Localized Linear Discriminant Analysis

Irina Czogiel[1], Karsten Luebke[2], Marc Zentgraf[2], and Claus Weihs[2]

[1] Graduiertenkolleg Statistische Modellbildung,
Lehrstuhl für Computergestützte Statistik,
Universität Dortmund, D-44221 Dortmund, Germany

[2] Lehrstuhl für Computergestützte Statistik,
Universität Dortmund, D-44221 Dortmund, Germany

Abstract. Despite its age, the Linear Discriminant Analysis performs well even in situations where the underlying premises like normally distributed data with constant covariance matrices over all classes are not met. It is, however, a global technique that does not regard the nature of an individual observation to be classified. By weighting each training observation according to its distance to the observation of interest, a global classifier can be transformed into an observation specific approach. So far, this has been done for logistic discrimination. By using LDA instead, the computation of the local classifier is much simpler. Moreover, it is ready for applications in multi-class situations.

Key words: classification, local models, LDA

1 Introduction

Statistical work on classification begins with the work proposed by Fisher (1936). For the dichotomous case, he suggests to reduce a multivariate classification problem to an univariate problem by linearly transforming the given observations into scalar values such that the separation of the transformed class means is maximized whilst the within class variances of the transformed observations are minimized. Although Fishers approach is distribution-free, it does implicitly assume that the covariance structure is the same in both classes, because a pooled estimate of the common covariance matrix is used. The resulting classification rule can alternatively be derived using Bayesian argumentation although here more restrictive assumptions are made: the data within each class are assumed to be normally distributed with class-specific means and a common covariance structure. Both approaches can be extended to multi-class situations and in each case, obtaining the actual decision functions for a given data set requires the estimation of the unknown model parameters, namely the class-specific means, the class priors, and the covariance matrix. Since the estimation is carried out without taking into account the nature of the problem at hand, i.e. the classification of a specific trial point, LDA can be considered a global classifier. Hand and Vinciotti (2003) argue that an approach like this can lead to poor results if the chosen model does not exactly reflect the underlying data generating process because then, a good fit in some parts of the data space may worsen the fit in other regions. Since in classification problems, accuracy is often not equally important throughout the entire data space, they suggest to improve the fit in regions where a good fit is crucial for obtaining satisfactory results – even if the fit elsewhere is degraded. For the dichotomous logistic discrimination, two approaches have been proposed to accomplish this. Hand and Vinciotti (2003) introduce a logistic regression model in which data points in the vicinity of the ideal decision surface are weighted more heavily than those which are far away. Another strategy is presented by Tutz and Binder (2005) who suggest to assign locally adaptive weights to each observation of the training set. By choosing the weights as decreasing in the (Euclidean) distance to the observation to be classified, and maximizing the corresponding weighted (log-)likelihood, a localized version of the logistic regression model can be obtained. The classifier is therefore adapted to the nature of each individual trial point which turns the global technique of logistic discrimination into an observation specific approach.

In this paper, we adopt the strategy of using locally adaptive weights to the context of LDA which comprises the advantage that localizing a classification rule can be accomplished without numerical methods like Fisher scoring. In the following we call this new approach LLDA (Localized Linear Discriminant Analysis). It will be proposed in Section 2. In Section 3, the benefit of LLDA will be shown on basis of a simulated data set containing local subclasses. The application of LLDA to the real-life problem of business phase classification is described in Section 4. A summary of the main results is provided in Section 5.

2 Localized Linear Discriminant Analysis

Let the training data consist of N observations $(\mathbf{x}_i, y_i)'$, where $\mathbf{x}_i \in \mathbb{R}^p$ is the set of explanatory variables for the ith observation and $y_i \in \{A_1, \ldots, A_G\}$ denotes the corresponding class membership. The objective now is to construct a classification rule on basis of the training sample which can then be used for predicting the unknown class of a new observation. In LDA, the classification is based on the posterior class probabilities of the considered trial point $\mathbf{x}$. To calculate these, the data is assumed to be normally distributed with class-specific mean vectors $\boldsymbol{\mu}_g$ and a common covariance matrix $\boldsymbol{\Sigma}$. Let π_g denote the prior probability of A_g. Choosing the class with the highest posterior probability for a given trial point $\mathbf{x}$ can then be shown to be equivalent to assigning $\mathbf{x}$ to the class with the largest value of the corresponding discriminant function

$$h_g(\mathbf{x}) = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_g\right)'\mathbf{x} - 0.5\,\boldsymbol{\mu}_g'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_g + \ln(\pi_g).$$

