Kernel discrimination and explicative features: an operative approach

Caterina Liberati, *Università degli Studi di Milano-Bicocca*, caterina.liberati@unimib.it
Furio Camillo, *Università di Bologna of Bologna*, furio.camillo@unibo.it
Gilbert Saporta, *Conservatoire National des Arts et Mètiers*, gilbert.saporta@cnam.fr

Abstract. Kernel-based methods such as SVMs and LS-SVMs have been successfully used for solving various supervised classification and pattern recognition problems in machine learning. Unfortunately, they are heavily dependent on the choice of the optimal kernel function and from tuning parameters. Their solutions, in fact, suffer of complete lack of interpretation in terms of input variables. That is not a banal problem, especially when the learning task is related with a critical asset of a business, like credit scoring, where deriving a classification rule has to respect an international regulation. The following strategy is proposed for solving problems using categorical predictors: replace the predictors by components issued from MCA, choice of the best kernel among several ones (linear ,RBF, Laplace, Cauchy, etc.), approximation of the classifier through a linear model. The loss of performance due to such approximation is balanced by better interpretability for the end user, employed in order to understand and to rank the influence of each category of the variables set in the prediction. This strategy has been applied to real risk-credit data of small enterprises. Cauchy kernel was found the best and leads to a score much more efficient than classical ones, even after approximation.

1 Introduction

Kernel methods have proved their efficiency to solve various problems of supervised classification and pattern recognition and are now a basic tool of machine learning. SVMs are well known and proved their superiority in many fields of applications : pattern recognition, industry and business including credit scoring [1, 22].

There are unfortunately some drawbacks which prevent from a general use: besides some technical difficulties (choice of the kernel, tuning of hyperparameters) , they appear like black-box methods without any direct interpretation in terms of input variables. This lack of comprehensibility is a major drawback and causes a reluctance to use the model. It goes even further: when credit has been denied to a customer, the Equal Credit Opportunity Act of the U.S. requires

that the financial institution provides specific reasons why the application was rejected; indefinite and vague reasons for denial are illegal. According to that, our goal in this study was not to derive a new classification model that discriminates better than others previously published, or illustrate a system that best estimate the performance of existing algorithms, but to employ, indirectly, the good Kernel discriminant classifications into an operative field. Therefore, we propose a simple but innovative way to improve the understanding of the classification function; we fit a linear model to the classifier using the input variables as predictors: linear regression if all predictors are numerical, or a general linear model if some or all predictors are categorical. The loss of performance due to such approximation is balanced by a better interpretability for the end user who may rank the influence of each category of the variables set in the prediction.

Instead of standard SVM, we use here LS-SVM (least-squares support vector machines). LS-SVM [17] boils down to FisherÕs linear discriminant function in the feature space which is a simple way to extend discriminant analysis to the nonlinear case [2] .

Since most practical problems are concerned with categorical predictors, we propose a kernelized version of Disqual, a technique developed by Saporta, 1977 with the following strategy: replace the categorical predictors by some components issued from MCA, then perform a LS-SVM and finally approximate the classifier through a linear model.

This strategy has been applied to real risk-credit data of small enterprises. We compared several kernels (linear , RBF, Laplace, Cauchy, etc.): Cauchy kernel was found the best and leads to a score much more efficient than classical ones, even after approximation.

2 Kernel discriminant analysis

The kernel machines provide an elegant way of designing nonlinear algorithms by reducing them to linear ones in a high-dimensional Feature Space $\mathcal{F}$ nonlinearly related to the Input sample space $\mathcal{X}$:

$$\Phi : \mathcal{X} \rightarrow \mathcal{F} \tag{1}$$

Naturally, $\mathcal{F}$ dimensionality could be arbitrarily large, possibly infinite, and that could be very complex task to be done. Fortunately, the exact $\phi(z)$ is not needed and the Feature Space can become implicit by using a positive definite kernel satisfying the Mercer's condition (Mercer, 1909):

$$K(z, x) =< \Phi(z), \Phi(x) > \tag{2}$$

The trick behind the methods is to replace dot products in $\mathcal{F}$ with a kernel function in the Input space so that the nonlinear mapping is performed implicitly in the new Space [19, 20]. Once mapped the data, a Fisher Linear Discriminant Analysis (FLDA), the most popular supervised classification technique, can be performed.

