Home Page

Title Page

Go Back

Full Screen

Close

Quit

Comparing several methods of Discriminant Analysis on the case of Wine Data *

Dimitar Vandev, Ute Römisch

Sofia University, TU - Berlin

Description of data
Traditional Methods
Features space
Software used
Results and conclusion
Home Page
Title Page
Go Back
Full Screen
Close
Quit

Comparing several methods of Discriminant Analysis on the case of Wine Data *

Dimitar Vandev, Ute Römisch

Sofia University, TU - Berlin

Abstract

Description of data
Traditional Methods
Features space
Software used
Results and conclusion
Home Page
Title Page
Go Back
Full Screen
Close
Quit

Comparing several methods of Discriminant Analysis on the case of Wine Data *

Dimitar Vandev, Ute Römisch

Sofia University, TU - Berlin

Abstract

We shortly describe the type of data collected in WINE-DB project and the problems of recognition which has to be solved. Then the procedures of Linear and Quadratic Discriminant analysis as well as a small improvement - mixture of both models are described.

General Discriminant Analysis is a nonparametric procedure. Support Vector Mashines (also known as Kernel Mashines) are procedures from the field of Mashine Learning.

We test these techniques on our data and comment the results.

*The research is supported by contracts: PRO-ENBIS: GTC1-2001-43031 and WINE DB: G6RD-CT-2001-00646

Description of data

Traditional Methods

Features space

Software used

Results and conclusion

Home Page

Title Page

Go Back

Full Screen

Close

Quit

1. Description of data

1.1. Two data sets

1. Description of data

1.1. Two data sets

1. East European wines: `TranWein35.sta`: 35 variables by 144 cases, from 5 countries:
   (in German spelling) Bulgarien, Rumänien, Ungarn, Mazedonien, Moldawien

1. Description of data

1.1. Two data sets

1. East European wines: `TranWein35.sta`: 35 variables by 144 cases, from 5 countries:
   (in German spelling) Bulgarien, Rumänien, Ungarn, Mazedonien, Moldawien

2. Oversee wines: `usw_tran_2.sta`: 45 variables by 274 cases from:
   (in German spelling) Kalifornien, Südafrika, Australien, Chile, Argentinien

1.2. Preliminary Data Processing

1. Description of data

1.1. Two data sets

1. East European wines: `TranWein35.sta`: 35 variables by 144 cases, from 5 countries:
   (in German spelling) Bulgarien, Rumänien, Ungarn, Mazedonien, Moldawien
2. Oversee wines: `usw_tran_2.sta`: 45 variables by 274 cases from:
   (in German spelling) Kalifornien, Südafrika, Australien, Chile, Argentinien

1.2. Preliminary Data Processing

- Transformations of some variables to normality.

1. Description of data

1.1. Two data sets

1. East European wines: `TranWein35.sta`: 35 variables by 144 cases, from 5 countries:
   (in German spelling) Bulgarien, Rumänien, Ungarn, Mazedonien, Moldawien

2. Oversee wines: `usw_tran_2.sta`: 45 variables by 274 cases from:
   (in German spelling) Kalifornien, Südafrika, Australien, Chile, Argentinien

1.2. Preliminary Data Processing

- Transformations of some variables to normality.
- The existent missing values was filled with within groups means.

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

1. continuous ξ with values $x \in R^p$;

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

1. continuous ξ with values $x \in R^p$;
2. discrete (or categorical) η with values $y \in \{1, 2, \ldots, G\}$.

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

1. continuous ξ with values $x \in R^p$;
2. discrete (or categorical) η with values $y \in \{1, 2, \ldots, G\}$.
3. they have joined distribution (DA model):

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

1. continuous ξ with values $x \in R^p$;
2. discrete (or categorical) η with values $y \in \{1, 2, \ldots, G\}$.
3. they have joined distribution (DA model):

- $\Pr(\eta = y) = p_y$

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

1. continuous ξ with values $x \in R^p$;
2. discrete (or categorical) η with values $y \in \{1, 2, \ldots, G\}$.
3. they have joined distribution (DA model):

- $\Pr(\eta = y) = p_y$
- Conditional distribution of $\xi \in R^p$ given $\eta = y$ is described by the density $\varphi(x, m_y, C_y)$.