Since the parameters of the assumed normal distribution are usually unknown, in practice the sample analogues of h_g are used:

$$\hat{h}_g(\mathbf{x}) = \left(\mathbf{S}^{-1}\bar{\mathbf{x}}_g\right)'\mathbf{x} - 0.5\,\bar{\mathbf{x}}_g'\,\mathbf{S}^{-1}\bar{\mathbf{x}}_g + \ln(p_g), \tag{1}$$

where $\bar{\mathbf{x}}_g$ denotes the mean vector and p_g denotes the proportion of the training observations belonging to A_g. The matrix $\mathbf{S}$ is the pooled estimate of the covariance matrix.

A version of (1) which is adaptive to the nature of the considered trial point can be obtained by introducing weights $w_i = w\big(\mathbf{x}, \mathbf{x}_i\big)$ to the sample estimates. For the mean vector and the proportion, this can be formulated as

$$\bar{\mathbf{x}}_{g_L} = \frac{\sum_i w_i \mathbf{x}_i I_{\{\mathbf{x}_i \in A_g\}}}{\sum_i w_i I_{\{\mathbf{x}_i \in A_g\}}} \quad \text{and} \quad p_{g_L} = \frac{w_i \mathbf{x}_i I_{\{\mathbf{x}_i \in A_g\}}}{\sum_i w_i I_{\{\mathbf{x}_i \in A_g\}}}.$$

To compute an analogous variant of $\mathbf{S}$, first a weighted estimate of the covariance matrix is calculated for each class:

$$\mathbf{S}_{g_L} = \frac{1}{1 - \sum_i w_i^2 I_{\{\mathbf{x}_i \in A_g\}}} \sum_i w_i \big[(\mathbf{x}_i - \bar{\mathbf{x}}_{g_L}) I_{\{\mathbf{x}_i \in A_g\}}\big] \big[(\mathbf{x}_i - \bar{\mathbf{x}}_{g_L}) I_{\{\mathbf{x}_i \in A_g\}}\big]'.$$

These matrices are then weighted with the number of training observations of the corresponding class and aggregated to

$$\mathbf{S}_L = \frac{N}{N-G} \sum_g p_g \mathbf{S}_{g_L}.$$

As suggested by Tutz and Binder (2005), the weights are chosen to be locally adaptive in the sense that they depend on the Euclidean distance of the considered trial point $\mathbf{x}$ and the training observations $\mathbf{x}_i$. This can be accomplished by using a kernel window

$$w_i = K\big(||\mathbf{x} - \mathbf{x}_i||\big).$$

In this context, various kernels can be used and the performance of a kernel function of course depends on the nature of the problem. In this paper, we will restrict the consideration to the kernel we found most robust against varying data characteristics:

$$K(z) = \exp(-\gamma \cdot z).$$

The quantity $\gamma \in \mathbb{R}^+$ is the flexible parameter of the LLDA algorithm which should be optimized before its usage.

LLDA is based on the local estimates of the model parameters described above. Applying them in (1) yields a set of localized discriminant functions $\hat{h}_{g_L}$ which can be used to construct the classification rule

$$\hat{A}(\gamma) = \arg\max_g \hat{h}_{g_L}\big(\mathbf{x}\big). \tag{2}$$

As in classical LDA, this approach may cause numerical problems if the considered trial point $\mathbf{x}$ extremely differs from all G class-specific mean vectors since then, the posterior probabilities for all classes are approximately equal to zero. Although this case is very rare, we augmented (2) in order to account for it:

$$\hat{A}(\gamma) = \begin{cases} \arg\max\limits_{g} \hat{h}_{g_L}(\mathbf{x}) & , \exists\, g : \exp\big(-0.5(\mathbf{x} - \bar{\mathbf{x}}_g)' \mathbf{S}_L^{-1} (\mathbf{x} - \bar{\mathbf{x}}_g)\big) > \frac{10^{-150}}{p_{g_L}}, \\ \arg\min\limits_{g} \|\mathbf{x} - \bar{\mathbf{x}}_g\| & , \text{otherwise.} \end{cases}$$

If classifying $\mathbf{x}$ on basis of (2) is rather questionable because of its position in the data space, the simple classification rule ’Choose the class with the closest centroit’ is applied. For programming the LLDA algorithm, we used the software package R (R Development Core Team, 2005)[1].