Assume that we are given the input data set $\mathcal{I}_{XY} = \{(x_1, y_1), ..., (x_n, y_n)\}$ of training vectors $\mathbf{x}_i \in \mathcal{X}$ and the corresponding values of $y_i \in \mathcal{Y} =\{1, 2\}$ be sets of indices of training vectors belonging to the first $y = 1$ and the second $y = 2$ class, respectively. The class separability in a direction of the weights $\alpha= [\alpha_1, \ldots, \alpha_n]'$ is obtained maximizing the Rayleigh coefficient:

$$J = \frac{\alpha' S_B^{\Phi} \alpha}{\alpha' S_W^{\Phi} \alpha}, \tag{3}$$

where S_B^Φ, S_W^Φ are respectively the Between and Within covariance matrices in the Feature Space:

$$\begin{aligned} S_B^\Phi &= (m_1^\Phi - m_2^\Phi)(m_1^\Phi - m_2^\Phi)' \\ &= (\overline{\kappa}_1 - \overline{\kappa}_2)(\overline{\kappa}_1 - \overline{\kappa}_2')', \end{aligned} \tag{4}$$

$$\begin{aligned} S_W^\Phi &= \sum_{i=1,2} \sum_{x \in X_i} (\Phi(x) - m_i^\Phi)(\Phi(x) - m_i^\Phi)' \\ &= KK' - \sum_{k=1}^{2} n_k \overline{\kappa}_k \overline{\kappa}_k' \end{aligned} \tag{5}$$

with

$$K = [\kappa(x_i, x_j)]_{(n \times n)}, \text{and}$$

$$\overline{\kappa}_k = \frac{1}{n_k} \sum_{j \in I_k} K_j,$$

where K_j is the j-th column of K and I_k the index set of group k. This problem can be solved by finding the leading eigenvectors of $(S_W^\Phi)^{-1} S_B^\Phi$. Therefore the kernel discriminant function $f(x)$ of the binary classifier can be written as

$$f(x) = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) + b \tag{6}$$

where b is the intercept (or the bias) of the discriminant hyperplane (6) is determined by setting the hyperplane to pass through the mid point of the two group means

$$b = -\alpha' \frac{(\overline{\kappa}_1 + \overline{\kappa}_2)}{2}. \tag{7}$$

Since the matrix S_W^Φ is at most of rank $n - 1$ the proposed setting is ill-posed, therefore a regularization method to overcome the singularity and instability has to be applied ([8], [14] [21]).

KDA settings

Optimal generalization of kernel-based method including KDA, still depends on the selection of a suitable kernel function and the values of regularization and kernel parameters ([4]). In literature, many kernel functions are present we can choose from. The most common are shown in the Table 1 below: where $c \in \mathrm{R}$ is the width can be reviewed as a variance indicator of the data. Except for the Polynomial kernel which presents, as unknown parameter, the degree of the functions that can be fixed by the user just choosing the degree of the transformation, with the other maps the estimation process has been addressed towards suited choices according to the data. Grid search or its variant is the procedure employed in those works when the parameters are not fixed a priori, but these types of procedures are too time consuming. In this paper we used the parameter selection first introduced in [12], by which the we learn directly from

Kernel Mapping	k(x,z)
Cauchy	$\frac{1}{1+\frac{\|\|\mathbf{x}-\mathbf{z}\|\|^2}{c}}$
Laplace	$\exp(-\sqrt{\frac{\|x-z\|^2}{c^2}})$
Multi-quadric	$\sqrt{\|\|\mathbf{x}-\mathbf{z}\|\|^2+c^2}$
Polynomial degree 2	$(x \cdot z)^2$
Gaussian (RBF)	$\exp(\frac{-\|x-z\|^2}{2c^2})$
Linear	$(x \cdot z)$

Table 1: Kernel Functions

the data the right value of c. We minimized the tuning parameter search space applying the *Jackknife Mahalanobis distance Data Depth* defined by [3] :

$$JMDD_G = \frac{1}{1+\sqrt{((x_i-\overline{x})S_i^{-1}(x_i-\overline{x}))}} \tag{8}$$

where S_i^{-1} is the inverse of the sample covariance matrix computed from the data set deleting the ith observation x_i.

