SUPPORT VECTOR MACHINES FOR BANKRUPTCY ANALYSIS

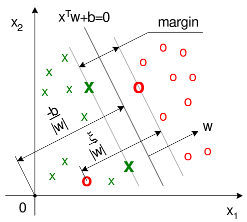
Wolfgang HÄRDLE [2]

Ralf KÖRNER [3]

Rouslan MORO [1,2]

Dorothea SCHÄFER [1]

[1] Deutsches Institut für Wirtschaftsforschung (DIW); [2] Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin; [3] Deutsche Bundesbank



Linear Discriminant Analysis

Fisher (1936); company scoring: Beaver (1966), Altman (1968)

Z-score:

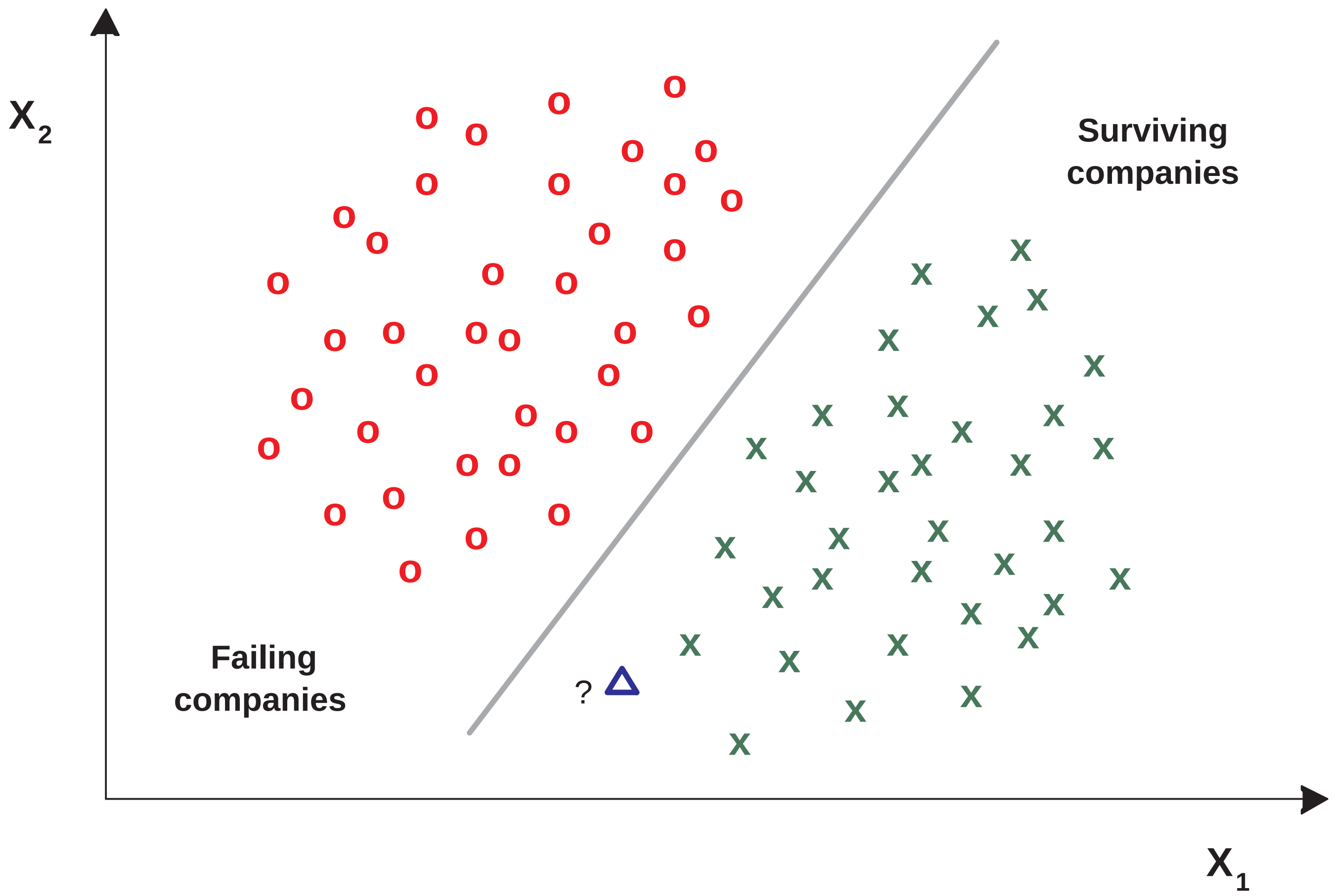
$$Z_i = a_1 x_{i1} + a_2 x_{i2} + \dots + a_d x_{id} = a^\top x_i,$$

where $x_i = (x_{i1}, \dots, x_{id})^\top \in \mathbb{R}^d$ are financial ratios for the i-th company.

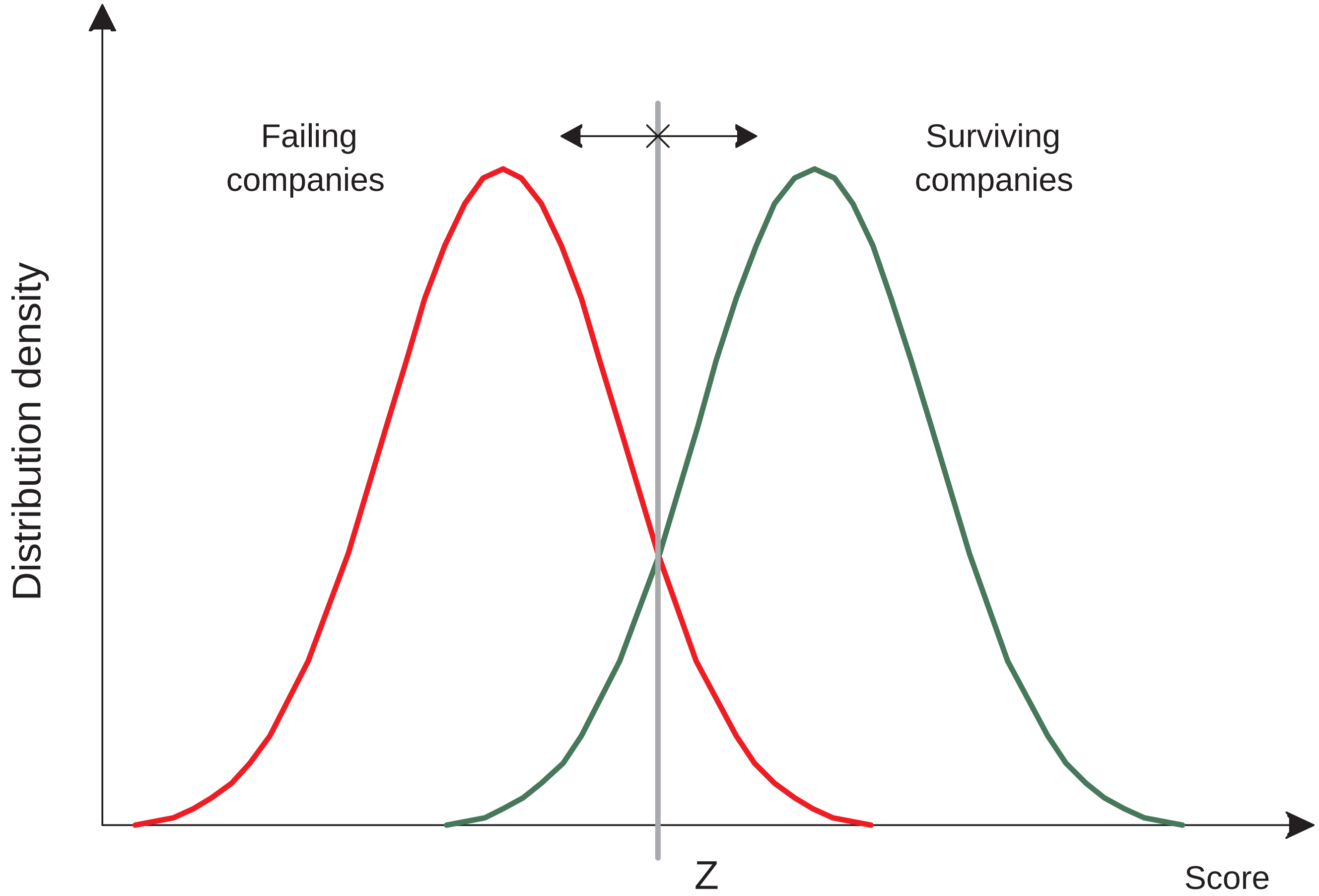
The classification rule:

