LINEAR DISCRIMINANT ANALYSIS - A BRIEF TUTORIAL

S. Balakrishnama, A. Ganapathiraju

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Box 9571, 216 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-8335, Fax: 601-325-3149
Email: {balakris, ganapath}@isip.msstate.edu

1. INTRODUCTION

There are many possible techniques for classification of data. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two commonly used techniques for data classification and dimensionality reduction. Linear Discriminant Analysis easily handles the case where the within-class frequencies are unequal and their performances has been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. The use of Linear Discriminant Analysis for data classification is applied to classification problem in speech recognition.We decided to implement an algorithm for LDA in hopes of providing better classification compared to Principal Components Analysis. The prime difference between LDA and PCA is that PCA does more of feature classification and LDA does data classification. In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA doesn't change the location but only tries to provide more class separability and draw a decision region between the given classes.This method also helps to better understand the distribution of the feature data. Figure 1 will be used as an example to explain and illustrate the theory of LDA.

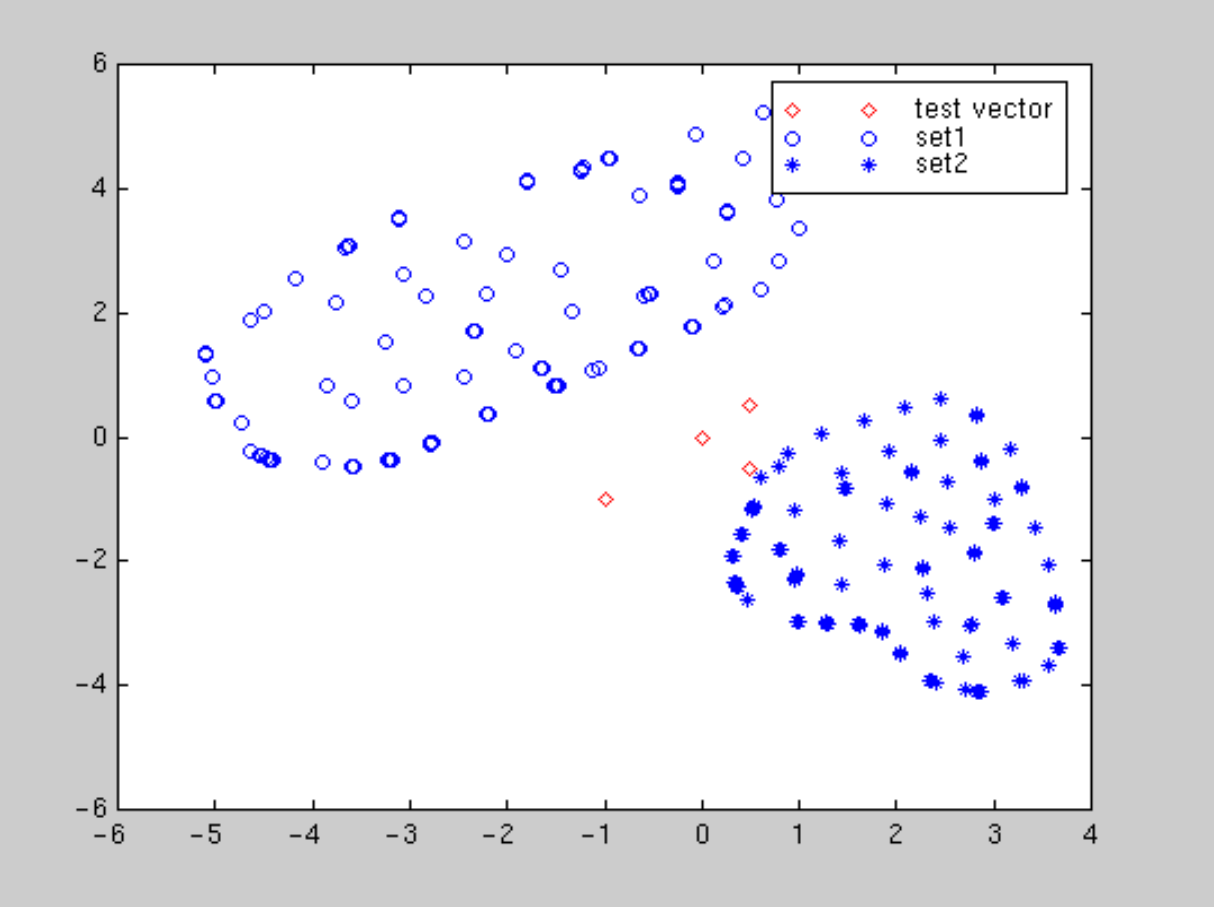

Figure 1. Figure showing data sets and test vectors in original

2. DIFFERENT APPROACHES TO LDA

Data sets can be transformed and test vectors can be classified in the transformed space by two different approaches.