Here φ is the density of Gauss distribution in R^p described by two parameters:

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

1. continuous ξ with values $x \in R^p$;
2. discrete (or categorical) η with values $y \in \{1, 2, \ldots, G\}$.
3. they have joined distribution (DA model):
    - $\Pr(\eta = y) = p_y$
    - Conditional distribution of $\xi \in R^p$ given $\eta = y$ is described by the density $\varphi(x, m_y, C_y)$.

Here φ is the density of Gauss distribution in R^p described by two parameters:

- mean - m_y;

2. Traditional Methods

2.1. Bayes Discriminant Analysis

Let us suppose that we have observed two random variables:

1. continuous ξ with values $x \in R^p$;
2. discrete (or categorical) η with values $y \in \{1, 2, \dots, G\}$.
3. they have joined distribution (DA model):

   - $\Pr(\eta = y) = p_y$
   - Conditional distribution of $\xi \in R^p$ given $\eta = y$ is described by the density $\varphi(x, m_y, C_y)$.

Here φ is the density of Gauss distribution in R^p described by two parameters:

- mean - m_y;
- covariance - C_y,

Suppose we know the parameters of this model:

Suppose we know the parameters of this model:

1. The prior probabilities - $\{p_y\}$;

Suppose we know the parameters of this model:

1. The prior probabilities - $\{p_y\}$;
2. Group means - $\{m_y\}$;

Suppose we know the parameters of this model:

1. The prior probabilities - $\{p_y\}$;
2. Group means - $\{m_y\}$;
3. Within group covariance matrices - C_y;

That is, the set of numbers: $\{p_y, m_y, C_y, y = 1, 2, \ldots \ldots, G\}$.

Suppose we know the parameters of this model:

1. The prior probabilities - $\{p_y\}$;
2. Group means - $\{m_y\}$;
3. Within group covariance matrices - C_y;

That is, the set of numbers: $\{p_y, m_y, C_y, y = 1, 2, \ldots\ldots, G\}$.

Then according of the famous formula of Bayes we may write down the conditional probability of $\eta = y$ given x:

$$\Pr(\eta = y|\xi = y) = q(y|x) = c(x).p_y.\varphi(x, m_y, C_y),$$

Suppose we know the parameters of this model:

1. The prior probabilities - $\{p_y\}$;
2. Group means - $\{m_y\}$;
3. Within group covariance matrices - C_y;

That is, the set of numbers: $\{p_y, m_y, C_y, y = 1, 2, \ldots \ldots, G\}$.

Then according of the famous formula of Bayes we may write down the conditional probability of $\eta = y$ given x:

$$\Pr(\eta = y | \xi = y) = q(y|x) = c(x).p_y.\varphi(x, m_y, C_y), \qquad (1)$$

where c is a normalizing constant, such that $\sum q(y|x) = 1$.

Suppose we know the parameters of this model:

1. The prior probabilities - $\{p_y\}$;
2. Group means - $\{m_y\}$;
3. Within group covariance matrices - C_y;

That is, the set of numbers: $\{p_y, m_y, C_y, y = 1, 2, \ldots\ldots, G\}$.

Then according of the famous formula of Bayes we may write down the conditional probability of $\eta = y$ given x:

$$\Pr(\eta = y|\xi = y) = q(y|x) = c(x).p_y.\varphi(x, m_y, C_y), \qquad (1)$$

where c is a normalizing constant, such that $\sum q(y|x) = 1$.

We call this probability posterior and say that the observation x belongs to the group y with probability $q(y|x)$.

According the maximum likelihood principle the classification rule should then be:

$$\widehat{y(x)} = \underset{h}{argmax} : q(h|x).$$

Suppose we know the parameters of this model:

1. The prior probabilities - $\{p_y\}$;
2. Group means - $\{m_y\}$;
3. Within group covariance matrices - C_y;

That is, the set of numbers: $\{p_y, m_y, C_y, y = 1, 2, \ldots\ldots, G\}$.

Then according of the famous formula of Bayes we may write down the conditional probability of $\eta = y$ given x:

$$\Pr(\eta = y|\xi = y) = q(y|x) = c(x).p_y.\varphi(x, m_y, C_y), \tag{1}$$

where c is a normalizing constant, such that $\sum q(y|x) = 1$.

We call this probability posterior and say that the observation x belongs to the group y with probability $q(y|x)$.

