Solving and Interpreting Binary Classification Problems in Marketing with SVMs

Georgi Nalbantov[1], Jan C. Bioch[2], and Patrick J. F. Groenen[2]

[1] Erasmus Research Institute of Management,
Erasmus University Rotterdam, Postbus 1738, 3000 DR Rotterdam, The Netherlands

[2] Econometric Institute, Faculty of Economics,
Erasmus University Rotterdam, Postbus 1738, 3000 DR Rotterdam, The Netherlands

Econometric Institute Report E.I. 2005-46

Abstract. Marketing problems often involve binary classification of customers into “buyers” versus “non-buyers” or “prefers brand A” versus “prefers brand B”. These cases require binary classification models such as logistic regression, linear, and quadratic discriminant analysis. A promising recent technique for the binary classification problem is the Support Vector Machine (Vapnik (1995)), which has achieved outstanding results in areas ranging from Bioinformatics to Finance. In this paper, we compare the performance of the Support Vector Machine against standard binary classification techniques on a marketing data set and elaborate on the interpretation of the obtained results.

1 Introduction

In marketing, quite often the variable of interest is dichotomous in nature. For example, a customer either buys or does not buy a product, visits or does not visit a certain shop. Some researchers and practitioners often approach such binary classification problems with traditional parametric statistical techniques, such as discriminant analysis and logistic regression (Lattin et al. (2003), Franses and Paap (2001)) and others employ semiparametric and nonparametric statistical tools, like kernel regression (Van Heerde et al. (2001), Abe (1991, 1995)) and neural networks (West (1997)). Nonparametric models differ from parametric in that they make no or less assumptions about the distribution of the data. A disadvantage of nonparametric tools in general is that they are considered to be “black boxes”. In many such cases, the model parameters are hard to interpret and often no direct probability estimates are available for the output binary variable. A discussion on the relative merits of both kind of techniques can be found, for instance, in Van Heerde et al. (2001) and West (1997).

In this paper, we employ the nonparametric technique of Support Vector Machine (SVM) (Vapnik (1995), Burges (1998), Müller et al. (2001)).

Some desirable features of SVM that are relevant for marketing include good generalization ability, robustness of the results, and avoidance of overfitting. One drawback of SVM is the inability to interpret the obtained results easily. In marketing, SVMs have been used by, for example, Bennett (1999), Cui (2003), and Evgeniou (2004).

Our aim is to assess the applicability of SVM for solving binary marketing problems and, even more importantly, to provide for the interpretation of the results. We compare SVM with standard marketing modelling tools of linear and quadratic discriminant analysis and the logit choice model on one empirical data set. In addition, we interpret the results of the SVM models in two ways. First, we report probability estimates for the realizations of the (binary) dependent variable, as proposed by Platt (1999) and implemented by Chang and Lin (2004). Second, we use these estimates to evaluate the (possibly nonlinear) effects of some independent variables on the dependent variable of interest. In this way, we can assess the effect of manipulating some marketing instruments on the probability of a certain choice between two alternatives.

The remainder of the paper is organized as follows. First, we describe the data used in this research. Next, we provide a brief overview of the construction of SVM for classification tasks. Sections 4 and 5 give an account of the obtained results and their interpretation and Section 6 gives a conclusion.

2 Data

We focus on a straightforward marketing problem: how to forecast holiday length on the basis of some general travelling and customer characteristics. These data have been collected by Erasmus University Rotterdam in 2003. Table 1 provides descriptive statistics for the data set. The dependent variable, holiday length, has been dichotomized into "not more than 14 days" and "more than 14 days". In total, there are 708 respondents. The outcome alternatives are quite balanced: 51.7% of the respondents have spent more than two weeks and 48.3% not more than two weeks of holidays. Eleven explanatory variables were available, some of which are categorical: destination, mode of transport, accommodation, full/nonfull board and lodging, sunshine availability, (other) big expenses, in/out of season, having/not having children, number of children, income group and age group.

3 Support Vector Machines for classification

Support Vector Machines (SVM) are rooted in statistical learning theory (Vapnik (1995)) and can be applied to both classification and regression problems. We consider here the supervised learning task of separating examples that belong to two classes. Consider a data set of n explanatory vectors $\{\mathbf{x}_i\}_{i=1}^n$ from $\mathbb{R}^m$ and corresponding classification labels $\{y_i\}_{i=1}^n$, where

Table 1. Descriptive statistics of the predictor variables for the holiday data set split by holiday length. For the categorical variables, the relative frequency is given (in %) and for numerical variables, the mean.

