational characteristics and Eigenvalue would give you the scaling characteristics of the Matrix.

When a vector is transformed by a Matrix, usually the matrix changes both direction and amplitude of the vector, but if the matrix applies to a specific vector, the matrix changes only the amplitude (magnitude) of the vector, not the direction of the vector. This specific vector that changes its amplitude only (not direction) by a matrix is called Eigenvector of the matrix.

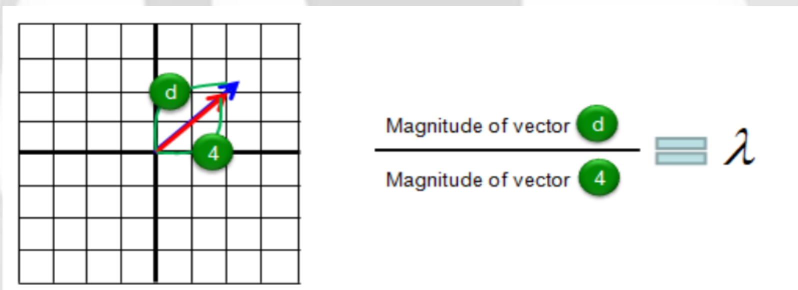
Let me try explaining the concept of eigenvector in more intuitive way. Let's assume we have a matrix called 'A'. We have 5 different vectors shown in the left side. These 5 vectors are transformed to another 5 different vectors by the matrix A as shown on the right side. Vector (1) is transformed to vector (a), Vector (2) is transformed to vector (b) and so on.



Compare the original vector and the transformed vector and check which one has changes both its direction and magnitude and which one changes its magnitude ONLY. The result in this example is as follows. According to this result, vector (4) is the eigenvector of Matrix 'A'.



Now we know eigenvector changes only its magnitude when applied by the corresponding matrix. Then the question is "How much in magnitude it changes?". Did it get larger? or smaller? exactly how much? The indicator showing the magnitude change is called Eigenvalue. For example, if the eigenvalue is 1.2, it means that the magnitude of the vector gets larger than the original magnitude by 20% and if the eigenvalue is 0.8, it means the vector got smaller than the original vector by 20 %. The graphical presentation of eigenvalue is as follows.



We can use principal component analysis (PCA) for the following purposes:

- To reduce the number of dimensions in the dataset.
- To find patterns in the high-dimensional dataset
- To visualize the data of high dimensionality
- To ignore noise
- To improve classification
- To gets a compact description
- To captures as much of the original variance in the data as possible

In summary, we can define **principal component analysis (PCA)** as the transformation of any high number of variables into a smaller number of uncorrelated variables called principal components (PCs), developed to capture as much of the data's variance as possible.