Class–Dependent Principal Component Analysis

Oleg Okun
SMARTTECCO, Stockholm, Sweden

INTRODUCTION

Principal Component Analysis (PCA) (Jolliffe, 2002) is one of the popular methods for dimensionality reduction that is often used in Predictive Analytics tasks. However, it is an unsupervised technique as it ignores class membership information when projecting data into a lower dimensional space from the original space. Therefore when PCA precedes data classification, one cannot always be certain if classification in the reduced space is more accurate than that in the original space as dimensionality reduction is unrelated to classification.

To address this problem, various authors proposed supervised PCA that utilizes the information about labels of instances (Bair, Hastie, Paul, & Tibshirani, 2006; Chen, Wang, Smith, & Zhang, 2008; Das & Nenadic, 2008; Barshan, Ghodsi, Azimifar, & Jahromi, 2011; Wu, Bowers, Huynh, & Souvenir, 2013; Cai, et al., 2013).

As an example of this type of algorithms, the work of Das and Nenadic (2008) is presented in detail in this chapter. Das & Nenadic (2008) proposed an algorithm, where principal subspace is found for each class of data, independently of other classes. Test data are then projected into each principal subspace and the Bayes rule judges which class in which subspace is associated with the maximum posterior probability. Thus, dimensionality reduction is combined with classification.

Das and Nenadic argued that partitioning the original space onto multiple linear subspaces leads to more accurate classification results than the conventional wholistic PCA where only one linear subspace is used for all classes of data. Their motivation was based on the assumption that the projection onto a single linear subspace will be inadequate if different classes are highly overlapped. In this case, class-dependent PCA would have better chances to succeed where the class-ignorant PCA failed.

BACKGROUND

Principal Component Analysis

Principal component analysis is a procedure for analyzing multivariate data which transforms the original variables $x_1,x_2,\ldots,x_N$ into new ones $y_1,y_2,\ldots,y_N$ that are uncorrelated and account for decreasing proportions of the variance in the data (Everitt, 2006). PCA can also be defined as the orthogonal projection of the data onto a lower dimensional space (principal subspace), such that the variance of the projected data is maximized (Hotelling, 1933).

The new variables, the principal components, are defined as linear functions of the original variables (Everitt, 2006). Thus, PCA is a linear dimensionality reduction technique.

The main assumption behind PCA is that rich information in the data space corresponds to directions of high variance. That is, one needs to search for directions of maximum variance in the reduced space, which implies that one supposes that:

1. The data exhibits higher variance in certain directions while smaller variance or no variance at all in other directions, and
2. The number of directions of high variance is much smaller compared to the original data dimensionality.

So, speaking in mathematical terms, it is necessary to maximize the variance of the projected data, subject to certain constraints, which renders this task as an optimization one. Let us define the variance of the projected data and for simplicity but without loss of generality let us consider the projection onto one-dimensional space as done in Bishop (2006, page 562).

Given the data x_i, $i=1,2,\ldots,N$ as D-dimensional column-vectors[1], this linear projection will take the form $u_1^T x_i$, where u_1 is the D-dimensional projection column-vector, "T" stands for transposition (u_1^T is thus a row-vector), and the subscript "1" means "first dimension," so that the result of $u_1^T x_i$ is a scalar (real number). For convenience (since we are interested in the direction of u_1 but not in its magnitude), let us impose such a constraint on u_1: $u_1^T u_1 = 1$ or equivalently, $\|u_1\|_2 = 1$, where $\|.\|_2$ is the Euclidean norm of a vector. This constraint means that the projection vector has unit length and it prevents $\|u_1\|_2 \to \infty$. It is common to define coordinate systems in terms of directions of unit-length vectors sometimes called basis vectors. Therefore one can think of the desired principle subspace as one of such coordinate systems.

The variance of the projected data is then given by

$$\frac{1}{N-1}\sum_{i=1}^{N}\left(u_1^T x_i - u_1^T \bar{x}\right)^2$$
$$= u_1^T \left\{ \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \right\} u_1 = u_1^T S u_1,$$

where

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i,$$

$$S = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2.$$

In these formulas, $\bar{x}$ and S are the sample mean ($D\times 1$ vector) and sample covariance $D\times D$ matrix. Notice that since each vector x_i is projected onto the principal subspace, the mean vector also needs to be projected. Moving u_1 out of the sum was possible because this vector is the same for all vectors x_i.

As a result, our optimization problem is formulated as

$$\text{Maximize } u_1^T S u_1, \quad \text{Subject to } \|u_1\|_2 = 1.$$

The standard approach to solve the constrained optimization problems is to compose the Lagrangian L by using the Lagrangian multiplier λ_1, to take the derivative $\dfrac{dL}{du_1}$ and to set it to zero:

$$L = u_1^T S u_1 + \lambda_1 \left(1 - u_1^T u_1\right),$$

$$\frac{dL}{du_1} = 2S u_1 - 2\lambda u_1 = 0.$$

Hence, the stationary point corresponds to $S u_1 = \lambda_1 u_1$, which implies that u_1 must be an eigenvector of the sample covariance matrix S, since the last equation describes the typical eigenproblem[2]. When left-multiplying the last equation by u_1 and keeping in mind the constraint $u_1^T u_1 = 1$, one obtains that the variance is $u_1^T S u_1 = \lambda_1$ so that it is maximized when u_1 is equal to the eigenvector of the sample covariance matrix, having the largest eigenvalue λ_1. This eigenvector is called the first principal component. As can be seen, the first largest eigenvalue is equal to the sample variance of the first principle component.

Other principal components can be incrementally added by searching for a new direction that maximizes the projected variance among all possible directions orthogonal to those already found. Therefore if the principal subspace is m-dimensional, we need to find the m eigenvectors of the data covariance matrix corresponding to the m

Related Content

Employer Brand Equity and Its Impact on the Application Intent of the Prospective Employees
Dipali Shrikant Dalvi (2021). *International Journal of Business Intelligence Research (pp. 1-25).*
www.irma-international.org/article/employer-brand-equity-and-its-impact-on-the-application-intent-of-the-prospective-employees/269444

A Case Study on Data Quality, Privacy, and Entity Resolution
William Decker, Fan Liu, John Talburt, Pei Wangand Ningning Wu (2014). *Information Quality and Governance for Business Intelligence (pp. 66-87).*
www.irma-international.org/chapter/a-case-study-on-data-quality-privacy-and-entity-resolution/96145

Optimizing Group Waiting Time in Service System with Learning Effect
Yuval Cohenand Shai Rozenes (2017). *International Journal of Business Analytics (pp. 18-35).*
www.irma-international.org/article/optimizing-group-waiting-time-in-service-system-with-learning-effect/169218

Digitalization in Accounting Through Changing Technology and Accounting Engineering as an Adaptation Proposal
Azize Esmerayand Murat Esmeray (2020). *Handbook of Research on Strategic Fit and Design in Business Ecosystems (pp. 354-376).*
www.irma-international.org/chapter/digitalization-in-accounting-through-changing-technology-and-accounting-engineering-as-an-adaptation-proposal/235581

The Importance of Process Thinking in Business Intelligence
Olivera Marjanovic (2010). *International Journal of Business Intelligence Research (pp. 29-46).*
www.irma-international.org/article/importance-process-thinking-business-intelligence/47194