According the maximum likelihood principle the classification rule should then be:

$$\widehat{y(x)} = \underset{h}{argmax} : q(h|x). \tag{2}$$

2.2. Linear and Quadratic DA

Suppose that within group covariances $C(g)$ are equal:

$$C(g) = C, \quad (g = 1, 2, \ldots, G)$$

2.2. Linear and Quadratic DA

Suppose that within group covariances $C(g)$ are equal:

$$C(g) = C, \quad (g = 1, 2, \ldots, G) \tag{3}$$

Than the maximum likelihood rule (2) becomes a set of inequalities:

$$p(\widehat{g}).f(x, m(\widehat{g}), C) \geq p(h).f(x, m(h), C), .$$

2.2. Linear and Quadratic DA

Suppose that within group covariances $C(g)$ are equal:

$$C(g) = C, \quad (g = 1, 2, \ldots, G) \tag{3}$$

Than the maximum likelihood rule (2) becomes a set of inequalities:

$$p(\widehat{g}).f(x, m(\widehat{g}), C) \geq p(h).f(x, m(h), C), . \tag{4}$$

or (what is the same) to:

$$L_g(x) = b(\widehat{g})'x + a(\widehat{g}) \geq L_h(x) = b(h)'x + a(h), \tag{5}$$

We decide that the observation x belongs to the group g, if for each h the inequality (5) holds.

2.2. Linear and Quadratic DA

Suppose that within group covariances $C(g)$ are equal:

$$C(g) = C, \quad (g = 1, 2, \ldots, G) \tag{3}$$

Than the maximum likelihood rule (2) becomes a set of inequalities:

$$p(\widehat{g}).f(x, m(\widehat{g}), C) \geq p(h).f(x, m(h), C), . \tag{4}$$

or (what is the same) to:

$$L_g(x) = b(\widehat{g})'x + a(\widehat{g}) \geq L_h(x) = b(h)'x + a(h), \tag{5}$$

We decide that the observation x belongs to the group g, if for each h the inequality (5) holds.The functions L are called discriminant functions.

When the assumption (3):$C(g) = C$ is not appropriate, the corresponding functions become quadratic.

When the assumption (3):$C(g) = C$ is not appropriate, the corresponding functions become quadratic.

If one has equal prior probabilities $p(h) = 1/G$, the solution of the classification problem (2) is equivalent to the minimization of so called Mahalanobis distances of the observation to the group means:

$$h(x, g) = (x - m(g))' C_g^{-1} (x - m(g))$$

When the assumption (3):$C(g) = C$ is not appropriate, the corresponding functions become quadratic.

If one has equal prior probabilities $p(h) = 1/G$, the solution of the classification problem (2) is equivalent to the minimization of so called Mahalanobis distances of the observation to the group means:

$$h(x, g) = (x - m(g))' C_g^{-1} (x - m(g)) \tag{6}$$

One uses Mahalanobis distances (6) to classify the observation to the closest group (so called nearest neighbors method):

$$\widehat{g} = \underset{h}{argmin}\, h(x, h).$$

When the assumption (3):$C(g) = C$ is not appropriate, the corresponding functions become quadratic.

If one has equal prior probabilities $p(h) = 1/G$, the solution of the classification problem (2) is equivalent to the minimization of so called Mahalanobis distances of the observation to the group means:

$$h(x,g) = (x - m(g))' C_g^{-1} (x - m(g)) \tag{6}$$

One uses Mahalanobis distances (6) to classify the observation to the closest group (so called nearest neighbors method):

$$\widehat{g} = \underset{h}{argmin}\, h(x,h).$$

In general however, the Bayes rule (1) is better if the supposition of normal distribution is fulfilled and its parameters can be estimated.

Description of data
Traditional Methods
Features space
Software used
Results and conclusion
Home Page
Title Page
Go Back
Full Screen
Close
Quit

2.3. Nonparametric DA

One way to attempt to overcome this problem is to try to obtain an estimation of these densities by nonparametric methods.

2.3. Nonparametric DA

One way to attempt to overcome this problem is to try to obtain an estimation of these densities by nonparametric methods.

Indeed, recently much attention has been given to the application of nonparametric methods in the classification problem, including methods such as neural networks (Ripley, 1994), classification and regression trees (Breiman et al., 1984), flexible discriminant analysis (Hastie, Tibshirani and Buja (1994)) and multivariate adaptive regression splines (Friedman (1991)).