$$S_i^{-1} = (n-2)[\frac{1}{n-1}S^{-1} + \frac{\gamma}{(n-1)}S^{-1}(x_i-\overline{x})(x_i-\overline{x})^T S^{-1}] \tag{9}$$

and

$$\gamma = \frac{n}{n-1}[1-(\frac{n}{(n-1)^2})^2(x_i-\overline{x})S^{-1}(x_i-\overline{x})] \tag{10}$$

As we highlighted in the section 2, matrix S_W^Φ is at most of rank $n-1$ the proposed setting is ill-posed. As such, regularization methods to overcome the singularity and instability are widely applied in the statistical domain. Instead of employing a ridge to the pooled covariance S_W^Φ, as usual in the literature, we employed a two stage stabilization and smoothing process that provides a well-conditioned covariance matrix that is both nonsingular and positive definite first applied in [12]. First it stabilizes the eigenvectors of the matrix S_W^Φ via its mean [18], then we smooth the stabilized covariance matrix using the Convex Sum Covariance estimator (CSE) ([12]) which is given by:

$$\hat{\Sigma}_{CSE} = \frac{n}{n+m}\hat{\Sigma} + (1-\frac{n}{n+m})\hat{D}_W, \tag{11}$$

where n is the sample size and $\hat{D}_W = (\frac{1}{p}tr\hat{\Sigma})\mathbf{I}_p$ with p number of variables. For $p \geq 2$, m is chosen to be

$$0 < m < \frac{2[p(1+\beta)-2]}{p-\beta}, \tag{12}$$

where

$$\beta = \frac{(tr\hat{\Sigma})^2}{tr(\hat{\Sigma}^2)}. \tag{13}$$

This estimator improves upon the $\hat{\Sigma}$ by shrinking all the estimated eigenvalues of $\hat{\Sigma}$ toward their common mean. Moreover, rather than just pick a ridge parameter, we let the data pick it for

us.

Our intelligent criterion to select for selecting the appropriate kernel function is Information Complexity index (ICOMP) [3, 12] whose MANOVA formulation is given by:

$$\begin{aligned} ICOMP(IFIM) &= -2\log L(\hat{\theta}) + 2C_1(F^{-1}(\hat{\theta})) && (14)\\ &= np\log(2\pi) + n\log|S_W^{\Phi}| + np + 2C_1(F^{-1}(\hat{\theta})) && (15)\end{aligned}$$

$\hat{S}_W$ is the estimated within-cluster covariance matrix, C_1 denotes the maximal information complexity, and $F^{-1}(\hat{\theta})$ is the estimated Inverse Fisher Information Matrix (IFIM) of the model.

3 K-disqual

When predictors are categorical, there is no simple metric to compute distances or similarities between units. A usual solution is to perform in a first step a Multiple Correspondence Analysis. Let us suppose that we have p predictors with m_j categories, $j = 1, .., p$. MCA provides $q = \sum_{j=1}^{p} m_j - p$ components which embeds units in an euclidean space of dimension q . Disqual (Discriminant analysis for qualitative variables) proposed by [16] is a kind a principal component regression where the response variable Y is binary, and the explanatory variables are components obtained by MCA instead of PCA components.

After a simple process of selecting r components among q by eg. cross-validation, Disqual computes the Fisher linear discriminant function as outlined as follows.

Let z^j be the MCA components and λ_j their variances (eigenvalues) . Since MCA components are orthogonal, it is straightforward to inverse the total covariance matrix V instead of the within covariance matrix S_W (it is well-known that solutions of the FisherÕs LDA with V^{-1} or S_W^{-1} are proportional). Fisher's score is

$$s = \sum_{j=1}^{r} u_j z^j \text{ where } u = \begin{pmatrix} . \\ . \\ u_j \\ . \\ . \end{pmatrix} = V^{-1}(g_1 - g_2) = \begin{pmatrix} . \\ . \\ \frac{\bar{z}_1^j - \bar{z}_2^j}{\lambda_j} \\ . \\ . \end{pmatrix} \quad (16)$$

and $\bar{z}_1^j$ $\bar{z}_2^j$ are the means of group 1 and group 2 for the j-th axis. Transition formulas say that $z^j = Xa^j$ where a^j is the vector of coordinates of the m_1+É.+m_p categories along j-th axis, hence:

$$s = \sum_{j=1}^{r} u_j X a^j = X \underbrace{\sum_{j=1}^{r} u_j a^j}_{scorecard} \quad (17)$$

Score s is a sum of partial scores (scorecard) corresponding to each predictor, which is easy to interpret.