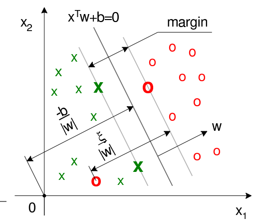
successful company:	$Z_i \geq z$
failure:	$Z_i < z$

Linear Discriminant Analysis

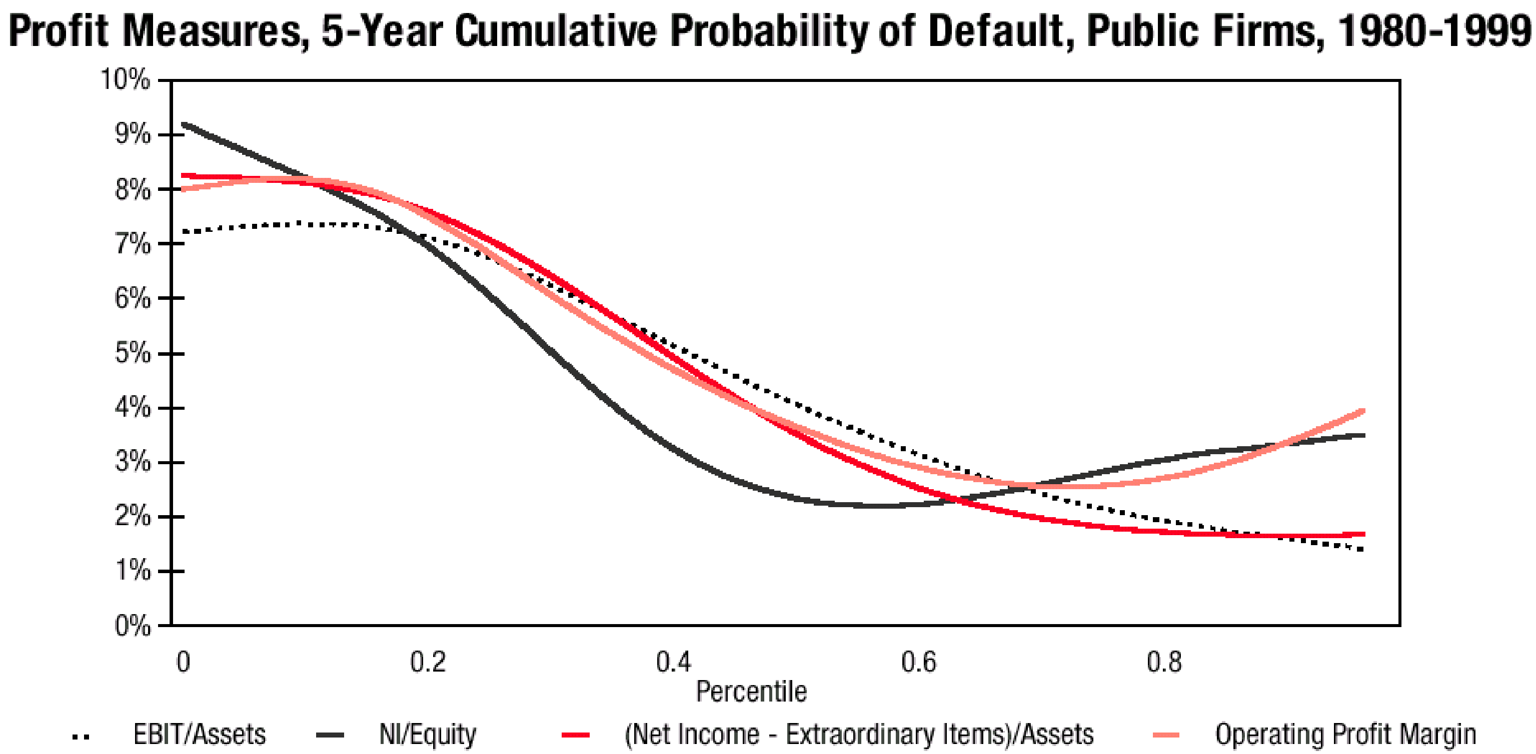


Linear Discriminant Analysis





Company Data: Probability of Default



Source: Falkenstein et al. (2000)