	Holiday length in days			Holiday length in days	
Variable	≤ 14	> 14	Variable	≤ 14	> 14
Transport			Destination		
Car	39.8	34.2	Inside Europe	87.7	66.7
Airplane	48.0	58.2	Outside Europe	12.3	33.3
Other	12.2	7.6	Accommodation		
Full board			Camping	17.5	27.9
Yes	25.7	18.3	Apartment	29.5	24.0
No	74.3	81.7	Hotel	33.6	27.6
Sunshine			Other	19.4	20.5
Important	83.9	88.5	Season		
Not important	16.1	11.5	High	38.6	43.2
Big expenses			Low	61.4	56.8
Made	26.0	26.5	Having children		
Not made	74.0	73.5	Yes	31.6	40.2
Mean no. of children	0.35	0.49	No	68.4	59.8
Mean age group	3.95	4.52	Mean income group	2.23	2.67

$y_i \in \{-1, 1\}$. Thus, in the marketing data set, -1 identifies short holiday length ($\leq$ 14 days) and 1 identifies long holiday length ($>$ 14 days). The SVM method finds the oriented hyperplane that maximizes the closest distance between observations from the two classes (the so-called "margin"), while at the same time minimizes the amount of training errors (Vapnik (1995), Cristianini and Shawe-Taylor (2000), Burges (1998)). In this way, good generalization ability of the resulting function is achieved, and therefore the problem of overfitting is mitigated.

The explanatory vectors $\mathbf{x}$ from the original space $\mathbb{R}^m$ are usually mapped into a higher dimensional, space, where their coordinates are given by $\mathbf{\Phi}(\mathbf{x})$. In this case, the optimal SVM hyperplane is found as the solution of the following optimization problem:

$$\begin{aligned} &\max_\alpha \ \textstyle\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ &\text{subject to } 0 \leq \alpha_i \leq C, \ i = 1, 2, \cdots, n, \ \text{and} \ \textstyle\sum_{i=1}^n y_i \alpha_i = 0, \end{aligned} \tag{1}$$

where $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i)'\mathbf{\Phi}(\mathbf{x}_j)$ is a kernel function that calculates dot products of explanatory vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in feature space. Intuitively, the kernel determines the level of proximity between any two points in the feature space. Common kernels in SVM are the linear $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j)$, polynomial $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^d$ and Radial Basis Function $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2)$ ones, where d and γ and manually adjustable parameters. The feature space implied by the RBF kernel is infinite-dimensional, while the linear kernel preserves the data in the original space. Maximizing the term $-\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$ corresponds to maximizing the margin between the two classes, which is equal to the distance between hyperplanes with equations $\sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b = -1$ and $\sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b = 1$. The manually adjustable constant C determines the trade-off between the margin and the amount of training errors. The α's are the weights associated with the observations. All observations with nonzero weights are called "support vectors", as they are the only ones that determine the position of the optimal SVM hyperplane. This hyperplane consists of all points $\mathbf{x}$ which satisfy $\sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b = 0$. The b parameter is found from the so-called Kuhn-Tucker conditions associated with (1).

The importance of binary classification methods lies in how well they are able to predict the class of a new observation $\mathbf{x}$. To do so with SVM, the optimal separation hyperplane $\sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b = 0$ that is derived from the solution $(\{\alpha_i\}_{i=1}^{n}, b)$ of (1) is used:

$$f(\mathbf{x}) = \mathrm{sign}(g(\mathbf{x})) = \mathrm{sign}\left(\sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b\right),$$

where $\mathrm{sign}(a) = -1$ if $a < 0$, $\mathrm{sign}(a) = 1$ if $a \geq 0$.

For interpretation, it is often important to know not only the predicted binary outcome, but also its probability. One way to derive posterior probabilities for the estimated class membership $f(\mathbf{x}_i)$ of observation $\mathbf{x}_i$ has been proposed by Platt (1999). His approach is to fit a sigmoid function to all estimated $g(\mathbf{x}_i)$ to derive probabilities of the form:

$$P(y = 1 | g(\mathbf{x}_i)) = p_i = (1 + \exp(a_1 g(\mathbf{x}_i) + a_2))^{-1},$$

where a_1 and a_2 are estimated by minimizing the negative log-likelihood of the training data:

$$\min_{a_1, a_2} \; -\sum_{i=1}^{n} \left(\frac{y_i + 1}{2} \log(p_i) + (1 - \frac{y_i + 1}{2}) \log(1 - p_i) \right).$$

4 Experiments and results

We define a training and a test sample, corresponding to 85% and 15% of the original data set, respectively. Our experiments have been carried out with the LIBSVM 2.6 software Chang and Lin (2004). We have constructed three SVM models, which differ in the transformation of the original data space, that is, using the linear, the polynomial of degree 2 ($d = 2$) and the RBF kernel. Table 2 shows detailed results of the SVM models as well as competing classification techniques in marketing such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and the logit choice model. The manually adjustable parameters C and γ have been estimated

Table 2. Hit rates (in %) of different learning methods for the vacation data set. Approximately 85% and 15% of each data set are used for training and testing, respectively. LDA, QDA and logit stand for Linear Discriminant Analysis, Quadratic Discriminant Analysis and logit choice model.

