



THEORETICAL REPORT

IFS-TR-006

USE OF CANONICAL DISCRIMINANT ANALYSIS WITHIN AMBER FOR DIMENSION REDUCTION

ABSTRACT

As we have already discussed in Theoretical Report IFS-TR-005, it is important to reduce the amount of redundant information in an input vector prior to its inclusion into a neural network. In that report, it was suggested that this issue could be tackled by using the Principal Components Analysis (PCA) transform as a method of reducing the input vector dimension. This was done by effectively de-correlating the components of the input vectors. A reduced dimension vector containing uncorrelated inputs was produced by projecting the input vector onto a sub-space spanned by the top N principal components.

However, for pattern recognition and regression problems, there is likely to be other information in the input vectors that, although not correlated with other input components, is nevertheless redundant **with respect to the problem of recognition or regression**. As the problem of predicting future values of a financial time series may be regarded as either a regression (estimates of expected future return) or classification (forecast positive or negative future returns) problem, this is important when considering inputs to neural networks within Amber.

In this report, we shall examine a method of removing redundant information using a technique called Canonical Discriminant Analysis (CDA) also known as Canonical Variates Analysis.

REPORT

BACKGROUND

Consider the problem of predicting future returns of the Dow Jones index based on a thirty dimensional vector of information. Suppose furthermore, that the last five components of the input vector contain information concerning sunspot activity over the last six months.

Now, it is quite possible (indeed likely) that the sunspot information may be uncorrelated with the other (hopefully more informative) components of the input vector. Because of this, a Principal Components Analysis (see Theoretical Report IFS-TR-005) would not eliminate the sunspot activity in its reduced dimension representation of the original vector. However, we would imagine that, in fact, the sunspot activity components are highly unlikely to contain information *relevant to the prediction of future returns of the Dow Jones Index*. Because of this we would prefer to remove them from our reduced dimension representation of the input vector.

In this report, we will look at a possible method of **eliminating irrelevant or redundant** information using the technique of Canonical Discriminant Analysis (CDA).

MATHEMATICAL DERIVATION OF CANONICAL DISCRIMINANT ANALYSIS

The CDA transform is essentially the generalisation of Fisher's linear discriminant function to multiple dimensions (see, for example, Duda & Hart [1]). The canonical discriminants of a set of labelled input vectors may be found in the following manner.

Let us consider a sample of labelled input vectors denoted $X = \{x_i^p, i = 1 \dots T\}$ of size T. Suppose that each (N x 1 dimensional) sample vector x_i^p is either a member of class 1, ($x_i^p \in H_1$), or class 2, ($x_i^p \in H_2$). Let us assume that the two classes refer to samples of either positive or negative future returns of a financial time series. Zero returns can either be excluded from the sample set or assumed to belong to the positive class.

Now let us define two matrix measures of the separability of the two classes in terms of the original un-transformed vectors. We shall then define a separability metric in terms of the ratio of the norms of those two matrices.

Within Class Scatter Matrix S_w

We define the Within Class Scatter Matrix S_w as

$$S_w = \sum_{c=1}^2 \left(\sum_{x_i^p \in H_c} (x_i^p - \mu_c)(x_i^p - \mu_c)^T \right)$$

where μ_c is the mean for class c, defined

$$\mu_c = \frac{1}{N_c} \sum_{x_i^p \in H_c} x_i^p$$

and N_c is the number of input vectors having class c.

Intuitively, the Within Class Scatter Matrix describes the overall spread or variance of the distribution of each class. To optimise the separability of classes we would wish the size (matrix norm) of this measure to be as low as possible.

Between Class Scatter Matrix , S_B

We define the Between Class Scatter Matrix, S_B

$$S_B = \sum_{c=1}^2 N_c (\mu_c - \mu)(\mu_c - \mu)^T$$

where μ is the mean of all input vectors i.e.

$$\mu = \frac{1}{T} \sum_{x_i} x_i$$

The Between Class Scatter Matrix describes the average distance between the means of vectors from the different classes and the overall mean for all classes. Intuitively, for good class separability we will want the size (matrix norm) of this matrix to be large.

