Dimensionality Reduction

BY MG ANALYTICS

Feature Engineering

Give me six hours to chop down a tree and I will spend the first four sharpening the axe.
—- **Abraham Lincoln**

The above quote has a great influence in the machine learning too. When it comes to modeling different machine learning models most of the time need to spend on data **preprocessing** and **feature engineering** stages.

Curse of Dimensionality

- Incomplete data with a few features - led to the development of the **common myth**.
- Having more features and more data will always improve the accuracy of solving the machine learning problem. In reality, this is a curse more than a gain.
- Lot of features with few data points. Fitting a model in this scenario often leads to a low accuracy model even with many features - called as the "Curse of dimensionality".

Curse of dimensionality –

- Phenomenon when an increase in the number of features results in decrease the model accuracy.
- Increase in the number of features increase the model complexity (more precise the model complexity increase exponentially)

Ways to stay away

Two ways to stay away from this curse of dimensionality.

- Add more data to the problem.
- Reduce the number of features in the data.
- Adding data may not be possible in many scenarios
- Reducing the number of features is more preferable. Such a technique is known as **“Dimensionality reduction”**

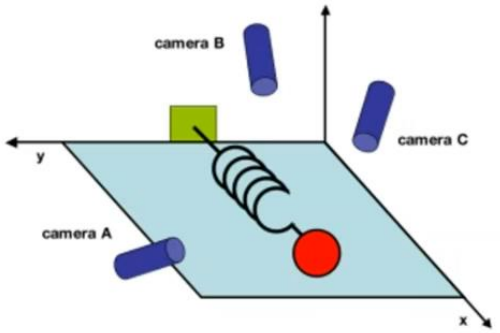
Feature Engineering

- Identify the influence features over all the available features. So the identified features used to train the model.
- Identifying the influence features doesn't mean picking the features in an analytical way. Some times converting latent features into a meaningful feature. This is known as **dimensionality reduction**.
- One famous approach know as **Principal component analysis** (PCA) to perform the dimensionality reduction

•Feature Elimination
•Feature Extraction

PCA explained in simple terms

- Consider simple problem such as recording the motion of a pendulum, which moves in only one direction.
- If one is unaware of the exact direction. The number of cameras required to record its movement will at least be three, given that we are able to place the cameras perpendicular to each other.
- If we do not have the knowledge of keeping the cameras perpendicular to each other, then more cameras will be required. The problem then keeps on increasing and more and more cameras are required as available information keeps on decreasing.



Principal Components

- The PCA technique transforms our features (or cameras) into a new dimensional space and represents it as a set of new orthogonal variables so that our problem is observed with a reduced set of features.

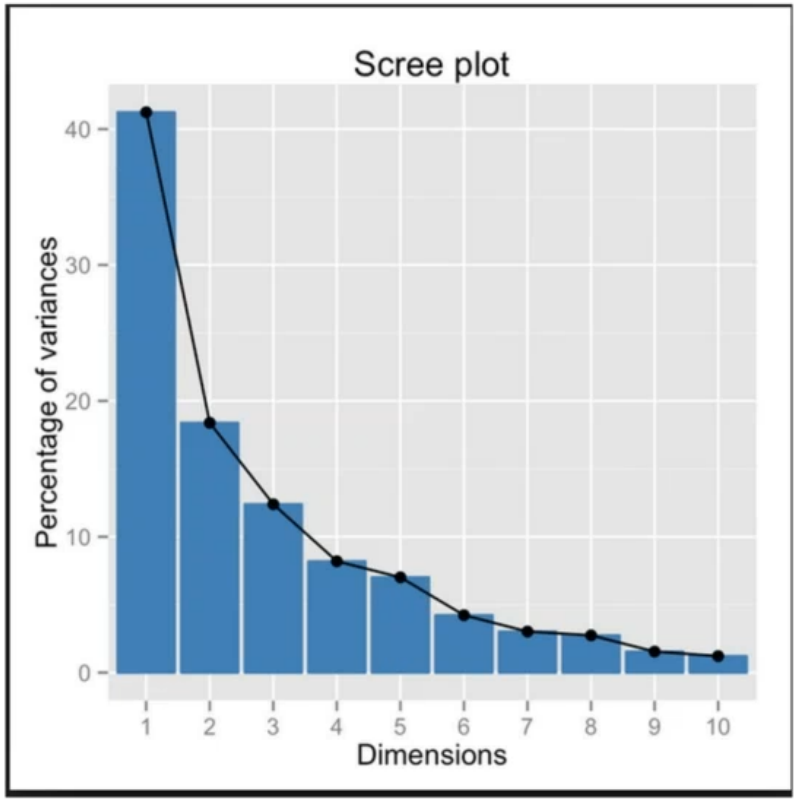
These orthogonal features are known as "**Principal Components".**