Sample		LDA	QDA	logit	lin SVM	poly SVM	RBF SVM
Training	≤ 14 days	68.2	69.2	63.3	73.0	78.9	77.5
	> 14 days	63.3	67.5	66.2	60.5	59.2	61.4
	Overall	65.7	68.3	64.8	66.5	68.7	69.8
Test	≤ 14 days	64.2	54.7	60.4	58.5	75.5	71.7
	> 14 days	56.4	54.6	65.5	49.1	45.5	52.7
	Overall	60.2	54.6	63.0	53.7	60.2	62.0

via a five-fold cross-validation procedure. As a result, the parameters for the linear, polynomial and RBF SVM models have been set as follows: $C = 2.5$, $C = 0.004$ and $d = 2$, $C = 3500$ and $\gamma = 0.0013$.

The overall performance of SVM on the test set is comparable to that of the standard marketing techniques. Among SVM models, the most flexible one (RBF-SVM) is also the most successful at generalizing the data. The average hit rate on the test set of all techniques considered centers at around 59%. There is no substantial distinction among the performance of all models, except for the QDA and linear SVM models, which relatively underperform. In such a setting we generally favor those models that can be better interpreted.

5 Interpreting the influence of the explanatory variables

The classical SVM appears to lack two main interpretation aspects shared by the standard models of LDA, QDA, and logit choice model. First, for the standard models, coefficient estimates for each explanatory variable are available and can be interpreted as the direct effect of a change in one of the independent variables on the dependent variable, while keeping all other independent variables fixed. The same interpretation is possible for the linear SVM model, since the original data space is preserved, and thus individual coefficient estimates are available. For all the other types of SVM this direct variable effect can be highly nonlinear and is not directly observable. The SVM with RBF kernel, for example, implies infinitely many number of explanatory variables, and thus infinitely many coefficients for each of these variables, which makes interpretation impossible at first sight.

Second, the coefficient estimates obtained from the standard models can be used to derive the effect of each explanatory variable on the probability of

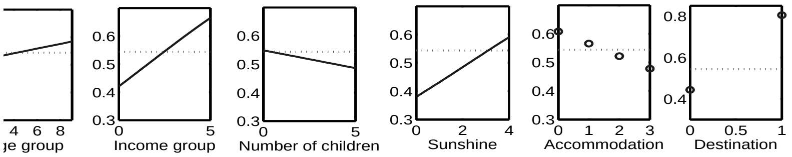

Fig. 1. Influences of individual explanatory variables on the probability to spend more than two weeks on a vacation for the logit model.

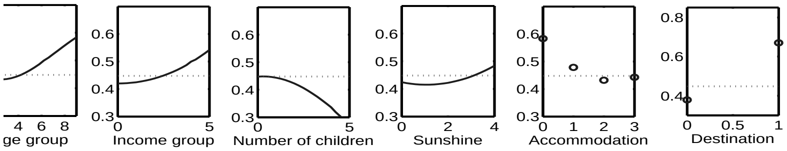

Fig. 2. Influences of individual explanatory variables on the probability to spend more than two weeks on a vacation for the RBF-SVM model.

a certain binary outcome. Although classical SVM does not output outcome probabilities, one can use here the proposed probability estimates by Platt (1999), discussed in Section 3. Interestingly, these probability estimates can help to derive individual variable effects also for the nonlinear SVM. For interpretation purposes, all that is needed is to visualize the relationship between a given explanatory variable and the probability to observe one of the two possible binary outcomes, while keeping the rest of the explanatory variables fixed. Thus, even for the SVM with RBF kernel it is not necessarily to know the coefficients for each data dimension in order to infer the influence of individual variables.

Next, we interpret the results of the SVM model with RBF kernel on the vacation data set and compare them with those from the logit model. Consider Figures 1 and 2 that show the relationships between some of the independent variables and the probability to go on a vacation for more than two weeks, for the logit and RBF-SVM models respectively. In each of the panels, the remaining explanatory variables are kept fixed at their average levels. The dashed lines denote the probability of the "average" person to go on a vacation for more than two weeks.