3 Simulation Study

In this section we use a simulated two-class discrimination problem to investigate the performance of LLDA. In our simulation study, the two classes A_1 and A_2 each consist of two subclasses, namely A_{11}, A_{12}, A_{21} and A_{22}. For the training data set, each class is chosen to contain 1000 two-dimensional observations which are equally divided into the two corresponding subclasses. The data points are generated as normally distributed with the common covariance matrix $\Sigma_{ij} = I_2\ \forall i, j$ and the following subgroup-specific means: $\mu_{11} = (1, 0)'$, $\mu_{12} = (-1, 0)'$, $\mu_{21} = (1.75, 0)'$ and $\mu_{22} = (-1.75, 0)'$. The entire data cloud building up class A_2 is then relocated by a shift defined through the vector $\mathbf{s} = (0.1, 0.3)'$ and rotated by 60°. This results in a data structure such as shown in Figure 1.

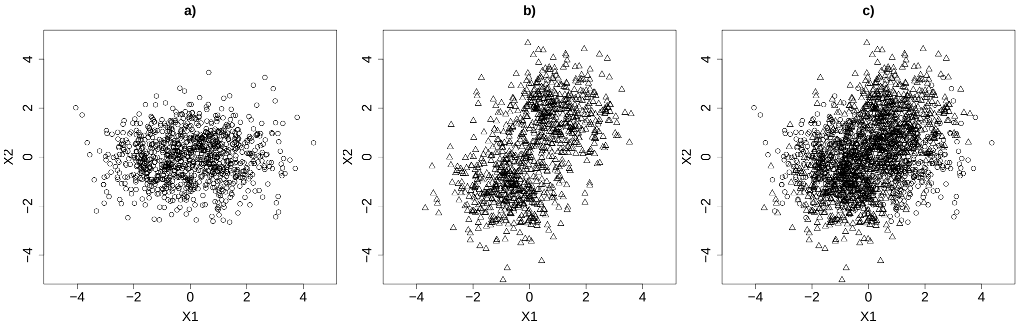

Fig. 1. Simulated Data: a) class A_1, b) class A_2 and c) combined data set

[1] On the used computer, the value 10^{-150} reflects the square root of the machine precision of R for distinguishing a number from zero.

Table 1. Performances of LDA, LLDA and MDA. The error rates are averaged over 10 simulations with the standard error for the average in parenthesis.

Technique	Test Error Rate
LDA	0.4629 (0.0101)
MDA	0.2705 (0.0101)
LLDA	0.2765 (0.0089)

The test data is generated independently from the training data in exactly the same way. It, too, consists of 1000 observations per class. Due to the local structure of the simulated data, LDA is not likely to perform well. We therefore chose MDA (Mixture Discriminant Analysis) as a further competitor for LLDA since this method is particulary designed to cope with local subgroups (Hastie and Tibshirani, 1996). To evaluate the performances of LLDA, the training data is randomly divided into a learning set and a validation set containing 1333 and 667 observations respectively. The optimal values for the flexible parameter γ is then obtained by minimizing the error rate on the validation data. Having done this, the entire training data is used to create a classification rule which is then evaluated on the test data. When using LDA, no optimal parameter values have to be obtained. The classification rule here is learned on basis of the training set and used for the classification of the test observations. The same is done for MDA where the number of subclasses is set to two for both classes. Table 1 contains the results obtained by ten simulations. As expected, all classifiers perform rather bad due to the high degree of overlapping of the two classes. In particular, the global LDA classifier fails to construct a good classification rule whereas MDA and LLDA result in error rates close to the Bayes risk which for the ten simulations on average is given by 0.2747.

4 Application to Business Phase Classification

The field of predicting economic phenomena is a diverse practical example where the data adaptive approach of LLDA is likely to outperform the classical LDA due to the fact that the economic situation develops over time. Assuming the same distribution for all observations of the same class can therefore be too restrictive. In this paper, we address the particular problem of classifying business phases. The data set we consider consists of 13 economic variables with quarterly observations describing the German business cycle from 1955/4 to 2000/4 (Heilemann and Muench, 1996). These variables are standardized and used to classify the business cycle corresponding to a four-phase scheme: upswing, upper turning point, downswing and lower turning point. For such kind of time related data, the key interest often is to find a reliable classification rule for e.g. the next six quarters. In order to evaluate classifiers with respect to this request, Luebke and Weihs (2005) propose the Ex-Post-Ante error rate (EPAER).