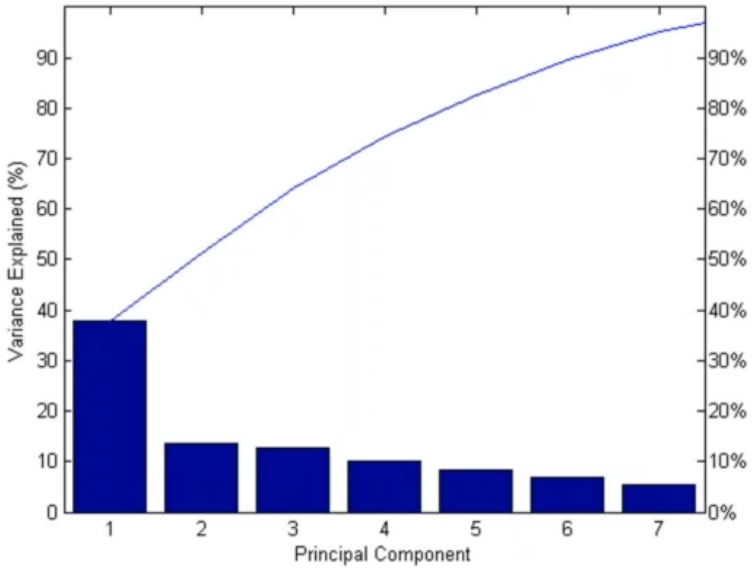
- In practice, our data is like the motion of the pendulum. If we had complete knowledge of the system, we will require a small number of features.
- We have to observe the system using a set of features which will convey maximum information if they are orthogonal. This is done using principal component analysis.
- The new set of features which are produced after PCA transformation are linearly uncorrelated as they are orthogonal. Moreover, the features are arranged in decreasing order of their importance.

Principal Components

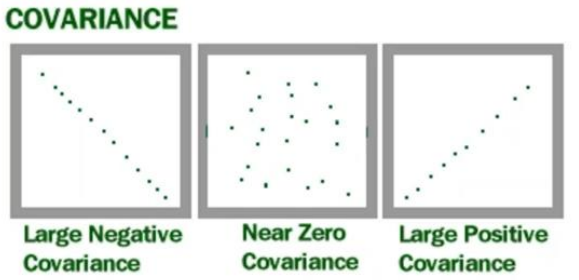
- This means that the first principal component alone will explain a very large component of the data.
- The second principal component will explain less than the first major component but more than all other components.
- The last principal component will explain only a small change in the data.
- Run PCA and take the top principal components such that they together explain most of the data.
- In most analytical problems, explaining **90-99%** of the is considered very high.

Principal Components – Optimum number





Terms



Covariance is a measure of the joint variability of two random variables.

If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values - covariance is positive

In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other - Covariance is negative.

The sign of the covariance shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables.