The first striking feature to observe is the great degree of similarity between both models. Although the RBF-SVM model is very flexible, the estimated effects for variables such as "Having children", "Big expenses", and "In season" are close to linear, just as the logit model predicts. The main difference between both techniques is best illustrated by the predicted effect of the "Age group" variable. The SVM model suggests that both relatively younger and relatively older holiday makers tend to have (on average) a higher probability to choose for the longer vacation option than the middle-aged ones, which makes sense intuitively. The logit model cannot capture such an effect by its definition as it imposes a monotonically increasing (or decreasing) relationship between the explanatory variables and the probability of a certain outcome. The RBF-SVM model, on the other hand, is free to impose a highly nonlinear such relationship via the mapping of the original data into a higher-dimensional space. Moreover, since the SVM model does not suffer from monotonicity restrictions, it reports nonmonotonically ordered outcome probabilities for each of the "Accommodation" variable categories (see Figure 2). Although one cannot conclude here that SVM is immune to the need to optimally scale the variables prior to model estimation, it is clear that it offers a better protection from arbitrary coding of unordered categorical variables than the logit model does.

The marketing implications of the results obtained by SVM can be derived directly from Figure 2. By considering the effects of changes in individual variables, marketeers can infer which ones are most effective and, as a result of this, streamline the advertising efforts accordingly. Thus, it seems most effective to offer longer-than-two-week vacations to customers with the following profile: relatively older, with high income, small number of children or no children at all, preferring to have sunshine available most of the time, and to a destination outside Europe.

6 Conclusion

We have analyzed a marketing classification problem with SVM for binary classification. We have also compared our results with those of standard marketing tools. Although the classical SVM exhibits superior performance, a general deficiency is that the results are hard to interpret, especially in the nonlinear case. To facilitate such an interpretation, we have constructed relationships between the explanatory and (binary) outcome variable by making use of probabilities for the SVM output estimates obtained from an approach proposed by Platt (1999). Ultimately, this allows for the possibility to evaluate the effectiveness of different marketing strategies under different scenarios. In terms of interpretation of the results, it appears that SVM models can give two advantages over standard techniques. First, highly nonmonotonic effects of the explanatory variables can be detected and visualized. And second, which comes as a by-product of the first, the SVM appears to model adequately the effects of arbitrarily coded unordered categorical variables.

References

ABE, M. (1991): A Moving Ellipsoid Method for Nonparametric Regression and Its Application to Logit Diagnostics With Scanner Data. *Journal of Marketing Research, 28, 339–346.*

ABE, M. (1995): A Nonparametric Density Estimation Method for Brand Choice Using Scanner Data. *Marketing Science, 14, 300-325.*

BENNETT, K.P., WU, S. and AUSLENDER, L. (1999): On Support Vector Decision Trees For Database Marketing. *IEEE International Joint Conference on Neural Networks (IJCNN '99), 2, 904–909.*

BURGES, C.J.C. (1998): A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2, 121–167.*

CHANG, C.C. and LIN, C.J. (2004): LIBSVM: a Library for Support Vector Machines. *Software available at*: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000): *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge.

CUI, D. (2003): Product Selection Agents: A Development Framework and Preliminary Application. *Unpublished doctoral dissertation.* University of Cincinnati, Business Administration: Marketing, Ohio. Retrieved April 5, 2005, from http://www.ohiolink.edu/etd/send-pdf.cgi?ucin1054824718

EVGENIOU, T. and PONTIL, M. (2004): Optimization Conjoint Models for Consumer Heterogeneity. INSEAD Working Paper, Serie No. 2004/10/TM, Fontaineblea: INSEAD.

FRANSES, P.H. and PAAP, R. (2001): *Quantitative Models in Marketing Research.* Cambridge University Press, Cambridge.

LATTIN, J., CARROLL, J. and GREEN, P. (2003): *Analyzing Multivariate Data.* Duxbury Press, Belmont, CA.

MÜELLER, K.-R., MIKA, S., RÄTSCH, G., TSUDA, K. and SCHÖLKOPF, B. (2001): An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks, 12(2), 181–201.*

PLATT, J. (1999): Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.): *Advances in Large Margin Classifiers.* MIT Press, Cambridge, MA, 61–74.

VAN HEERDE, H., LEEFLANG, P., and WITTINK, D. (2001): Semiparametric Analysis to Estimate the Deal Effect Curve. *Journal of Marketing Research, 38, 197-215.*

VAPNIK, V.N. (1995): *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc., (2nd edition, 2000).

WEST, P.M., BROCKETT, P.L. and GOLDEN, L.L. (1997): A Comparative Analysis of Neural, Networks and Statistical Methods for Predicting Consumer Choice. *Marketing Science, 16, 370–391.*