Class-dependent transformation: This type of approach involves maximizing the ratio of between class variance to within class variance. The main objective is to maximize this ratio so that adequate class separability is obtained. The class-specific type approach involves using two optimizing criteria for transforming the data sets independently.

Class-independent transformation: This approach involves maximizing the ratio of overall variance to within class variance. This approach uses only one optimizing criterion to transform the data sets and hence all data points irrespective of their class identity are transformed using this transform. In this type of LDA, each class is considered as a separate class against all other classes.

3. MATHEMATICAL OPERATIONS

In this section, the mathematical operations involved in using LDA will be analyzed the aid of sample set in Figure 1. For ease of understanding, this concept is applied to a two-class problem. Each data set has 100 2-D data points. Note that the mathematical formulation of this classification strategy parallels the Matlab implementation associated with this work.

1. Formulate the data sets and the test sets, which are to be classified in the original space. The given data sets and the test vectors are formulated, a graphical plot of the data sets and test vectors for the example considered in original space is shown in Figure 1. For ease of understanding let us represent the data sets as a matrix consisting of features in the form given below:

$$set1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \dots & \dots \\ \dots & \dots \\ a_{m1} & a_{m2} \end{bmatrix} \qquad set2 = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ \dots & \dots \\ \dots & \dots \\ b_{m1} & b_{m2} \end{bmatrix} \tag{1}$$

2. Compute the mean of each data set and mean of entire data set. Let μ_1 and μ_2 be the mean of set 1 and set 2 respectively and μ_3 be mean of entire data, which is obtained by merging set 1 and set 2, is given by Equation 1.

$$\mu_3 = p_1 \times \mu_1 + p_2 \times \mu_2 \tag{2}$$

where p_1 and p_2 are the apriori probabilities of the classes. In the case of this simple two class problem, the probability factor is assumed to be 0.5.

3. In LDA, within-class and between-class scatter are used to formulate criteria for class separability. Within-class scatter is the expected covariance of each of the classes. The scatter measures are computed using Equations 3 and 4.

$$S_w = \sum_j p_j \times (cov_j) \tag{3}$$

Therefore, for the two-class problem,

$$S_w = 0.5 \times cov_1 + 0.5 \times cov_2 \tag{4}$$

All the covariance matrices are symmetric. Let $cov1$ and $cov2$ be the covariance of set 1 and set 2 respectively. Covariance matrix is computed using the following equation.

$$cov_j = (\boldsymbol{x}_j - \mu_j)(\boldsymbol{x}_j - \mu_j)^T \tag{5}$$

The between-class scatter is computes using the following equation.

$$S_b = \sum_j (\mu_j - \mu_3) \times (\mu_j - \mu_3)^T \tag{6}$$

Note that S_b can be thought of as the covariance of data set whose members are the mean vectors of each class. As defined earlier, the optimizing criterion in LDA is the ratio of between-class scatter to the within-class scatter. The solution obtained by maximizing this criterion defines the axes of the transformed space. However for the class-dependent transform the optimizing criterion is computed using equations (5) and (6). It should be noted that if the LDA is a class dependent type, for *L*-class L separate optimizing criterion are required for each class. The optimizing factors in case of class dependent type are computed as

$$criterion_j = inv(cov_j) \times S_b \tag{7}$$

For the class independent transform, the optimizing criterion is computed as

$$criterion = inv(S_w) \times S_b \tag{8}$$

4. By definition, an eigen vector of a transformation represents a 1-D invariant subspace of the vector space in which the transformation is applied. A set of these eigen vectors whose corresponding eigen values are non-zero are all linearly independent and are invariant under the transformation. Thus any vector space can be represented in terms of linear combinations of the eigen vectors. A linear dependency between features is indicated by a

zero eigen value. To obtain a non-redundant set of features all eigen vectors corresponding to non-zero eigen values only are considered and the ones corresponding to zero eigen values are neglected. In the case of LDA, the transformations are found as the eigen vector matrix of the different criteria defined in Equations 7 and 8.

5. For any *L*-class problem we would always have *L-1* non-zero eigen values. This is attributed to the constraints on the mean vectors of the classes in Equation 2. The eigen vectors corresponding to non-zero eigen values for the definition of the transformation.