The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

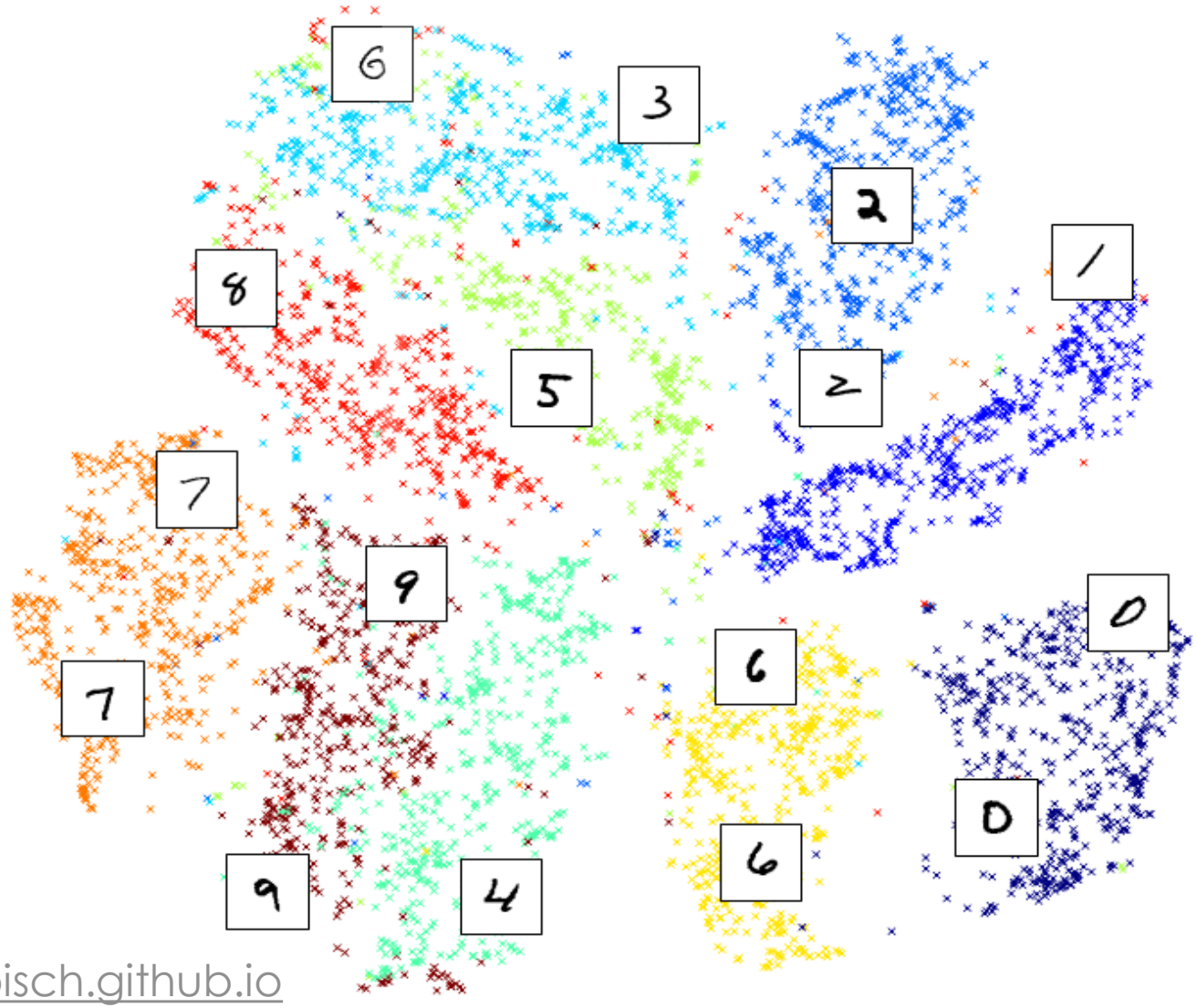
Principle Components

- **How do we transform a given set of features into a new feature set such that they are orthogonal?**
- The answer is the eigenvectors of the matrix.
- We know that eigenvectors are orthogonal to each other so transforming our features in the direction of the eigenvectors will also make them orthogonal.
- **But wait!** Before transforming a matrix, it is always recommended to normalize.
- If the matrix is not normalized, our transformation will always be in favor of the feature with the largest scale of values. This is why **PCA is sensitive** to the relative scaling of the original variables.

Why?

- to treat multicollenearity
- Reduce number of features
- to visualize Data

NIST dataset – Two-dimensional embedding of 70,000 handwritten digits witl



▶http://alexanderfabisch.github.io/t-sne-in-scikit-learn.html