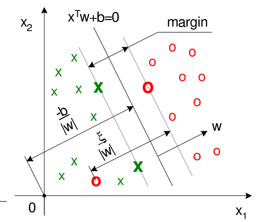
RiskCalc Private Model

Moody's default model for private firms

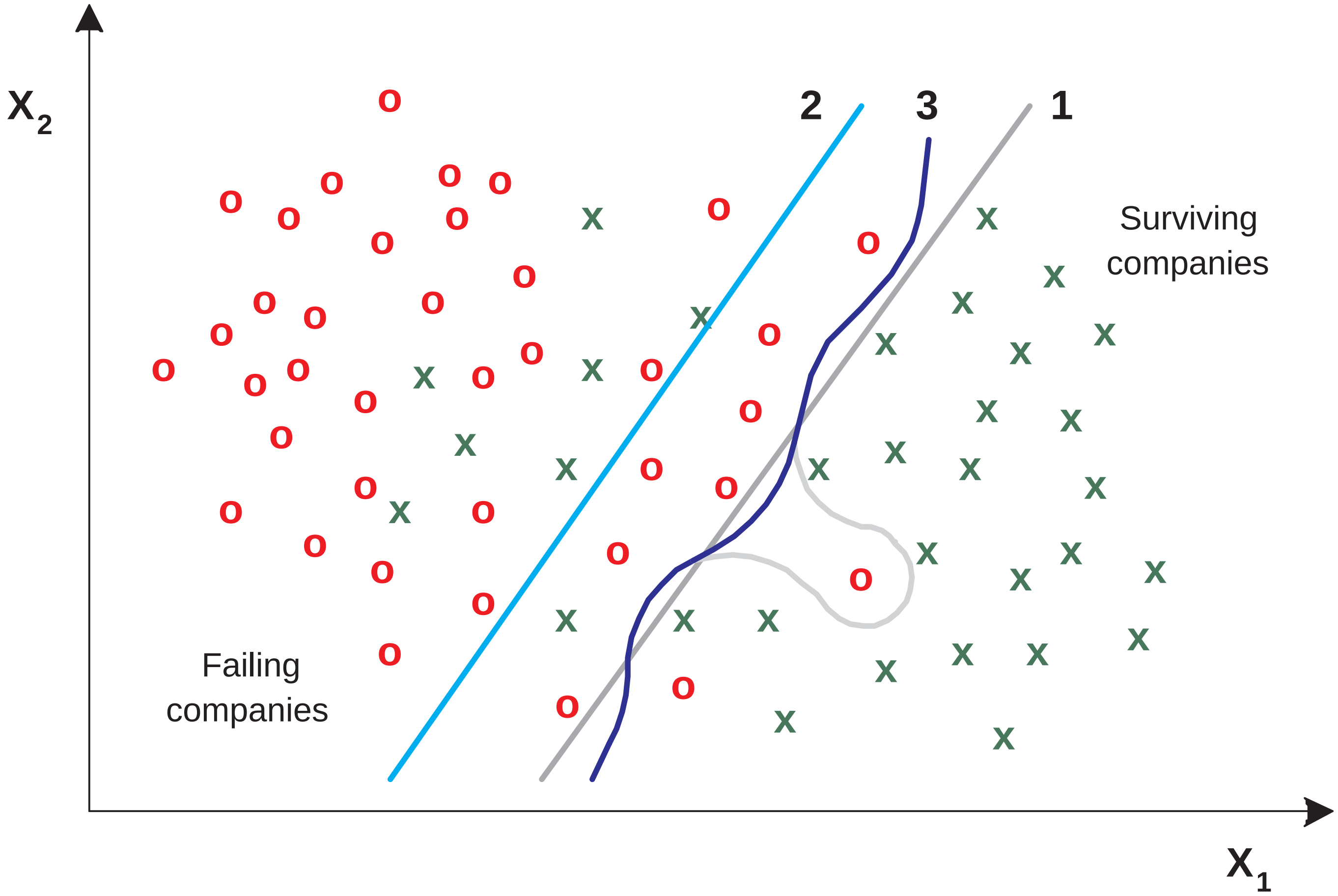
A semi-parametric model based on the probit regression

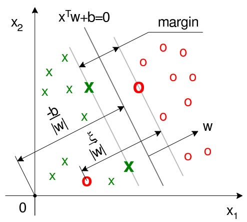
$$E[y_i|x_i] = \Phi\{a_0 + \sum_{j=1}^{d} a_j f_j(x_{ij})\}$$

f_j are estimated non-parametrically on univariate models



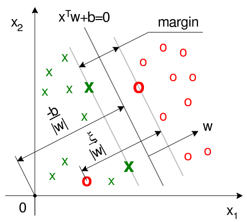
Linearly Non-separable Classification Problem





Outline

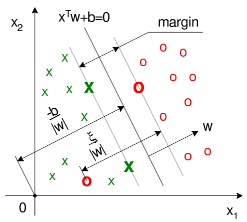
✓ 1. Motivation
2. Support Vector Machines and their Properties
3. Expected Risk vs. Empirical Risk Minimization
4. Realization of an SVM
5. Non-linear Case
6. Company Classification and Rating with SVMs



Support Vector Machines (SVMs)

SVMs are a group of methods for classification (and regression) that make use of classifiers providing “high margin”.

- SVMs possess a flexible structure which is not chosen a priori
- The properties of SVMs can be derived from statistical learning theory
- SVMs do not rely on asymptotic properties; they are especially useful when d/n is big, i.e. in most practically significant cases
- SVMs give a unique solution and often outperform Neural Networks



Classification Problem

Training set: $\{(x_i, y_i)\}_{i=1}^{n}$ with the distribution $P(x_i, y_i)$.

Find the class y of a new object x using the classifier $f : \mathbb{R}^d \mapsto \{+1; -1\}$, such that **the expected risk** $R(f)$ **is minimal**.

$x_i \in \mathbb{R}^d$ is the vector of the i-th object characteristics;

$y_i \in \{-1; +1\}$ or $\{0; 1\}$ is the class of the i-th object.

Expected Risk Minimization

Expected risk

$$R(f) = \int \frac{1}{2}|g(x) - y| dP(x,y) \quad = \mathrm{E}_{P(x,y)}[L(x,y)]$$

is minimized wrt $f(x)$, where $g(x) = \mathrm{sign}\{f(x)\}$:

$$f_{opt} = \arg\min_{f \in \mathcal{F}} R(f)$$

$$L(x,y) = \frac{1}{2}|g(x) - y| = \begin{cases} 0, & \text{if classification is correct,} \\ 1, & \text{if classification is wrong.} \end{cases}$$