4.1 The Ex-Post-Ante Error Rate for Time Related Data

Let the training data $\left\{(\mathbf{x}_t, y_t)'\right\}_{t=1}^{T}$ consist of T successive p-dimensional observations $\mathbf{x}_t$ with a known class membership $y_i \in \{A_1, \ldots, A_G\}$, and let pre denote the number of future time points for which an appropriate classification rule is required. The EPAER at time $t < T$ then has the form

$$epa(t; pre) = \frac{\sum_{i=t}^{\min(t+pre,T)} I_{\{A_i \neq \hat{A}_i^t\}}}{\min(pre, T-t)},$$

where A_i and $\hat{A}_i^t$ are the true and the estimated class for observation i respectively. The quantity $epa(t; pre)$ denotes the error rate of the classification rule which is based on the first t training observations and then applied to the next pre training observations. This approach therefore produces a time series $\left\{epa(t; pre)\right\}_t$ of error rates which can then be condensed to an overall accuracy measure for the classification of the next pre time points:

$$\hat{e}_{t_0,pre} = \sum_{t=t_0}^{T-1} \tilde{w}(t)\, epa(t; pre), \tag{3}$$

where t_0 denotes the starting point for calculating the EPAERs. A suitable weight function for the problem at hand is

$$\tilde{w}(t; t_0, T) = \frac{t}{\sum_{t=t_0}^{T-1} t},$$

which gives more weight to recently calculated error rates.

4.2 Results

In order to obtain a benchmark for the performance of LLDA, we first applied the classical LDA to the data set described above. The dashed line in Figure 2 shows the resulting time series of EPAERs for the prediction interval length of $pre = 6$ quarters and the starting point $t_0 = 20$. The plot reveals the interesting property that historic events which had an impact on the German business cycle can be identified by peaks of the corresponding error rates. For example the reunification of Germany (1990) changed the business cycle so that the next few phases cannot be predicted properly. Also the start of the oil-crisis (oil price increased after 1971) and the second oil crisis (1979) cause problems for LDA.

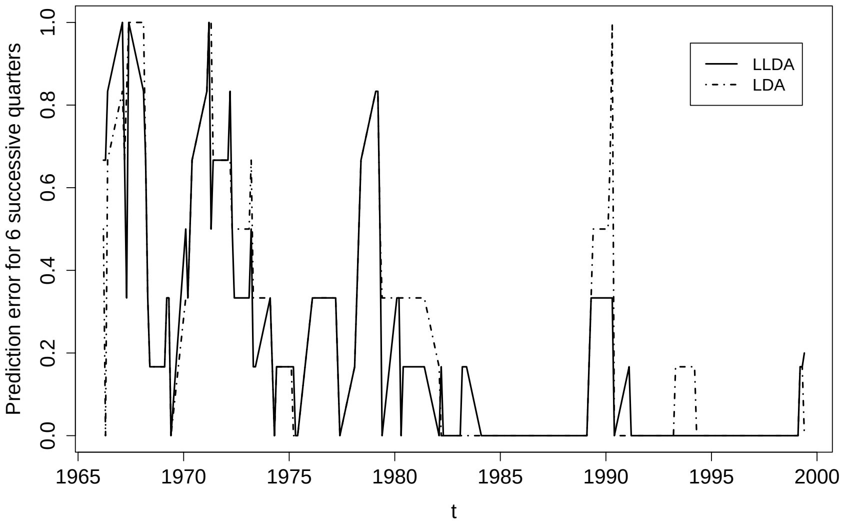

Fig. 2. Time Series of Ex-Post-Ante error rates for LDA and LLDA

Aggregating the time series of error rates corresponding to (3) leads to an estimated overall accuracy for predicting the next six quarters of $\hat{e}_{20,6} = 0.1527$.

As described in Section 2, when using LLDA, the performance of a classification rule is influenced by the value of γ which should therefore be optimized with respect to the chosen accuracy measure. A possible way to accomplish this is minimizing the function $\hat{e}_{t_0,pre} = \hat{e}_{t_0,pre}(\gamma)$ with respect to γ. The optimum found by doing so (setting $t_0 = 20$ and $pre = 6$) yields $\hat{e}_{20,6}(\gamma^{\text{opt}}) = 0.1144$. This result, however, is clearly due to overfitting and therefore gives an overoptimistic estimate of the accuracy. To get a more realistic impression about the benefit of optimizing γ, we applied the stepwise procedure shown in Algorithm 1.