   For our 2-class example, Figures 2 and 3 show the direction of the significant eigen vector along which there is maximum discrimination information. Having obtained the transformation matrices, we transform the data sets using the single LDA transform or the class specific transforms which ever the case may be. From the figures it can be observed that, transforming the entire data set to one axis provides definite boundaries to classify the data. The decision region in the transformed space is a solid line separating the transformed data sets thus

   For the class dependent LDA,

$$transformed\_set\_j = transform\_j^{T} \times set\_j \tag{9}$$

   For the class independent LDA,

$$transformed\_set = transform\_spec^{T} \times data\_set^{T} \tag{10}$$

   Similarly the test vectors are transformed and are classified using the euclidean distance of the test vectors from each class mean.

   The two Figures 4 and 5 clearly illustrate the theory of Linear Discriminant Analysis applied to a 2-class problem. The original data sets are shown and the same data sets after transformation are also illustrated. It is quite clear from these figures that transformation provides a boundary for proper classification. In this example the classes were properly defined but cases where there is overlap between classes, obtaining a decision region in original space will be very difficult and in such cases transformation proves to be very essential. Transformation along largest eigen vector axis is the best transformation.

   Figures 6 and 7, are interesting in that they show how the linear transformation process can be viewed as projecting data points onto the maximally discriminating axes represented by the eigen vectors.

6. Once the transformations are completed using the LDA transforms, Euclidean distance or RMS distance is used to classify data points. Euclidean distance is computed using Equation 11 where μ_{ntrans} is the mean of the transformed data set, n is the class index and $\boldsymbol{x}$ is the test vector. Thus for n classes, n euclidean distances are obtained for each test point.