$\mathcal{F}$ is an a priori defined set of (non)linear classifier functions

Empirical Risk Minimization

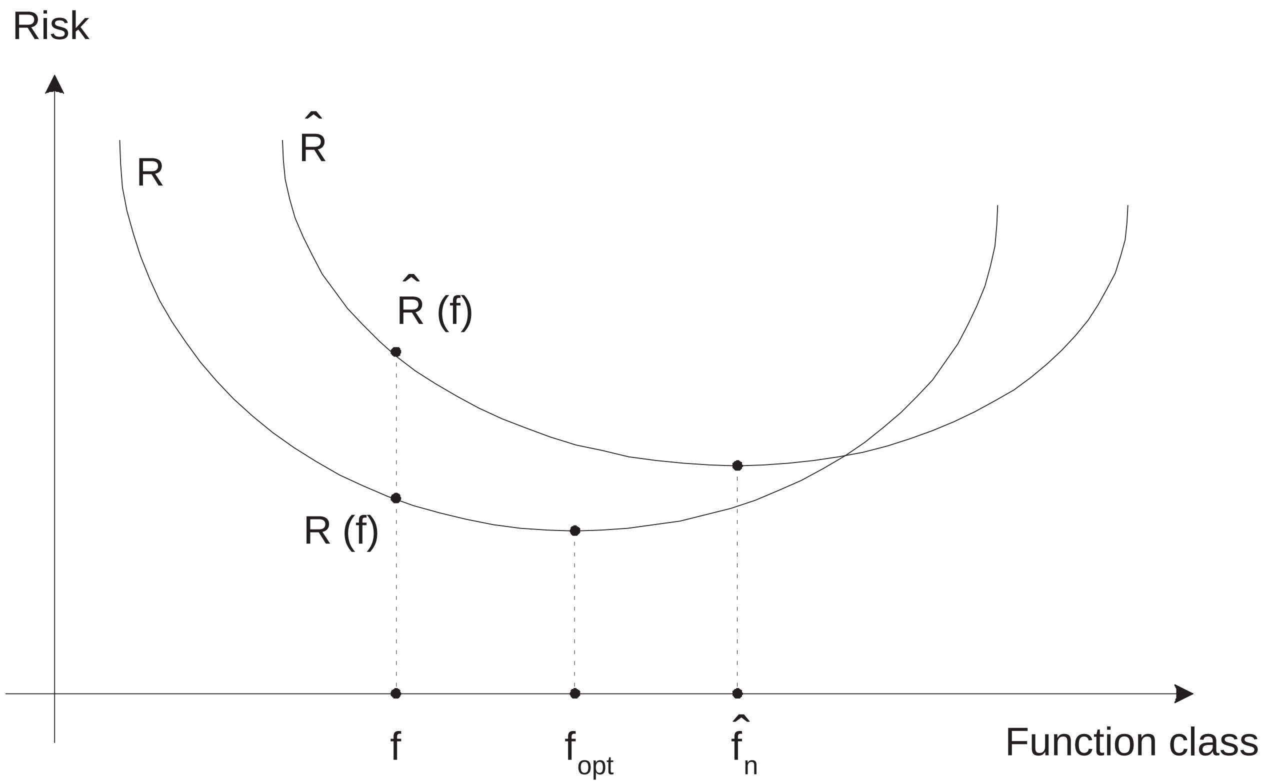
In practice $P(x, y)$ is usually **unknown**: use *Empirical Risk*

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}|g(x_i) - y_i|$$

Minimization (ERM) over the training set $\{(x_i, y_i)\}_{i=1}^{n}$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$$

Empirical Risk vs. Expected Risk



Convergence

From the law of large numbers

$$\lim_{n \to \infty} \hat{R}(f) = R(f)$$

In addition ERM satisfies

$$\lim_{n \to \infty} \min_{f \in \mathcal{F}} \hat{R}(f) = \min_{f \in \mathcal{F}} R(f)$$

if "$\mathcal{F}$ is not too big".

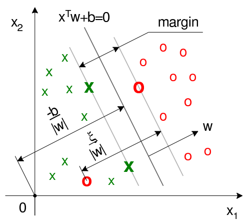
Vapnik-Chervonenkis (VC) Bound

Basic result of Statistical Learning Theory (for linear classifiers):

$$R(f) \leq \hat{R}(f) + \phi\left(\frac{h}{n}, \frac{\log(\eta)}{n}\right)$$

where the bound holds with probability $1 - \eta$ and

$$\phi\left(\frac{h}{n}, \frac{\log(\eta)}{n}\right) = \sqrt{\frac{h(\log \frac{2n}{h} + 1) - \log(\frac{\eta}{4})}{n}}$$



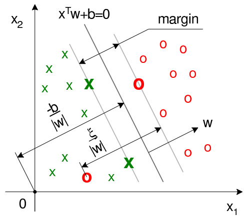
Structural Risk Minimization

Search for the model structure $\mathcal{S}_h$,
$\mathcal{S}_{h1} \subseteq \mathcal{S}_{h2} \subseteq \ldots \subseteq \mathcal{S}_h \subseteq \ldots \subseteq \mathcal{S}_{hk} \subseteq \mathcal{F}$, such that $f \in \mathcal{S}_h$ minimizes the expected risk upper bound.

h **is VC dimension**. It is a measure of complexity. S_h is a set of classifier functions with the same complexity described by h.

Example 1: $P(1) \subseteq P(2) \subseteq P(3) \subseteq \ldots \subseteq \mathcal{F}$, where $P(i)$ are polynomials of degree i (here $P(1) = S_{h1}$, $P(2) = S_{h2}$, etc.).

The functional class $\mathcal{F}$ is given **a priori**



Example 2:

$$\mathrm{SVM}(h_1) \subseteq \mathrm{SVM}(h_2) \subseteq \ldots \subseteq \mathrm{SVM}(h_i) \subseteq \ldots \subseteq \mathcal{F} = \mathrm{SVM}(h_N),$$

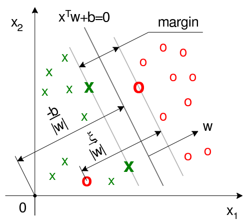
where h_i are function f complexities (VC dimensions):

$$0 = h_1 \leq h_2 \leq \ldots \leq h_i \leq \ldots \leq h_N = d + 1.$$

This corresponds to:

$$\infty = \mathrm{margin}_1 \geq \mathrm{margin}_2 \geq \ldots \geq \mathrm{margin}_i \geq \ldots \geq \mathrm{margin}_N = 0,$$