2.3. Nonparametric DA

One way to attempt to overcome this problem is to try to obtain an estimation of these densities by nonparametric methods.

Indeed, recently much attention has been given to the application of nonparametric methods in the classification problem, including methods such as neural networks (Ripley, 1994), classification and regression trees (Breiman et al., 1984), flexible discriminant analysis (Hastie, Tibshirani and Buja (1994)) and multivariate adaptive regression splines (Friedman (1991)).

2.4. Independent Component Discriminant Analysis

(Amato, Antoniadis et al., 2002; Alfano, Amato et al., 2002) proposed so called ICDA - a nonparametric discriminant analysis method that is a simple generalization of the model assumed by linear and quadratic discriminant analysis. This generalization relies upon a transformation of the data based on independent component analysis (ICA), a statistical method for transforming an observed multivariate vector into components that are stochastically as independent as possible from each other. ICA was proposed in (Hyvärinen, 1997) and an algorithm in (Hyvärinen, 1999).

3. Features space

This section (see (Navarrete and del Solar, 2002)) is focused on the so called features space and methods connected with its use.

3. Features space

This section (see (Navarrete and del Solar, 2002)) is focused on the so called features space and methods connected with its use.

3.1. Kernels approach and features space

The set of vectors $\vec{x}_1, ..., \vec{x}_n \in R^n$, (our observations) is mapped into a feature space F by a set of functions $\{\Phi_j(\vec{x}), j = 1, ..., M\}$.

3. Features space

This section (see (Navarrete and del Solar, 2002)) is focused on the so called features space and methods connected with its use.

3.1. Kernels approach and features space

The set of vectors $\vec{x}_1, ..., \vec{x}_n \in R^n$, (our observations) is mapped into a feature space F by a set of functions $\{\Phi_j(\vec{x}), j = 1, ..., M\}$.

It is better that these functions are eigenfunctions of a given kernel (i.e., satisfying the Mercer's condition).

3. Features space

This section (see (Navarrete and del Solar, 2002)) is focused on the so called features space and methods connected with its use.

3.1. Kernels approach and features space

The set of vectors $\vec{x}_1, ..., \vec{x}_n \in R^n$, (our observations) is mapped into a feature space F by a set of functions $\{\Phi_j(\vec{x}), j = 1, ..., M\}$.

It is better that these functions are eigenfunctions of a given kernel (i.e., satisfying the Mercer's condition).

We suppose that $M > p$. In fact, this is an important purpose of kernel machines in order to give a good generalization ability to the system (Vapnik, 1995).

3. Features space

This section (see (Navarrete and del Solar, 2002)) is focused on the so called features space and methods connected with its use.

3.1. Kernels approach and features space

The set of vectors $\vec{x}_1, ..., \vec{x}_n \in R^n$, (our observations) is mapped into a feature space F by a set of functions $\{\Phi_j(\vec{x}), j = 1, ..., M\}$.

It is better that these functions are eigenfunctions of a given kernel (i.e., satisfying the Mercer's condition).

We suppose that $M > p$. In fact, this is an important purpose of kernel machines in order to give a good generalization ability to the system (Vapnik, 1995).

The aim of kernel machines is to work with the set of mapped vectors: $\Phi(x_i)$. Denote by Φ the matrix composed by them $\Phi = \{\Phi(\vec{x}_1), \ldots, \Phi(\vec{x}_n)\}$. Then, the correlation matrix of vectors Φ is defined as:

$$R = \frac{1}{n-1}\Phi\Phi' \tag{7}$$

The Fundamental Correlation Problem (FCP) for the matrix R, in its Primal form, consists in solving the eigensystem:

$$Rw_k = \lambda_k w_k, \quad ||w_k|| = 1, \quad k = 1, \ldots, M \tag{8}$$

However, R is an uncomputable matrix and then (8) cannot be solved. In this situation we need to introduce the Dual form of the Fundamental Correlation Problem for R:

$$Kv_k = \lambda_k v_k, \quad ||v_k|| = 1, \quad k = 1, \ldots, n, \tag{9}$$

where K is so called kernel matrix:

$$K = \frac{1}{n-1}\Phi'\Phi. \tag{10}$$

This can be shown by pre-multiplying (9) by Φ, and using (10).