Algorithm 1 Obtaining the EPAERs based on stepwise optimal values of γ.

```
$t = t_0$
**while** t < (T-1) **do**
    select the first $t$ training observations $\left\{(\mathbf{x}_i, y_i)'\right\}_{i=1}^{t}$ as learning data
    find the corresponding optimal value for $\gamma$:
    $\gamma_t^{\text{opt}} = \arg\min_{\gamma} \tilde{w}(t'; [\frac{t}{5}], t) \cdot \sum_{t'=[\frac{t}{5}]}^{t-1} \frac{\sum_{i=t'}^{\min\{t'+pre,t\}} I_{\{A_i \neq \hat{A}_i^{t'}(\gamma)\}}}{\min\{pre,t-t'\}}$
    use $\gamma_t^{\text{opt}}$ to calculate the EPAER for time point $t$:
    $epa(t; pre; \gamma_t^{\text{opt}}) = \frac{\sum_{i=t}^{\min\{t+pre,T\}} I_{\{A_i \neq \hat{A}_i^{t}(\gamma_t^{\text{opt}})\}}}{\min\{pre,T-t\}}$
    $t \leftarrow t + 1$
**end while**
```

Since here in each case, the value γ_t^{opt} is obtained from data points prior to t and used for predicting the class membership of upcoming observations (which mimics a real-life situation), the resulting time series of EPAERs $\{epa(t; pre, \gamma_t^{\text{opt}})\}_t$ does not suffer from overfitting. It is shown as the solid line of Figure 2. Since its shape roughly resembles the one obtained by LDA, it, too, can be explained historically. Corresponding to (3), an overall measure for the goodness of LLDA for predicting the next six quarters is given by $\hat{e}_{20,6}^{\text{opt}} = 0.1277$. Compared to the classical LDA, the observation specific approach of LLDA utilizing an optimal value of γ therefore leads to an improvement of 16.37% in terms of the Ex-Post-Ante-Error rate.

For completeness sake we also applied MDA to the problem at hand. Since an optimization of the flexible MDA parameters, namely the number of local subclasses for each class, would be very time intensive, we assumed the number of subclasses to be equal to a constant s for all four classes. Optimizing s with respect to the EPAER resulted in $s^{\text{opt}} = 1$, i.e. the classical LDA approach. The advantage of LLDA compared to LDA is therefore not due to the existence of local subclasses which shows that LLDA is beneficial in a diverse range of local settings.

5 Summary

By introducing locally adaptive weights to the global LDA technique, the new observation specific classifier LLDA has been developed. Its benefit could be evaluated on basis of a simulated data set containing local subclasses and in the real life application of classifying business phases. LLDA outperforms LDA in both cases as well as MDA in the business classification problem. In the simulation study, it yields similar results as MDA which is noteworthy since MDA is the method particulary developed for such data structure.

Acknowledgment

This work has been supported by the Collaborative Research Center 475 'Reduction of Complexity for Multivariate Data Structures' of the German Research Foundation (DFG).

References

FISHER, R.A. (1936): The use of multiple measurements in taxomonic problems. *Annals of Eugenics*, 7, 179–188.

HAND, D.J. and VINCIOTTI, V. (2003): Local versus global models for classification problems: Fitting models where it matters. *The American Statistician*, 57(2), 124–131.

HASTIE, T. and TIBSHIRANI, R. (1996): Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58(1), 155–176.

HEILEMANN, U. and MUENCH, J.M. (1996): West german business cycles 1963–1994: A multivariate discriminant analysis. In: *CIRET-Conference in Signapoure*, CIRET-Studien 50.

LUEBKE, K. and WEIHS, C. (2005): Prediction Optimal Classification of Business Phases. *Technical Report 41/05, SFB 475, Universität Dortmund*.

R DEVELOPMENT CORE TEAM (2005): *R: A language and environment for statistical computing.* R Foundation for Statistcal Computing, Vienna.

TUTZ, G. and BINDER, H. (2005): Localized classification. *Statistics and Computing*, 15, 155–166.