$i = 1, 2, \ldots, N$



Vapnik-Chervonenkis (VC) Dimension

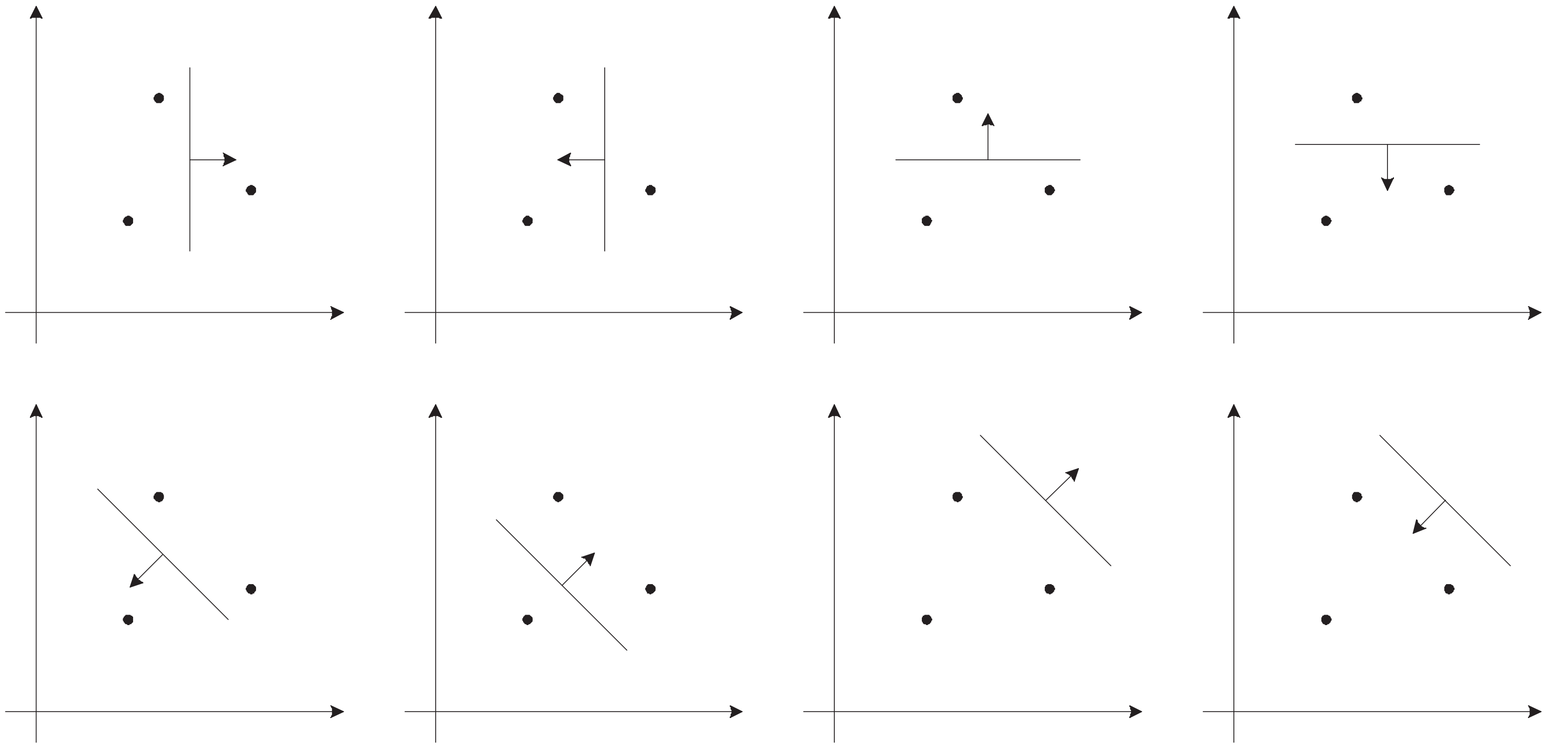
Definition. h is VC dimension of a set of functions if there exists a set of points $\{x_i\}_{i=1}^{h}$ such that these points can be separated in all 2^h possible configurations, and no set $\{x_i\}_{i=1}^{q}$ exists where $q > h$ satisfies this property.

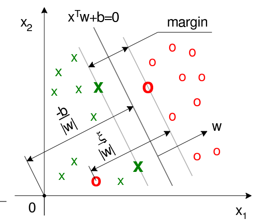
Example 1. The functions $f = A \sin \theta x$ have an infinite VC dimension.

Example 2. Three points on a plane can be shattered by a set of linear indicator functions in $2^h = 2^3 = 8$ ways (whereas 4 points cannot be shattered in $2^q = 2^4 = 16$ ways). The VC dimension equals $h = 3$.

Example 3. The VC dimension of $f = \{\text{Hyperplane} \in \mathbb{R}^d\}$ is $h = d + 1$.

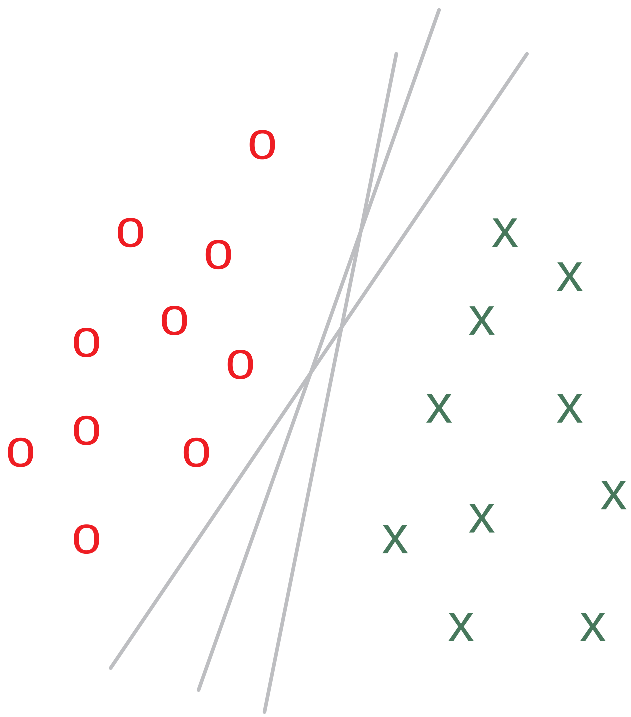
VC Dimension (d=2, h=3)

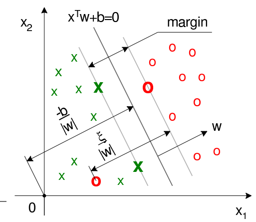




Linearly Separable Case

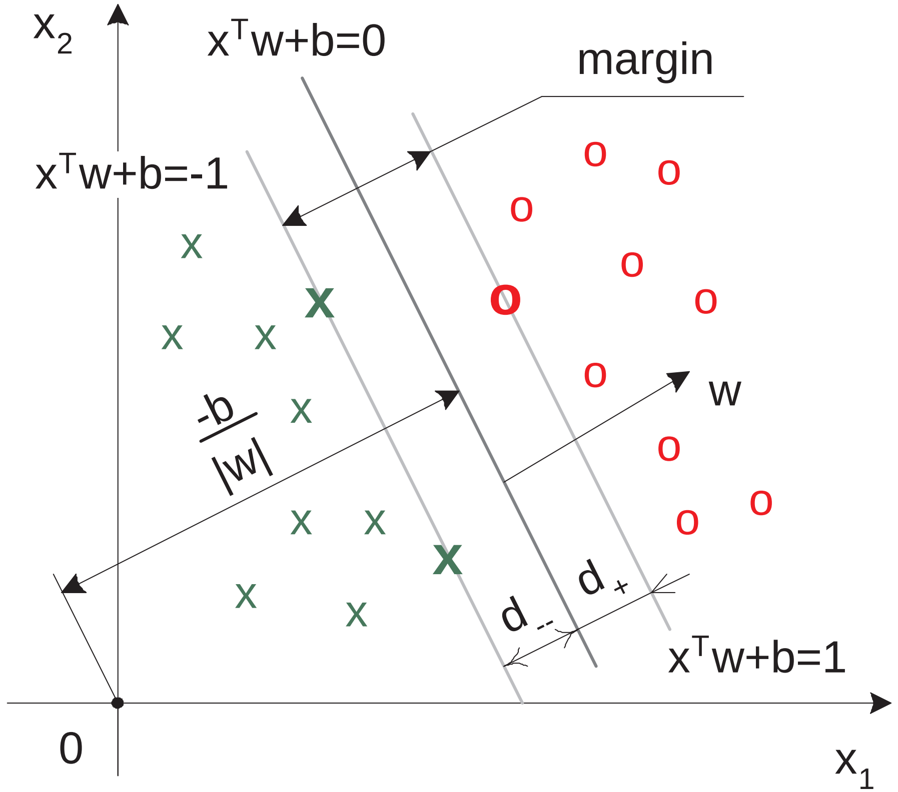
The training set: $\{(x_i, y_i)\}_{i=1}^{n}$, $y_i = \{+1; -1\}$, $x_i \in \mathbb{R}^d$. Find the classifier with the highest "margin" – the gap between parallel hyperplanes separating two classes where the vectors of neither class can lie. **Margin maximization minimizes the VC dimension.**