In order to obtain a kernelized version of Disqual "K-Disqual" , it suffices to project the data (z variables) in the feature space generated by some kernel K and performs a LS-SVM, which is nothing but a Fisher LDA in the feature space. But there is no simple relationship with the input variables, which prevents direct interpretation.

4 Getting back to input space from kernel space

Kernel-based methods, try to devise algorithms that solve complex tasks by learning a solution rather than by engineering it. The success of this approach is to build machines that are able to learning the relationship among objects without needing a lot of expert knowledge built-in. Moreover its mathematical framework is so flexible to convert non linear problem into a linear one by mapping data onto the Feature Space $\mathcal{F}$. This latter is defined as the space of all functions mapping from $\mathcal{X} \to \mathbb{R}$, i.e. $\mathbb{R}^{\mathcal{X}} = \mathcal{F} = \{f | f : \mathcal{X} \to \mathbb{R}\}$ then we can define:

$$\begin{aligned} \phi : \mathcal{X} &\to \mathbb{R}^{\mathcal{X}} \\ x &\to k(\cdot, x) \end{aligned} \tag{18}$$

where $\phi(x)$ denotes the function that assigns the value $k(x', x)$ to $x' \in \mathcal{X}$, i.e., $\phi(x)(\cdot) = k(\cdot, x)$. Therefore embedding the data into $\mathcal{F}$ has two advantages:

- It allows us to deal with patterns now represented by their similarity to all other points in the input domain, which is a very rich representation
- it gives us freedom to choose the mapping Φ that enable us to design a large variety of learning algorithms.

Although such setting shows a ease of use and theoretical appeal, it presents some drawbacks from an applied statistical perspective. The data transformation obtained via kernel trick changes semantically the meaning of the information employed to build the classification rule: this is not a crucial issue when model fit or misclassification error rate is the only objective of our analysis, but such issue could be very important when the assessment of the model has to take into account also of its explicative power. All the explorative techniques (as ACP or Clustering) need to be feed with variables easy to read and characterized by known metrics, same case if we estimate a predictive rule for economic or business applications. Kernel machines suffer of a completely lack of interpretation because of the dot product operation which represents its major innovation. Dot product of input instances generates implicitly the mapping of sample vectors into the Feature Space in which, however, it not possible to distinguish between objects (rows) and variables (columns) of the Input Space. In addiction the dual Space obtained does not allow to get back to the original one (for the properties of the dot product) therefore seems that good performance of kernel-based methods and interpretation of results can not be joined. In this paper we try to reach this ambitious aim via a two step strategy:

1. selection of the best kernel discriminant solution which becomes our benchmark result
2. approximation of the kernel rule via input variables in order to get decision function where the weight of every single variable's category can be derived.

Of course the step 2. might involve a lost of efficiency in classification or prediction because the reconstruction phase requires to fit a new model. Trade off between the best prediction and the easy of interpretation and application will lead to identification of the new model.

5 Credit scoring and kernel methods

Modern data-mining techniques have been employed to construct credit scoring models. Linear discriminant [15], logistic regression [9], decision tree [6] and neural networks [13] have been wild and successfully applied in order to increase the accuracy of the classification. Recently Kernel-based techniques have been employed in dealing with credit scoring problems and their performances have been much more superior to that of traditional approaches, especially in non-linear pattern classification [10, 7].
Although, every little improvement in misclassification rate causes noteworthy bank cost savings (kernel discriminants or SVM, therefore, are giving significant contributions in lending decision), such models focus only on good prediction and are completely missing of rule interpretation. Moreover they do not deal directly with mixed data (qualitative and quantitative customer attributes): often in such studies socio demographic variables were collected together with indexes which approximate the consumer behaviors in order to get a complete information set. This empirical evidence generates a lack of application, because a credit scoring of a bank, when implemented, has to return set of classification rules which have to be easily interpretable in terms of the original input variables. This makes Kernel-solutions difficult to use as is required in official documentations of European bank system, to a rejected applicant has to be given a reason for being rejected. Therefore as [5] underlined in their work, a possible solution for this issue is the use of SVM rule extraction techniques or the use of hybrid-SVM model combining with other more interpretable models.