The Fundamental Correlation Problem (FCP) for the matrix R, in its Primal form, consists in solving the eigensystem:

$$Rw_k = \lambda_k w_k, \quad ||w_k|| = 1, \quad k = 1, \dots, M \tag{8}$$

However, R is an uncomputable matrix and then (8) cannot be solved. In this situation we need to introduce the Dual form of the Fundamental Correlation Problem for R:

$$Kv_k = \lambda_k v_k, \quad ||v_k|| = 1, \quad k = 1, \dots, n, \tag{9}$$

where K is so called kernel matrix:

$$K = \frac{1}{n-1}\Phi'\Phi. \tag{10}$$

This can be shown by pre-multiplying (9) by Φ, and using (10).

The kernel function, $k(\vec{x}, \vec{x}')$ specify an inner product in the feature space

$$\Phi(\vec{x}).\Phi(\vec{x}') = k(\vec{x}, \vec{x}').$$

As we want to compute the solutions for which $\lambda_k > 0, k = 1, \ldots,$ we can go further and write the expression:

$$w_k = \frac{1}{\sqrt{\lambda_k(n-1)}}\Phi v_k, \quad k = 1, \ldots, q. \tag{11}$$

We are going to see that the solution of a general kind of kernel machines can be written in terms of K, and then we are going to call it the Fundamental Kernel Matrix (FKM).

As we want to compute the solutions for which $\lambda_k > 0, k = 1, ...,$ we can go further and write the expression:

$$w_k = \frac{1}{\sqrt{\lambda_k(n-1)}}\Phi v_k, \quad k = 1, \ldots, q. \tag{11}$$

We are going to see that the solution of a general kind of kernel machines can be written in terms of K, and then we are going to call it the Fundamental Kernel Matrix (FKM).

Dual FCP is also an ill-posed problem, and requires some kind of regularization as well. For the same reason the eigenvalues of R will decay gradually to zero, and then we need to use some criterion in order to determine q. An appropriate criterion is to choose q such that the sum of the unused eigenvalues is less than some fixed percentage (e.g. 5%) of the sum of the entire set (residual mean square error). Then, using (11), the set of primal eigenvectors $R\ W \in M^{M\times q}$ can be written as:

$$W = \frac{1}{n-1}\Phi V \Lambda^{-1/2}. \tag{12}$$

Here the matrix V and the diagonal matrix Λ are correspondingly q-truncated.

3.2. Support Vector Classification

The support vector machine (Boser, Guyon et al., 1992; Cortes and Vapnik, 1995), given labelled training data

$$\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{n}, \quad \vec{x}_i \in \vec{X} \subset \mathbb{R}^d, \quad y_i \in \vec{Y} = \{-1, +1\},$$

constructs a maximal margin linear classifier in a high dimensional feature space, $\Phi(\vec{x})$, defined by a positive definite kernel function.

3.2. Support Vector Classification

The support vector machine (Boser, Guyon et al., 1992; Cortes and Vapnik, 1995), given labelled training data

$$\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{n}, \quad \vec{x}_i \in \vec{X} \subset \mathbb{R}^d, \quad y_i \in \vec{Y} = \{-1, +1\},$$

constructs a maximal margin linear classifier in a high dimensional feature space, $\Phi(\vec{x})$, defined by a positive definite kernel function.

A common kernel is the Gaussian radial basis function (RBF),

$$k(\vec{x}, \vec{x}') = e^{-||\vec{x}-\vec{x}'||^2/2\sigma^2}.$$

3.2. Support Vector Classification

The support vector machine (Boser, Guyon et al., 1992; Cortes and Vapnik, 1995), given labelled training data

$$\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{n}, \quad \vec{x}_i \in \vec{X} \subset \mathbb{R}^d, \quad y_i \in \vec{Y} = \{-1, +1\},$$

constructs a maximal margin linear classifier in a high dimensional feature space, $\Phi(\vec{x})$, defined by a positive definite kernel function.