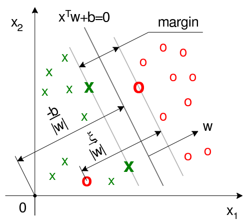




Linear SVMs. Separable Case

The **margin** is $d_+ + d_- = 2/\|w\|$. To maximize it minimize the Euclidean norm $\|w\|$ subject to the constraint (1).





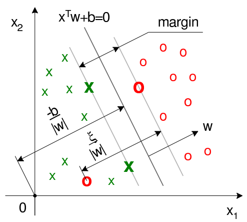
Let $x^\top w + b = 0$ be a separating hyperplane. Then d_+ (d_-) will be the shortest distance to the closest objects from the classes $+1$ (-1).

$x_i^\top w + b \geq +1$ for $y_i = +1$
$x_i^\top w + b \leq -1$ for $y_i = -1$

combine them into one constraint

$$y_i(x_i^\top w + b) - 1 \geq 0 \qquad i = 1, 2, ..., n \tag{1}$$

The canonical hyperplanes $x_i^\top w + b = \pm 1$ are parallel and the distance between each of them and the separating hyperplane is
$d_+ = d_- = 1/\|w\|$.



The Lagrangian Formulation

The Lagrangian for the primal problem

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \{ y_i (x_i^\top w + b) - 1 \}$$

The Karush-Kuhn-Tucker (KKT) Conditions

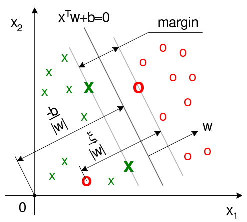
$$\frac{\partial L_P}{\partial w_k} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i x_{ik} = 0 \qquad k = 1, ..., d$$

$$\frac{\partial L_P}{\partial b} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$y_i (x_i^\top w + b) - 1 \geq 0 \qquad i = 1, ..., n$$

$$\alpha_i \geq 0$$

$$\alpha_i \{ y_i (x_i^\top w + b) - 1 \} = 0$$



Substitute the KKT conditions into L_P and obtain the Lagrangian for the dual problem

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

The primal and dual problems are

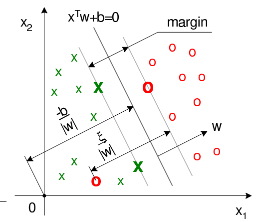
$$\min_{w_k, b} \max_{\alpha_i} L_P$$

$$\max_{\alpha_i} L_D$$

s.t.

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Since the optimization problem is convex the dual and primal formulations give the same solution.



The Classification Stage

The classification rule is:

$$g(x) = \operatorname{sign}(f) = \operatorname{sign}(x^\top w + b)$$

where

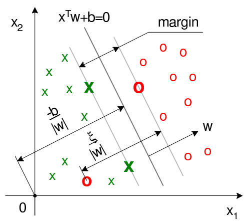
$w = \sum_{i=1}^{n} \alpha_i y_i x_i$

$b = \frac{1}{2}(x_+ + x_-)^\top w$

x_+ and x_- are any support vectors from each class

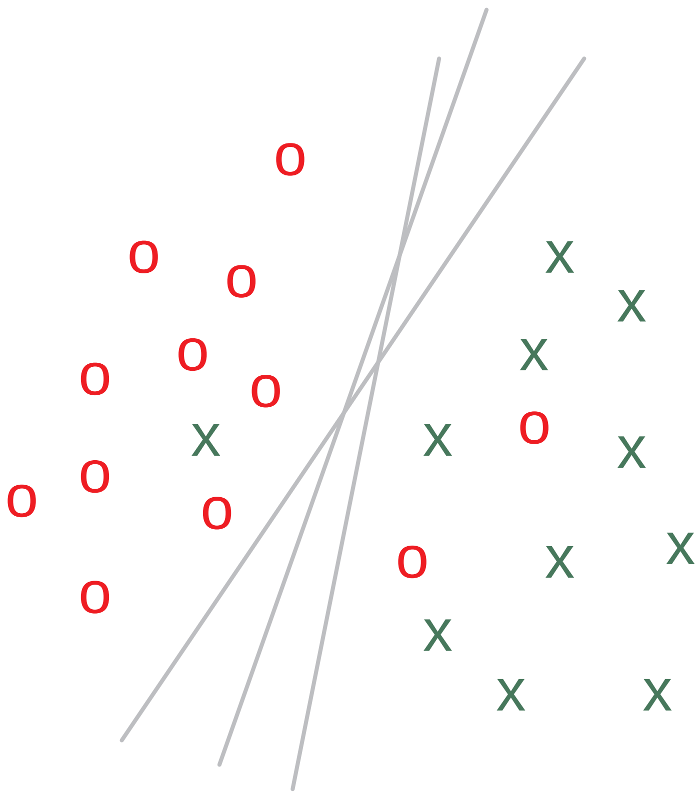
$$\alpha_i = \arg\max_{\alpha_i} L_D$$

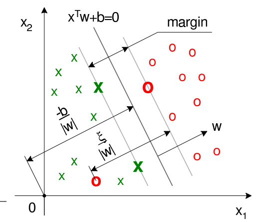
subject to the constraint $y_i(x_i^\top w + b) - 1 \geq 0 \qquad i = 1, 2, ..., n.$