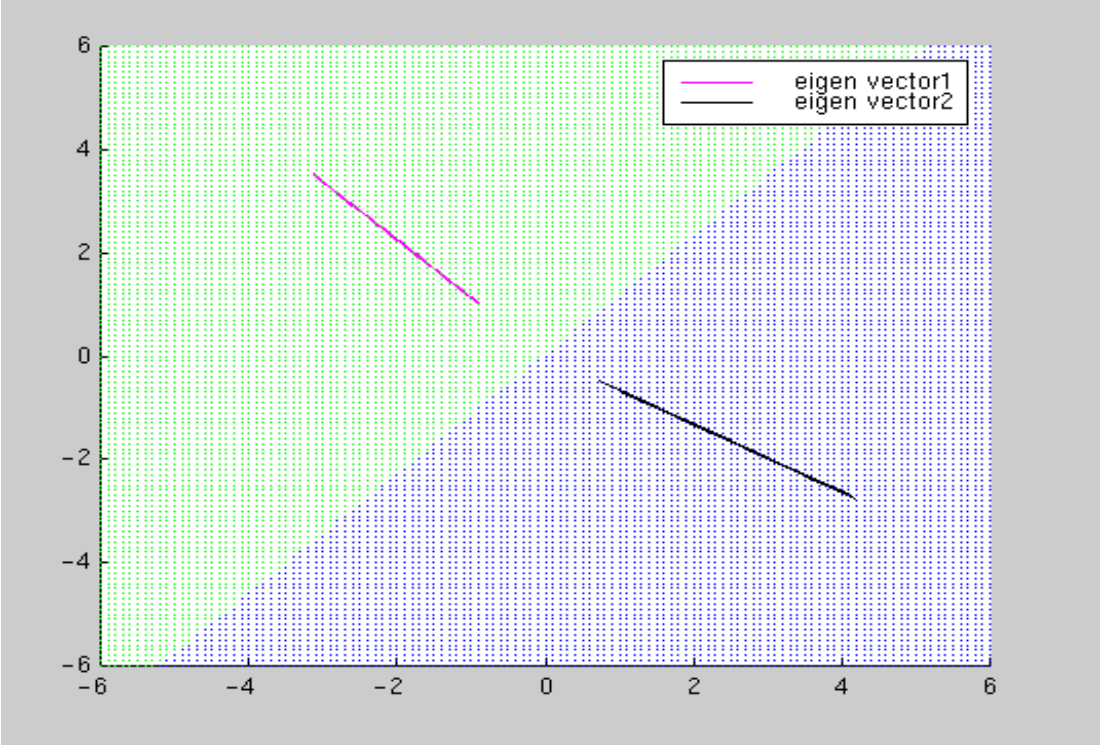

Figure 2. Figure for eigen vector direction in class dependent type

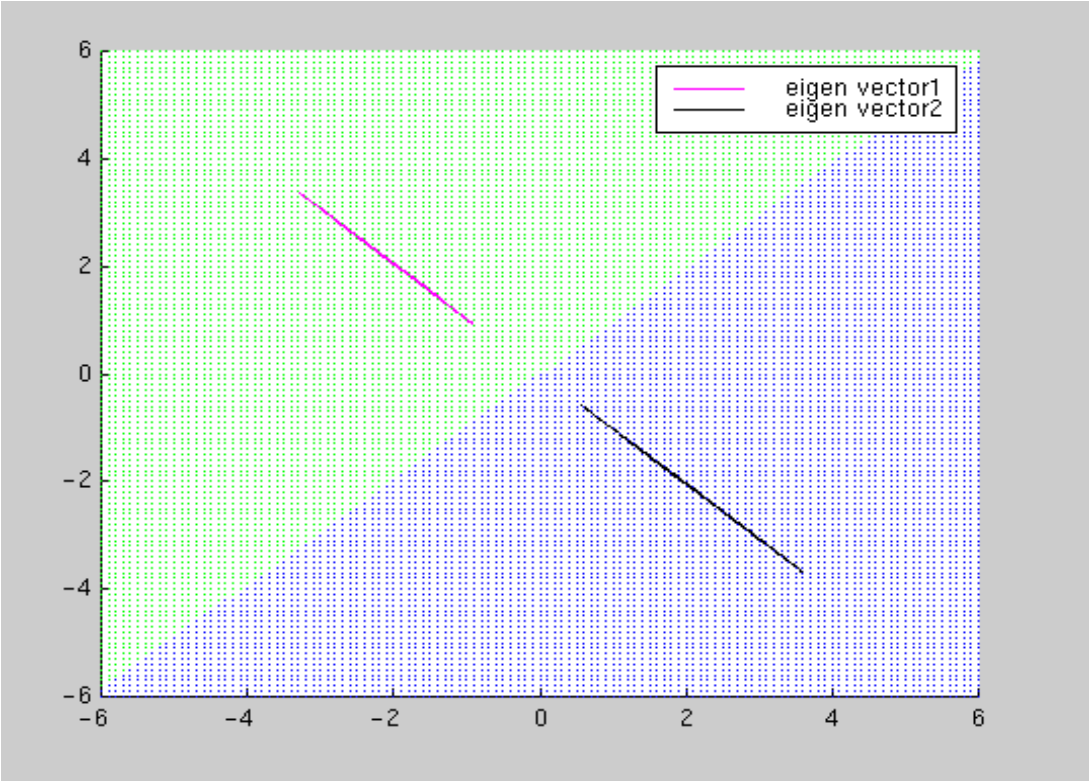

Figure 3. Figure for eigen vector direction in class independent type

$$dist\_n = (transform\_n\_spec)^T \times \boldsymbol{x} - \mu_{ntrans} \quad (11)$$

7. The smallest Euclidean distance among the n distances classifies the test vector as belonging to class n.

4. CONCLUSIONS

We have presented the theory and implementation of LDA as a classification technique. Throughout the tutorial we have used a 2-class problem as an exemplar. Two approaches to LDA, namely, class independent and class dependent, have been explained. The choice of the type of LDA depends on the data set and the goals of the classification problem. If generalization is of importance, the class independent transformation is preferred. However, if good discrimination is what is aimed for, the class dependent type should be the first choice. As part of our future work, we plan to work on a Java-based demonstration which could be used to visualize LDA based transformations on user defined data sets and also help the user appreciate the difference between the various classification techniques.