6 Application to an Italian bank data

The application presented in this work is related to a real problem of credit risk for small businesses of a major Italian bank. A real dataset which collects information on over 85000 firms, has been analyzed in order to provide a guideline on approach. Different data mining models have been considered in order to predict the default probability of a new potential client (a firm as regards the practical part) asking for a credit to the bank. According to the past qualitative and quantitative information, an estimation of the predicted default probability is computed. Based on that, clients are classified in one of the two groups, namely good or bad clients, and they receive credit or not. Indeed, following the Basel procedures it is more correct to refer to the default probability rather than bad or good clients. Thus, for the purpose of this research, we classify as "good" clients those that are predicted not in default over the next twelve months after the credit receiving, while the "bad" clients are those predicted in default. In order to deal to the bias of the defaults distribution, which is very overbalanced in favor of "good" clients, we oversampled the bad instances according to the classical approach of the "rare event" of "default situation", therefore, the final distribution between "bad" (y=1) and "good" (y=2) clients, was 29% and 71%, of a very large sample composed by 15000 units.
According to the k-Disqual method, a multiple correspondence analysis was performed in order to obtain factor coordinates that have been employed as input variable in the predictive methods. Going into more details, for as regards lending credit to businesses, during the lending application process, a customerÔs creditworthiness is assessed on the basis of the analysis of the following elements: a) operating, financial and cash flow data; b) qualitative information regarding the companyÔs competitive position, its corporate and organizational structure (only for business customers in the Corporate area); c) geographical and sector characteristics (only for business

customers in the Corporate area); performance data at bank and industry levels (e.g., the Central Risk Bureau). All those elements are considered when assigning a rating, meaning the borrowerÔs default probability that is calculated over a time horizon of one year. Our database is composed by 10 qualitative variables: 4 coming from information collected by a questionnaire administered to corporate customers, 2 representing the behavior of bank customers, and finally 4 variables related to measures of solvency derived from some external measures of risk assessment by the Central Risk Bureau. Via k-Disqual approach, a categorical segmentation of each original variable was realized and, submitting the original data matrix to a Multiple Correspondence Analysis (MCA), orthogonal continuous axes were obtained.
The outline of steps of the overall strategy can be summarized in the as follows:

1. transformation of original variables in factorial coordinates via Multiple Correspondence Analysis (MCA)

2. application of the K-Disqual approach using the Kernel Mapping in table 1 (from Cauchy to Linear)

3. choice of kernel transformation by the minimum misclassification error rate; the choice was made by analyzing the characteristics of equilibrium and balance of predictive content in the confusion matrix, with a starting point: the correct classification average rate.

4. study of the selected kernel discriminant function using the original categorical variables and each of their category

5. rebuilding the kernel discriminant function by including as exogenous variables, the original categorical variables. Study of the goodness of fit of the model reconstruction.

Table 2 shows the 5 confusion matrices (1=bad client, 2=good client) related to the 4 discriminant functions used (3 kernel transformations and the linear functions were used) and a classical logistic regression which represents our benchmark solution. It is evident that the Cauchy solution can be considered the “best performance” on our data.

cauchy	Percent			
	Classified into y			
Avg-correct	From y	1	2	Total
87.87	1	88.34	11.66	100
	2	12.31	87.69	100
	Total	33.44	66.56	100

linear	Percent			
	Classified into y			
Avg-correct	From y	1	2	Total
63.42	1	62.71	37.29	100
	2	36.31	63.69	100
	Total	43.65	56.35	100

laplace	Percent			
	Classified into y			
Avg-correct	From y	1	2	Total
81.63	1	82.58	17.42	100
	2	18.74	81.26	100
	Total	36.48	63.52	100

rbf	Percent			
	Classified into y			
Avg-correct	From y	1	2	Total
78.93	1	78.40	21.6	100
	2	20.87	79.13	100
	Total	36.86	63.14	100

logistic regression	Percent			
	Classified into y			
Avg-correct	From y	1	2	Total
74.06	1	52.12	47.88	100
	2	17.49	82.51	100
	Total	27.12	72.88	100

Table 2: Confusion matrices of different KDA and a logistic regression

Observing the 2-class distributions plots, it is evident how the discriminant function defined by the Cauchy-kernel transformation is more effective in the separation of the two target groups (Fig. 1).