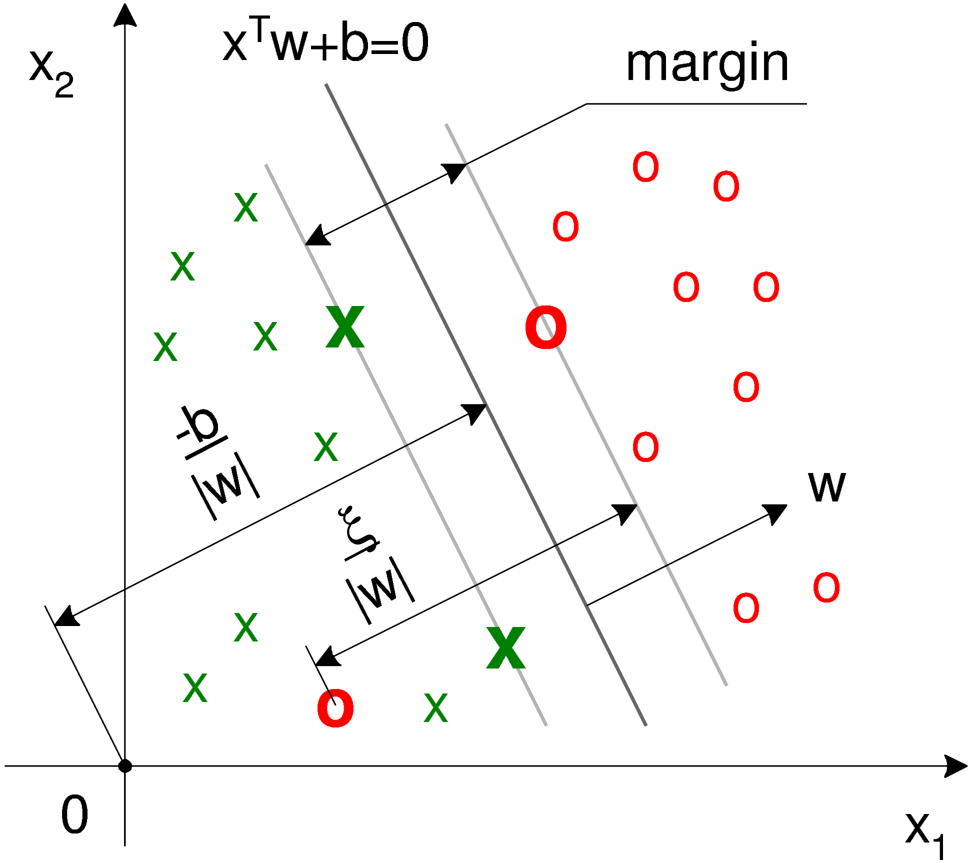
Linear SVMs. Non-separable Case

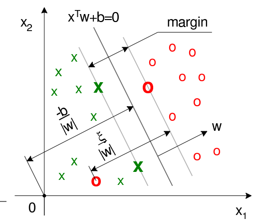
In the non-separable case it is impossible to separate the data points with hyperplanes without an error.





Linear SVM. Non-separable Case





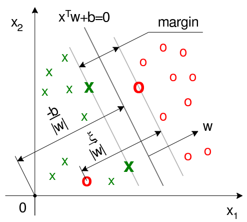
The problem can be solved by introducing **positive slack variables** $\{\xi_i\}_{i=1}^n$ into the constraints

$$
\begin{aligned}
x_i^\top w + b &\geq 1 - \xi_i \quad \text{for} \quad y_i = 1\\
x_i^\top w + b &\leq -1 + \xi_i \quad \text{for} \quad y_i = -1\\
\xi_i &\geq 0 \qquad \forall i
\end{aligned}
$$

If an error occurs, $\xi_i > 1$. The objective function:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i$$

where C ("capacity") controls the tolerance to errors on the training set. Under such a formulation the problem is **convex**



The Lagrangian Formulation

The Lagrangian for the primal problem for $\nu = 1$:

$$L_P = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i(x_i^\top w + b) - 1 + \xi_i\} - \sum_{i=1}^{n}\xi_i\mu_i$$

The primal problem:

$$\min_{w_k, b, \xi_i} \max_{\alpha_i, \mu_i} L_P$$

The KKT Conditions

$$\frac{\partial L_P}{\partial w_k} = 0 \quad \Leftrightarrow \quad w_k = \sum_{i=1}^{n} \alpha_i y_i x_{ik} \qquad k = 1, ..., d$$

$$\frac{\partial L_P}{\partial b} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \quad \Leftrightarrow \quad C - \alpha_i - \mu_i = 0$$

$$y_i(x_i^\top w + b) - 1 + \xi_i \geq 0$$

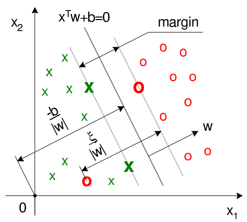
$$\xi_i \geq 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0$$

$$\alpha_i \{y_i(x_i^\top w + b) - 1 + \xi_i\} = 0$$

$$\mu_i \xi_i = 0$$



The dual Lagrangian does not contain ξ_i or their Lagrange multipliers

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \tag{2}$$

The dual problem is

$$\max_{\alpha_i} L_D$$

subject to

$$0 \leq \alpha_i \leq C$$

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

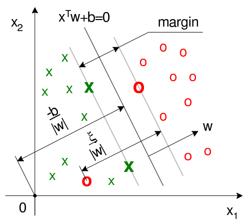
Non-linear SVMs

Map the data to a Hilbert space $\mathcal{H}$ and perform classification there

$$\Psi : \mathbb{R}^d \mapsto \mathcal{H}$$

Note, that in the Lagrangian formulation (2) the training data appear only in the form of **dot products** $x_i^\top x_j$, which can be mapped to $\Psi(x_i)^\top \Psi(x_j)$.

If a *kernel function* K exists such that $K(x_i, x_j) = \Psi(x_i)^\top \Psi(x_j)$, then we can use K **without knowing** Ψ **explicitly**



Hilbert spaces $\mathcal{H}$ are highly or infinitely (e.g. produced by Gaussian kernels) dimensional.