Ratio Of The Between And Within Class Scatter Matrices

Consider the ratio of the two matrices defined above,

$$J = \frac{|S_B|}{|S_W|}$$

This value will be large if $|S_W|$ is small and $|S_B|$ is large. Recalling the arguments above, J is a good indication of the (linear) separability of the classes. Therefore we should strive to ensure that the reduced dimensional vectors maintain a high value of J after projecting them into the reduced dimensional sub-space. In fact there are many other choices we might make for J. For examples see, for instance, Fukunaga [2]

Selecting The Optimal Transformation

Recall that our objective is to reduce the dimension of the input vectors whilst at the same time maintaining, as far as possible, the discriminatory information contained in the original vector. Let us represent the transformation of the original vectors x_i onto the reduced M dimensional subspace as the linear transformation

$$y_i = A^T x_i$$

where A is an $(N \times M)$ dimensional matrix whose rows are the axes of the subspace onto which the original vectors, \mathbf{x}_i are being projected. For a given choice of reduced dimension, M , we wish to find the optimal A .

Let P_W denote the Within Class Scatter Matrix of the projected vectors \mathbf{y}_i and P_B denote the Between Class Scatter Matrix of the projected vectors. The Within Class and Between Class Scatter Matrices for the projected vectors (P_W and P_B) can be written in terms of the transformation matrix, A , and the original Within and Between Class Scatter Matrices (S_W and S_B) i.e.

$$P_W = A^T S_W A$$

$$P_B = A^T S_B A$$

Recalling the discussion from the previous section, a good indication of the linear separability of the classes is the metric J . We therefore wish to find the matrix A that will optimise the ratio $\frac{|P_B|}{|P_W|}$ (i.e. give us large inter-class separations between tightly clustered class distributions). It turns out that this optimisation can be performed by solving the generalised eigenvalue problem.

$$S_B \mathbf{x}_i = \lambda_i S_W \mathbf{x}_i$$

The eigenvectors \mathbf{x}_i corresponding to the top M eigenvalues of the solution to the above are then the rows of the optimal transformation matrix A . Care must be taken in calculating the Matrix $(S_W^{-1} S_B)$, as it is generally non-symmetric. This can sometimes lead to numerical instabilities in the calculation of the eigenvector/eigenvalues.

DISCUSSION

In this report we have shown a method of finding an optimal projection of a set of input vectors onto a reduced dimensional sub-space in terms of maintaining the maximum possible inter-class separation. This reduced dimensional projected vector should then provide good discriminatory information if it is included in a neural network. However, there are a number of practical issues to bear in mind when using the Canonical Discriminant Analysis module within Amber.

1.) The measure of separation used is appropriate for **linear** classifiers. As we are using non-linear predictors (neural nets) the reduced dimension form of the vector may not be optimal.
2.) In the example given above we have separated future returns into two classes, positive and negative future returns. In reality there are a spread of returns ranging from strongly positive returns to strongly negative returns. In simplifying this to two classes we may lose the ability, for instance, to distinguish between very large

positive returns and very small positive returns. This can be rectified, to an extent, by subdividing the returns into more classes. For instance we may consider the five classes of returns shown in the table below.

3.) Because of the rank of the matrix $(S_W^{-1}S_B)$ it is only possible to project the original vectors onto a reduced dimension of at most $C - 1$, where C is the number of distinct classes we have. This problem can be alleviated by dividing the future returns into more classes, as described in (2) above.
4.) CDA was really designed for pattern recognition, not regression, problems. If we wish to discriminate between different magnitudes of positive and negative returns then the specially designed Non-Parametric Discriminant Analysis transform may be better suited to the task. Details of this (proprietary) IFS technique may be found in Theoretical Report IFS-TR-008.

The CDA transform has been fully implemented within Amber and is available as a feature transform component. For more details of the functioning of the module, please consult the Amber Reference Guide.

Author: Darren Toulson

Revision: v 1.00 4 January, 1997

See ALSO: Theoretical Reports IFS-TR-005, IFS-TR-007, IFS-TR-008 The Amber Reference Guide

Bibliography

- [1] R. O. Duda P. E. Hart **Pattern Classification and Scene Analysis** Wiley, 1973.
- [2] K Fukunaga **Statistical Pattern Recognition (2nd Edition)** Academic Press, 1990.