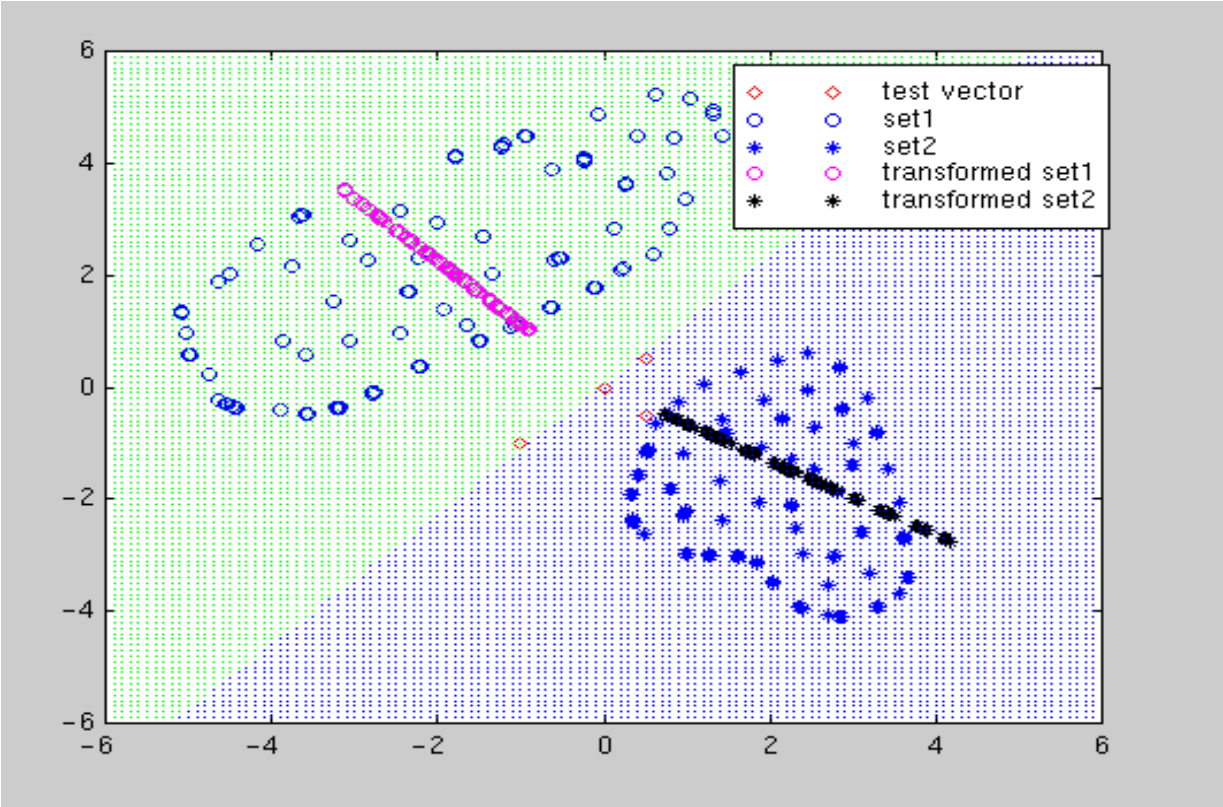

Figure 4. Data sets in original space and transformed space along with the tranformation axis for class dependent LDA of a 2-class problem