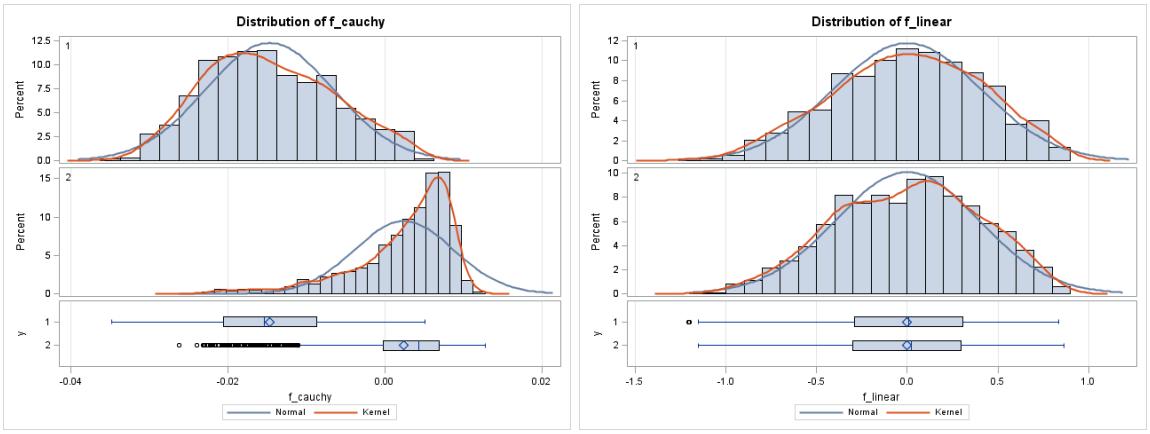

Figure 1: 2-classes distributions

As mentioned in section 5, the use of predictive models in credit scoring applications must be transparent with respect to the type of model adopted and the weights (parameters) that have the covariates in the predictive mechanism of each applicant customer [30]. One of the problems of operative applications of the KDA is thus right to not provide a solution directly interpretable by original exogenous variables. Therefore in our approach, the descriptive step of results obtained in calculating of kernel discriminant function can be performed by a particular characterization analysis of a quantitative variable (our kernel discriminant function) using categories of the same set of variables that gave rise to the original kernel transformation [11]. Table 3 shows the result of a F-test of significance for each of 10 categorical variables with which it has attempted a reconstruction of Cauchy discriminant function. It is evident that the 4 variables related to opinions collected through a survey (DOM1-DOM4) explain much less than two indicators of risk arising from the Central Bureau (CB1_AZ-CB2_AZ). These indicators, in a similar fitting process for the linear discriminant function, are much less important. In the definition of linear discriminant function variables arising from the opinion survey research are the most important.

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
DOM1	3	0.0002	6.115E-05	2.09	0.0997
DOM2	3	0.0018	0.0006	20.79	<.0001
DOM4	3	0.0077	0.0026	87.30	<.0001
DOM3	3	0.0084	0.0028	96.07	<.0001
ANAG	3	0.0029	0.0010	33.53	<.0001
CB1_AZ	4	0.0175	0.0044	149.43	<.0001
CB1_COLL	4	0.0024	0.0006	20.09	<.0001
CB2_AZ	4	0.0750	0.0188	640.35	<.0001
CB2_COLL	4	0.0049	0.0012	41.72	<.0001
CERI	4	0.0165	0.0041	141.15	<.0001

Table 3: F test and p-values

[30] For that reason, we choose to not include in our model no interaction among original variables, in order to get results more easy to be interpreted.

Table 4 illustrates the parameters of the linear fitting of Cauchy kernel discriminant function using the set of available covariates. The set of parameters for each variable can be clearly interpreted as a set of contrasts

