Principal Component Analysis

Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

24 Jul 2022

1. Introduction

Principal component analysis (PCA) is a linear dimensionality reduction algorithm. It finds the principal components of a given dataset. It is one of the most popular dimensionality reduction algorithms and is often used for exploratory data analysis.

Principal components of a dataset are a sequence of mutually perpendicular lines that best fit the data. In other words, the variation in the data is best explained when the data points are projected onto the principal components. Most of the variation is often explained by the first few principal components (when the data is not noisy) and these are the only ones used, thus resulting in dimensionality reduction.

2. Mathematical Formulation

Suppose we are given m samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ where $\mathbf{x}_i \in \mathbb{R}^n \ \forall i \in [m]$
The centroid of the data points is given by

$$\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_m \tag{2.1}$$

The n principal components are a sequence of n mutually perpendicular lines passing through $\bar{\mathbf{x}}$

$$\mathbf{x} = \bar{\mathbf{x}} + \lambda \mathbf{w}_k \qquad \lambda \in \mathbb{R}, \mathbf{w}_k \in \mathbb{R}^n, \|\mathbf{w}_k\| = 1 \ \forall k \in [n] \tag{2.2}$$

$$\text{or, } \mathbf{w}_k^\top(\mathbf{x} - \bar{\mathbf{x}}) = \lambda \tag{2.3}$$

The first principal component (PC1) is the line passing through $\bar{\mathbf{x}}$ that best fits the data. For finding such a line, we minimize the sum of squares of distances of all the points from the line which is equivalent to maximizing the sum of squares of the projections of the points on the line.

$$\hat{\mathbf{w}}_1 = \underset{\|\mathbf{w}_1\|=1}{\arg\max} \sum_{i=1}^{m} \left(\mathbf{w}_1^\top(\mathbf{x}_i - \bar{\mathbf{x}})\right)^2 \tag{2.4}$$

The corresponding unit vector (whose components are known as loading values) obtained is known as the singular vector or the eigenvector for PC1 and the sum of squares is known as the eigenvalue for PC1. The square root of the eigenvalue is known as the singular value.

The other principal components can be found out using a similar procedure. The k^{th} principal component maximizes the variation about it under the constraint that is perpendicular to all the previous $k-1$ principal components.

$$\hat{\mathbf{w}}_k = \underset{\|\mathbf{w}_k\|=1}{\arg\max} \sum_{i=1}^{m} \left(\mathbf{w}_k^\top(\mathbf{x}_i - \bar{\mathbf{x}})\right)^2 \ \text{ s.t. } \mathbf{w}_k^\top \mathbf{w}_j = 0 \ \forall j \in [k-1] \tag{2.5}$$

3. Choosing the Number of Principal Components

The eigenvalues are a measure of how much of the variation in data is explained by a particular principal component. They decrease monotonically with increase in k. The proportion of variation that the k^{th} principal component accounts for is given by

$$\frac{\sum_{i=1}^{m}\left(\mathbf{w}_k^\top(\mathbf{x}_i-\bar{\mathbf{x}})\right)^2}{\sum_{j=1}^{n}\sum_{i=1}^{m}\left(\mathbf{w}_j^\top(\mathbf{x}_i-\bar{\mathbf{x}})\right)^2} \tag{3.1}$$

The line plot of these proportions is known as a scree plot, which is often used to determine the number of principal components to use to reduce the dimensionality. Either, we choose those principal components whose cumulative variance crosses a certain threshold (say 90%) or we choose the elbow of the scree plot as our cut-off. After choosing the appropriate number of principal components, we project the original data points onto these principal components to get our transformed dataset.

4. Questions

i) What is the purpose of dimensionality reduction?
ii) Explain the curse of dimensionality.
iii) Why is PCA a linear dimensionality reduction algorithm?
iv) What are the results of PCA when used on noisy data?
v) Mention a drawback of PCA.

5. Answers

i) The main purpose of dimensionality reduction is to prevent overfitting by making the model less complex. It also removes noise and correlated features. Moreover, the computational requirements are reduced.
ii) The curse of dimensionality refers to various undesirable phenomena that happen when dealing with high-dimensional spaces. As the number of dimensions increases, we need more and more data to generalize accurately. Also, the concept of Euclidean distance becomes meaningless when the number of dimensions is very high because there is little difference between the distances between different points. Dimensionality reduction prevents the curse of dimensionality.
iii) The principal components are nothing but equations of lines, which are linear combinations of the features of the dataset. Therefore, the transformation brought about by PCA is also linear.
iv) When the data is noisy, most of the variation in the data will not be explained by the first few principal components and even the latter components will have significant contributions. However, PCA can still be used to identify clusters in noisy data.
v) PCA can only capture linear correlations between the features. When there are hidden non-linear patterns in the data, PCA will not be able to identify them and will instead give misleading results.