A common kernel is the Gaussian radial basis function (RBF),

$$k(\vec{x}, \vec{x}') = e^{-||\vec{x}-\vec{x}'||^2/2\sigma^2}.$$

The function implemented by a support vector machine is given by

$$f(\vec{x}) = \left\{ \sum_{i=1}^{n} \alpha_i y_i k(\vec{x}_i, \vec{x}) \right\} - b. \tag{13}$$

That is if we consider the two classes $I = \{i : y_i = 1\}$ and $\overline{I} = \{i : y_i = -1\}$ the equation (13) may be rewritten as definition of two functions ("densities"):

$$
\begin{aligned}
f_I(\vec{x}) &= \left\{ \sum_{i \in I} \alpha_i k(\vec{x}_i, \vec{x}) \right\} \\
f_{\overline{I}}(\vec{x}) &= \left\{ \sum_{i \in \overline{I}} \alpha_i k(\vec{x}_i, \vec{x}) \right\}.
\end{aligned}
$$

Thus the problem is like a nonparametric DA problem. The observation is classified into the class with higher density.

Description of data
Traditional Methods
Features space
Software used
Results and conclusion
Home Page
Title Page
Go Back
Full Screen
Close
Quit

To find the optimal coefficients, $\vec{\alpha}$, of the expansion (13) it is sufficient to maximise the functional,

$$W(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(\vec{x}_i, \vec{x}_j), \tag{14}$$

in the non-negative quadrant,

$$0 \leq \alpha_i \leq C, \qquad i = 1, \ldots, n, \tag{15}$$

subject to the constraint,

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{16}$$

C is a regularisation parameter, controlling a compromise between maximising the margin and minimising the number of training set errors.

The Karush-Kuhn-Tucker (KKT) conditions can be stated as follows:

$$\alpha_i = 0 \implies y_i f(\vec{x}_i) \geq 1, \tag{17}$$
$$0 < \alpha_i < C \implies y_i f(\vec{x}_i) = 1, \tag{18}$$
$$\alpha_i = C \implies y_i f(\vec{x}_i) \leq 1. \tag{19}$$

These conditions are satisfied for the set of feasible Lagrange multipliers, $\vec{\alpha}^0 = \{\alpha_1^0, \alpha_2^0, \ldots, \alpha_n^0\}$, maximising the objective function given by equation 14. The bias parameter, b, is selected to ensure that the second KKT condition is satisfied for all input patterns corresponding to non-bound Lagrange multipliers.

Note that in general only a limited number of Lagrange multipliers, $\vec{\alpha}$, will have non-zero values; the corresponding input patterns are known as support vectors. Let $\mathcal{I}$ be the set of indices of patterns corresponding to non-bound Lagrange multipliers,

$$\mathcal{I} = \{i \ : \ 0 < \alpha_i^0 < C\},$$

and similarly, let $\mathcal{J}$ be the set of indices of patterns with Lagrange multipliers at the upper bound C,

$$\mathcal{J} = \{i \ : \ \alpha_i^0 = C\}.$$

Note that in general only a limited number of Lagrange multipliers, $\vec{\alpha}$, will have non-zero values; the corresponding input patterns are known as support vectors. Let $\mathcal{I}$ be the set of indices of patterns corresponding to non-bound Lagrange multipliers,

$$\mathcal{I} = \{i \ : \ 0 < \alpha_i^0 < C\},$$

and similarly, let $\mathcal{J}$ be the set of indices of patterns with Lagrange multipliers at the upper bound C,

$$\mathcal{J} = \{i \ : \ \alpha_i^0 = C\}.$$

Equation 13 can then be written as an expansion over support vectors,

$$f(\vec{x}) = \left\{ \sum_{i \in \{\mathcal{I}, \mathcal{J}\}} \alpha_i^0 y_i k(\vec{x}_i, \vec{x}) \right\} - b. \tag{20}$$

For a full exposition of the support vector method, see the any of the excellent books (Vapnik, 1995; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000).

3.2.1. Multiclass Strategies in SVM

For multiclass discriminant analysis problems there are some additional steps to be performed. One has to extend the two–class solution given above to this case.

3.2.1. Multiclass Strategies in SVM

For multiclass discriminant analysis problems there are some additional steps to be performed. One has to extend the two–class solution given above to this case.

- One-against-all,The earliest used implementation for SVM multiclass classication is probably the one-against-all method (for example, (Bottou, Cortes et al., 1994)). It constructs G SVM models where G is the number of classes. The i-th SVM is trained with all of the samples in the i-th class with positive labels, and all other samples with negative labels.