Parameter Estimates

Variable	DOM1	DOM2	ANAG	CB1_AZ	CB1_COLL	CB2_AZ	CB2_COLL	CERI	DOM4	DOM3	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept											1	0.0062	0.0005	11.47	<.0001	.	0
DOM1	1										1	0.0003	0.0004	0.84	0.4002	0.2799	3.5730
	2										1	0.0002	0.0003	0.62	0.5367	0.3693	2.7076
	3										1	0.0006	0.0002	2.37	0.0176	0.5570	1.7954
	4										0	0	.	.	.	.	.
DOM2		1									1	-0.0023	0.0006	-3.78	0.0002	0.7660	1.3056
		2									1	-0.0026	0.0003	-7.46	<.0001	0.4523	2.2107
		3									1	-0.0004	0.0003	-1.39	0.1646	0.5434	1.8403
		4									0	0	.	.	.	.	.
DOM4									1		1	-0.0035	0.0003	-13.05	<.0001	0.4215	2.3727
									2		1	-0.0020	0.0003	-6.54	<.0001	0.5625	1.7777
									3		1	-0.0005	0.0003	-1.71	0.0880	0.4890	2.0448
									4		0	0	.	.	.	.	.
DOM3										1	1	-0.0060	0.0005	-11.62	<.0001	0.4101	2.4383
										2	1	-0.0002	0.0004	-0.48	0.6310	0.1650	6.0601
										3	1	-0.0006	0.0004	-1.45	0.1470	0.1803	5.5477
										4	0	0	.	.	.	.	.
ANAG			1								1	-0.0030	0.0003	-9.97	<.0001	0.3651	2.7387
			2								1	-0.0022	0.0003	-7.77	<.0001	0.4387	2.2794
			3								1	-0.0018	0.0003	-6.45	<.0001	0.5158	1.9387
			4								0	0	.	.	.	.	.
CB1_AZ				1							1	-0.0051	0.0003	-17.79	<.0001	0.4576	2.1853
				2							1	-0.0008	0.0003	-2.77	0.0056	0.5242	1.9078
				3							1	0.0008	0.0003	2.48	0.0130	0.5867	1.7045
				4							1	0.0007	0.0003	2.55	0.0108	0.6021	1.6610
				5							0	0	.	.	.	.	.
CB1_COLL					1						1	-0.0032	0.0006	-5.53	<.0001	0.2624	3.8114
					2						1	-0.0001	0.0006	-0.26	0.7973	0.3773	2.6504
					3						1	-0.0004	0.0006	-0.78	0.4329	0.3945	2.5350
					4						1	0.0010	0.0006	1.75	0.0808	0.4042	2.4738
					5						0	0	.	.	.	.	.
CB2_AZ						1					1	-0.0094	0.0003	-33.04	<.0001	0.4166	2.4001
						2					1	-7.068E-06	0.0003	-0.02	0.9807	0.5415	1.8467
						3					1	0.0025	0.0003	8.34	<.0001	0.5534	1.8070
						4					1	0.0028	0.0003	9.13	<.0001	0.5432	1.8408
						5					0	0	.	.	.	.	.
CB2_COLL							1				1	-0.0030	0.0006	-5.16	<.0001	0.2361	4.2351
							2				1	0.0023	0.0006	3.94	<.0001	0.3314	3.0179
							3				1	0.0003	0.0006	0.52	0.6003	0.3755	2.6631
							4				1	0.0018	0.0006	3.21	0.0013	0.4288	2.3320
							5				0	0	.	.	.	.	.
CERI								1			1	-0.0057	0.0003	-22.53	<.0001	0.7664	1.3049
								2			1	-0.0006	0.0002	-2.64	0.0084	0.8511	1.1750
								3			1	-0.0011	0.0003	-3.16	0.0016	0.9104	1.0984
								4			1	0.0007	0.0003	2.33	0.0198	0.8777	1.1393
								5			0	0	.	.	.	.	.

Table 4: Estimates of Linear Reconstruction Discriminant

Figure 2 shows the relationship between the real kernel discriminant function and the function linearly fitted using the available covariates.

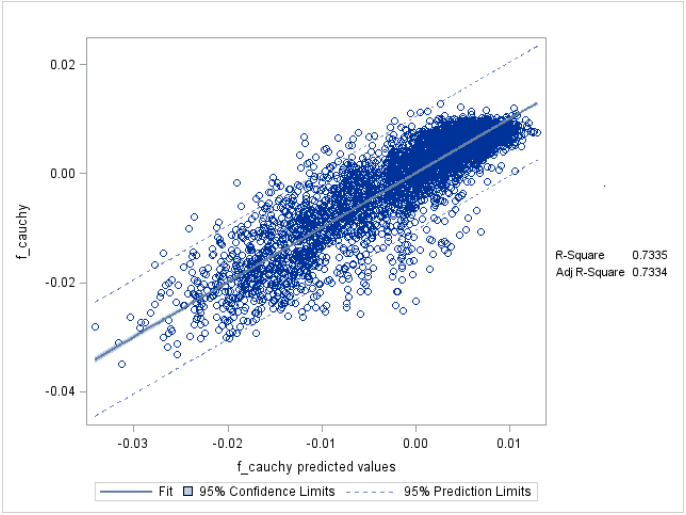

Figure 2: Cauchy Kernel Discriminant vs Linear reconstruction Discriminant

Final result in terms of confusion matrix is very satisfactory, both in the training-set and in a sample of observations treated as a test-set(Tab. 5).

