Mutual Principal Components, reduction of dimensionality in statistical classification

Carlos Cuevas-Covarrubias

Anahuac University, Mexico

Abstract

Linear discriminant analysis (LDA) and principal components analysis (PCA) are two fundamental tools of multivariate statistics. Given a p-dimensional random variable $\mathbf{X}$, PCA finds its optimal representation in a lower dimensional space. LDA assumes that the sample space of $\mathbf{X}$ is partitioned into two different categories. Given $\mathbf{x}$, a particular realization of $\mathbf{X}$, LDA lets us infer whether $\mathbf{x}$ comes from one category or the other. We present an original combination of PCA and LDA where the area under the ROC curve appears as the link between both methods; we call this *Mutual Principal Components*. Our objective is to represent $\mathbf{X}$ in terms of a small number of non correlated factors and maximum separability. Assuming that $\mathbf{X}$ is distributed according to a Gaussian mixture, a parametric approach selects those components with maximum contribution to the area under the ROC curve of an optimal linar discriminant function. A distribution free alternative shows that this principle is equivalent to maximize the square cosine between this discriminant function and the vector space generated by the colums of the resulting principal components transformation matrix.

Keywords

Classification, Linear score, ROC curve, PCA, Reduction of dimensionality.

References

[1] Anderson, T.W. and R.R. Bahadur (1962). Classification into two multivariate normal distributions with different covariance matrices. *A*nn. Math. Statist. 33, 420–431.

[2] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed). John Wiley and Sons.

[3] Chang, W.C. (1983). Using Principal Components before separating a mixture of two multivariate normal distributions. *Appl. Statist. 32(3)*, 267–275.

[4] Krzanowski, W.J. and D.J. Hand (2009). *ROC Curves for Continuous Data*. CRC Press.