Principal Component Analysis

2019年5月9日 11:51

About PCA source from:

PCA数学原理: http://blog.codinglabs.org/articles/pca-tutorial.html

A One-Stop Shop for Principal Component Analysis: https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Python Documentation for PCA within the sklearn library: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Principal Component Analysis in 3 Simple Steps: http://sebastianraschka.com/Articles/2015_pca_in_3_steps.html

A good Visualize: http://setosa.io/ev/principal-component-analysis/

A Tutorial on Principal Component Analusis:

A Tutorial on Princi...

PCA(Principal Component Analysis) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Reducing the dimension of the feature space is called "dimensionality reduction". There are many ways to achieve dimensionality reduction, but most of these techniques fall into one of two classes:

1. Feature Elimination
2. Feature Extraction

Feature elimination is what it sounds like: we reduce the feature space by eliminating features. Advantages of feature elimination methods include simplicity and maintaining interpretability of your variables. As a disadvantages, though, you gain no information from those variables you've dropped. By eliminating features, we've also entirely eliminated any benefits those dropped variables would bring.

Feature extraction is create "new" independent variables, which are combination of each of the "old" independent variables. These new independent variables in a specific way and order these new variables by how well they predict our dependent variable.

Principal component analysis is a technique for feature extraction. It combines our input variables in a specific way, then drop the "least important" variables while still retaining the most valuable parts of all of the variables. As an added benefit, each of the "new" variables after PCA are all independent of one another.

The sheer size of data in the modern age is not only a challenge for computer hardware but also a main bottleneck for the performance of many machine learning algorithms. The main goal of a PCA analysis is to identify patterns in data; PCA aims to detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense. In a nutshell, this is what PCA is all about: Finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.

PCA VS **LDA**

Both Linear Discriminant Analysis(LDA) and PCA are linear transformation methods. PCA yields the directions(principal components) that maximize the variance of the data, whereas LDA also aims to find the directions that maximize the separation(or discrimination) between different classes, which can be useful in pattern classification problem(PCA "ignores" class labels).

In other words, PCA projects the entire dataset onto a different feature (sub)space, and LDA tries to determine a suitable feature (sub)space in order to distinguish between patterns that belong to different classes.

PCA and Dimensionality Reduction

Often, the desired goal is to reduce the dimensions of a d-dimensional dataset by projecting it onto a (k)-dimensional subspace(where k < d) in order to increase the computational efficiency while retaining most of the information. An important question is "what is the size of k that represents the data 'well'?"

Later, we will compute eigenvectors (the principal components) of a dataset and collect them in a projection matrix. Each of those eigenvectors is associated with an eigenvalue which can be interpreted as the “length” or “magnitude” of the corresponding eigenvector. If some eigenvalues have a significantly larger magnitude than others, then the reduction of the dataset via PCA onto a smaller dimensional subspace by dropping the “less informative” eigenpairs is reasonable.

A Summary of the PCA Approach

1. Standardize the data
2. Obtain the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix, or perform Singular Vector Decomposition(SVD)
3. Sort eigenvalues in descending order and choose the k eigenvectors that correspond to the k largest eigenvalues where k is the number of dimensions of the new feature subspace(k < d)
4. Construct the projection matrix W from the selected k eigenvectors
5. Transform the original dataset X vis W to obtain a k-dimensional feature subspace Y