p_cauchy training	Percent Classified into y			
Avg-correct	From y	1	2	Total
78.55	1	80.06	19.94	100
	2	22.03	77.97	100
	Total	38.16	61.84	100

p_cauchy test	Percent Classified into y			
Avg-correct	From y	1	2	Total
77.04	1	79.81	20.19	100
	2	24.03	75.97	100
	Total	40.10	59.90	100

Table 5: Cauchy KDA vs Linear Reconstruction Discriminant confusion matrices

7 Conclusion

Linearizing a kernel classifier allows to easily interpret SVM in the Input Space. Experiments have shown that the loss of precision is not that large, and provides much better results than a direct linear technique. A kernelized version of Disqual has been developed to deal with categorical predictors. Such an approach can be applied to any algorithm translatable in a function of score or in a probability function.

Bibliography

[1] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54:627 – 635, 2003.

[2] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.

[3] H. Bozdogan. Exploring multivariate modality by unsupervised mixture of cubic b-splines in 1-d using model selection criteria. In *Data Analysis: Scientific Model and Pratical Application*, pages 105–119. Springer, 2000.

[4] G.C. Cawley and N. L. C. Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592, November 2003.

[5] H. Cheng-Lung, C. Mu-Chen, and W. Chieh-Jen. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33:847–856, 2007.

[6] R. H. Davis, D. B. Edelman, and A. J. Gammerman. Machine learning algorithms for credit-card applications. *Journal of Mathematics Applied in Business and Industry*, 4:43–51, 1992.

[7] C. Fei-Long and L. Feng-Chia. Combination of feature selection approaches with svm in credit scoring. *Expert Systems with Applications*, 2010.

[8] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[9] W. E. Henley and D. J. Hand. A k-nearest neighbor classifier for assessing consumer credit risk. *Statistician*, 44(1):77–95, 1996.

[10] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558, 2004.

[11] L. Lebart, A. Morineau, and K. Warwick. *Multivariate Descriptive Statistical Analysis.* Wiley and sons, 2004.

[12] C. Liberati, A. Howe, and H. Bozdogan. Data adaptive simultaneous parameter and kernel selection in kernel discriminant analysis (kda) using information complexity. *Journal of Pattern Recognition Research*, 4(1):119–132, 2009.

[13] R. Malhotra and D. K. Malhotra. Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research*, 136(1):190–211, 2002.

[14] S. Mika, J. Weston, B. Schölkopf, and K.R. Müller. Fisher discriminant analysis with kernels. In In Y.-H. Hu, E. Wilson J. Larsen, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, 1999.

[15] A. K. Reichert, C. C. Cho, and G. M. Wagner. An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business and Economic Statistics*, 1(2):101–114, 1983.

[16] G. Saporta. Une méthode et un programme d'analyse discriminante sur variables qualitatives. In E. Diday, editor, *Analyse des Données et Informatique*, pages 201–210. INRIA, 1977.

[17] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines.* World Scientific Pub. Co., Singapore, 2002.

[18] C.E. Thomaz, J.P. Boardman, D.L.G. Hill, J.V. Hajnal, D.D. Edwards, M.A. Rutherford, D.F. Gillies, and D. Rueckert. Using a maximum uncertainty lda-based approach to classify and analyse mr brain images. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2004*, pages 291–300. Springer Berlin / Heidelberg, 2004.

[19] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York., 1995.

[20] V. Vapnik. *Statistical Learning Theory.* Wiley, New York., 1998.

[21] J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu. Efficient model selection for regularized linear discriminant analysis. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 532–539, New York, NY, USA, 2006. ACM.

[22] L. Yu, S. Wang, K.K. Lai, and L. Zhou. *Bio-Inspired Credit Risk Analysis, Computational Intelligence with Support Vector Machines.* Springer, 2008.