3.2.1. Multiclass Strategies in SVM

For multiclass discriminant analysis problems there are some additional steps to be performed. One has to extend the two–class solution given above to this case.

- One-against-all,The earliest used implementation for SVM multiclass classication is probably the one-against-all method (for example, (Bottou, Cortes et al., 1994)). It constructs G SVM models where G is the number of classes. The i-th SVM is trained with all of the samples in the i-th class with positive labels, and all other samples with negative labels.

- One-against-one For each pair of classes i and j a classification model is created. Then for the test sample the class with largest number of votes wins.

4. Software used

4. Software used

- Our program `LDAgui` for Linear and Quadratic DA (LQDA) , (D.Vandev,)

4. Software used

- Our program `LDAgui` for Linear and Quadratic DA (LQDA) , (D.Vandev,)

- Generalised DA (GDA) (Baudat and Anouar, 2000) using PCA in the feature space.

4. Software used

- Our program `LDAgui` for Linear and Quadratic DA (LQDA) , (D.Vandev,)
- Generalised DA (GDA) (Baudat and Anouar, 2000) using PCA in the feature space.
- Support Vector Machine Toolbox (SVM) with renewed QP optimizer:

4. Software used

- Our program `LDAgui` for Linear and Quadratic DA (LQDA) , (D.Vandev,)
- Generalised DA (GDA) (Baudat and Anouar, 2000) using PCA in the feature space.
- Support Vector Machine Toolbox (SVM) with renewed QP optimizer:

```
Version 2.0-Aug-1998,  Support Vector Classification,
Steve Gunn (S.R.Gunn@ecs.soton.ac.uk)
Image Speech and Intelligent Systems Group,
University of Southampton
```

4. Software used

- Our program `LDAgui` for Linear and Quadratic DA (LQDA) , (D.Vandev,)
- Generalised DA (GDA) (Baudat and Anouar, 2000) using PCA in the feature space.
- Support Vector Machine Toolbox (SVM) with renewed QP optimizer:

```
Version 2.0-Aug-1998,  Support Vector Classification,
Steve Gunn (S.R.Gunn@ecs.soton.ac.uk)
Image Speech and Intelligent Systems Group,
University of Southampton
```

- LS-SVM Library (LSVM) (Chang and Lin, 2001) with One–To–One strategy for combining outputs of binary classifying.

All programs were feed with exactly the same training and test data sets.

5. Results and conclusion

All cited programs failed in comparison with QDA. When QLDA has 4-5%- errors over the test set, they achieved minimum of 17

The reason for such unexpectedly bad result may be in the fact that the test sets were generated with a model exactly the same as the model produced by QDA.

References

Alfano, B., Amato, U., et al. (2002). Segmentation of mr brain images through discriminant analysis. Tecnical Report 262, Istituto per le Applicazioni del Calcolo "Mauro Picone". URL http://www.iam.na.cnr.it/rapporti/2002/RT262.pdf

Amato, U., Antoniadis, A., et al. (2002). Independent component discriminant analysis. Tecnical Report 254, Istituto per le Applicazioni del Calcolo "Mauro Picone". URL http://www.iam.na.cnr.it/rapporti/2002/RT254_02.pdf

Baudat, G. and Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):pp. 2385–2404.

Boser, B., Guyon, I., et al. (1992). A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on computational learning theory*, pp. 144–152. ACM, Pittsburgh.

Bottou, L., Cortes, C., et al. (1994). Comparison of classier methods: a case study in handwriting digit recognition.

In: *International Conference on Pattern Recognition*, pp. 77–87. IEEE Computer Society Press.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20:pp. 1–25.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge University Press, Cambridge, U.K.

D.Vandev (). Interactive discriminant analysis in matlab. In: *Proceedings of the Seminar on Statistical Data Analysis 2003*.

Hyvärinen, A. (1997). Independent component analysis by minimization of mutual information. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, p. 3917–3920.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:pp. 626–634.

Navarrete, P. and del Solar, J. R. (2002). *Pattern Recognition with Support Vector Machines*, volume 2388 of *Lecture Notes*, chapter On the Generalization of Kernel Machines, pp. 24–39. Springer.
URL http://tamarugo.cec.uchile.cl/~jruizd/papers/svm2002.pdf

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York. ISBN 0-387-94559-8.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control. Wiley, New York.