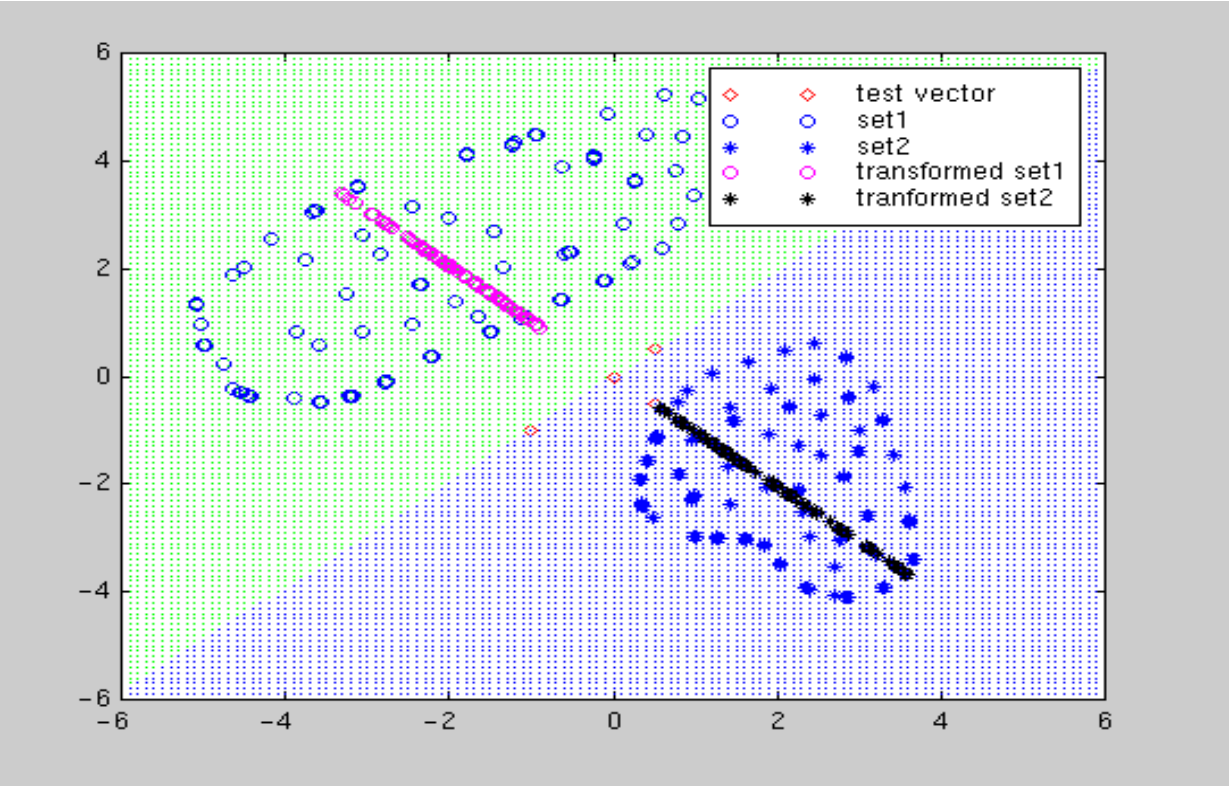

Figure 5. Data sets in original space and transformed space for class independent type of LDA of a 2-class problem

5. REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, California, 1990.

[2] S. Axler, *Linear Algebra Done Right*, Springer-Verlag New York Inc., New York, New York, 1995.

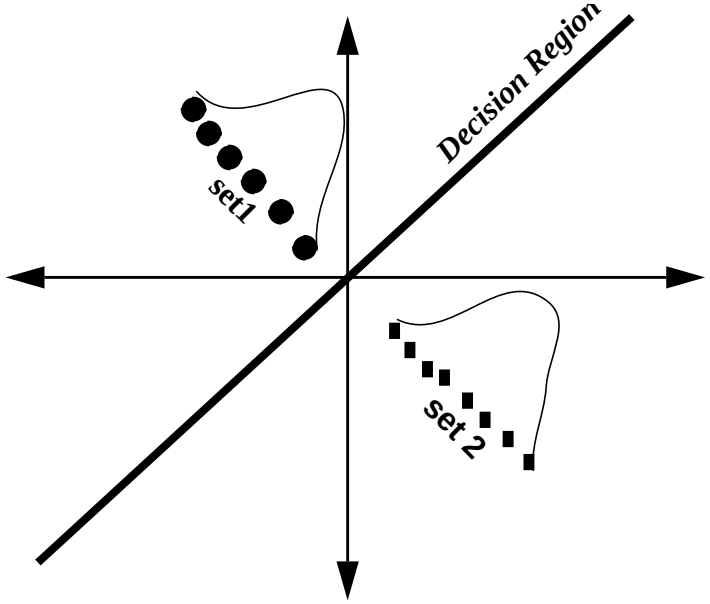

Figure 6. Figure showing histogram plot of transformed data with decision region in class independent type and the amount of class separability obtained in transformed space

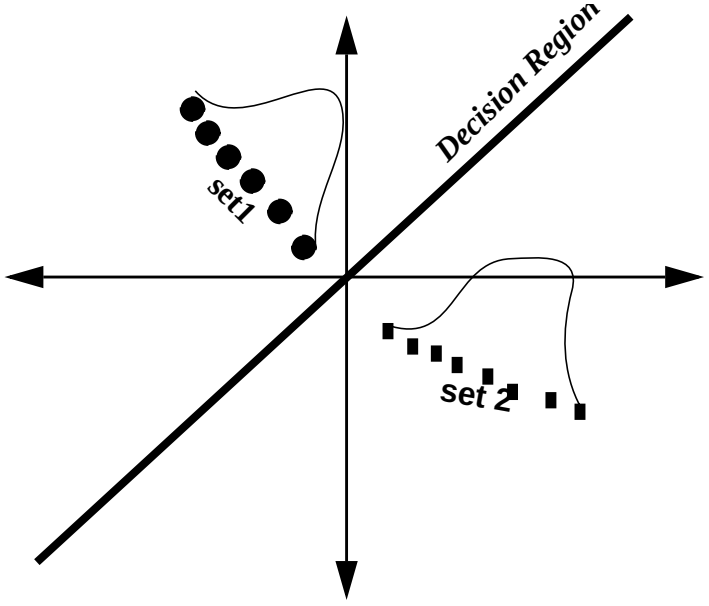

Figure 7. Histogram plot of transformed data with decision region in class dependent type and the amount of class separability obtained in transformed space