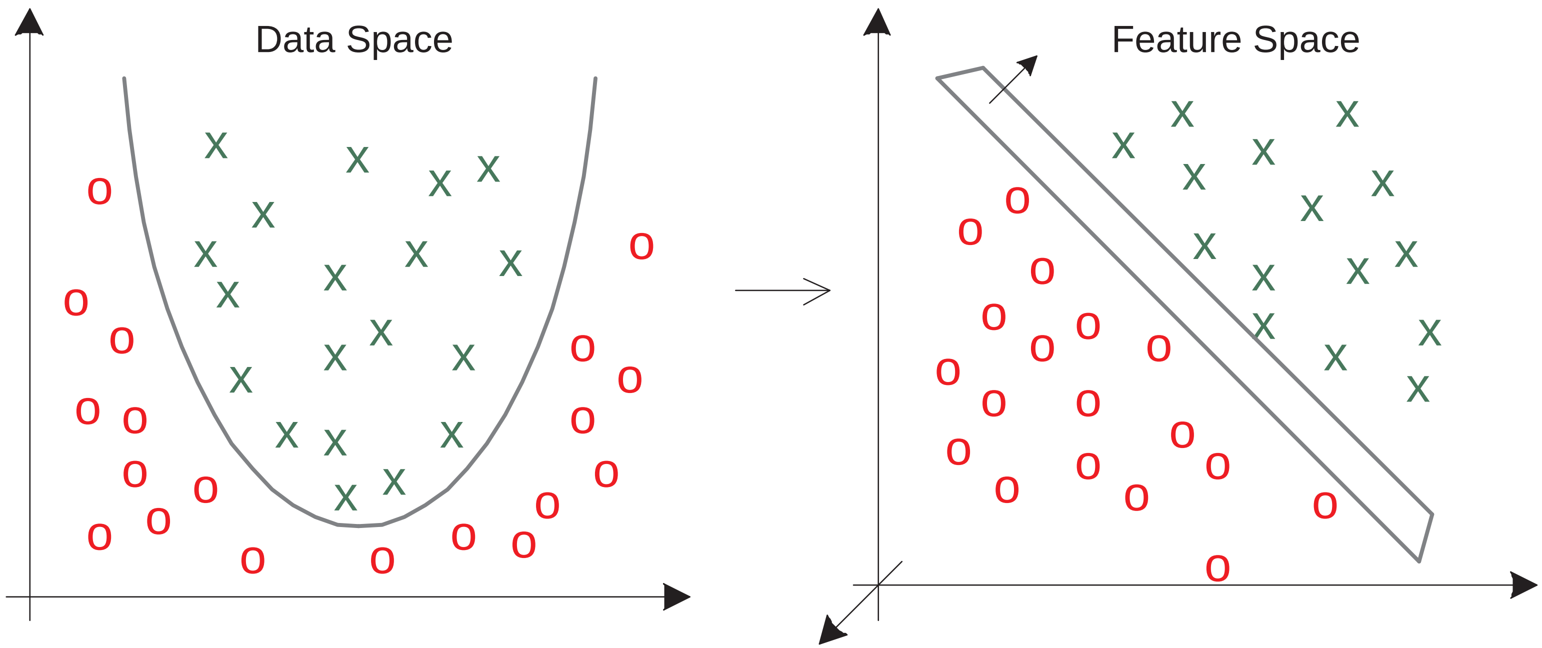
The SVM is well suited to be applied in $\mathcal{H}$ since:

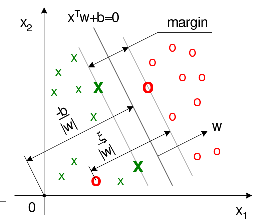
- it is not necessary to know Ψ and Ψ^{-1}. Only the kernel function $K(x_i, x_j)$ is required that is a scalar product in $\mathcal{H}$, and only $K(x_i, x_j)$ will appear in the SVM Lagrangian formulation
- the SVM is a regularized method that under reasonable parameters is not overfitted when $d \rightarrow \infty$

Mapping into the Feature Space. Example

$\mathbb{R}^2 \mapsto \mathbb{R}^3$,

$$\Psi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top, \quad K(x_i, x_j) = (x_i^\top x_j)^2$$





Mercer's Condition (1909)

A necessary and sufficient condition for a symmetric function $K(x_i, x_j)$ to be a kernel is that it must be positive definite, i.e. for any $x_1, ..., x_n \in \mathbb{R}^d$ and any $\lambda_1, ..., \lambda_n \in \mathbb{R}$ the function K must satisfy:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j K(x_i, x_j) \geq 0$$

Examples of kernel functions:

$K(x_i, x_j) = e^{-(x_i - x_j)^\top (x_i - x_j)/2\sigma^2}$	isotropic Gaussian kernel
$K(x_i, x_j) = e^{-(x_i - x_j)^\top \Sigma^{-1} (x_i - x_j)/2}$	anisotropic Gaussian kernel
$K(x_i, x_j) = (x_i^\top x_j + 1)^p$	polynomial kernel

Classes of Kernels

Stationary kernel is a kernel which is translation invariant:

$$K(x_i, x_j) = K_S(x_i - x_j)$$

Isotropic (homogeneous) kernel is one which depends only on the distance between two data points:

$$K(x_i, x_j) = K_I(\|x_i - x_j\|)$$

Local stationary kernel is a kernel of the form:

$$K(x_i, x_j) = K_1(\frac{x_i + x_j}{2})K_2(x_i - x_j)$$

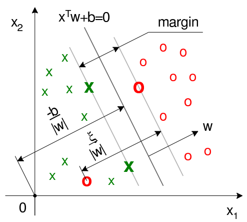
where K_1 is a non-negative function, K_2 is a stationary kernel.

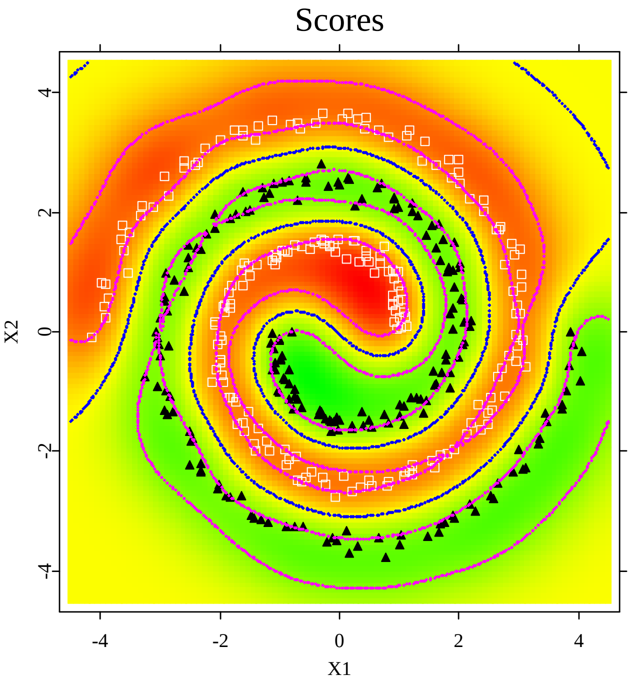
The SVM is Described By

- the kernel $K(x_i, x_j)$ that determines Ψ
- the capacity C

Here we are using $K(x_i, x_j) = \mathrm{pdf}\,\{\mathrm{N}_d(x_i - x_j, \Sigma)\}$ with the estimate $\hat{\Sigma} = cov(x)$ defining an anisotropic radial basis.

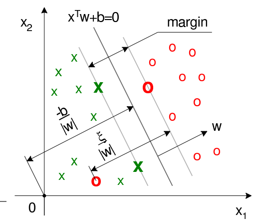
The capacity is a **bandwidth** parameter determining the width of the margin zone (the higher is C the narrower is the margin zone). It has to be estimated by out-of-sample prediction

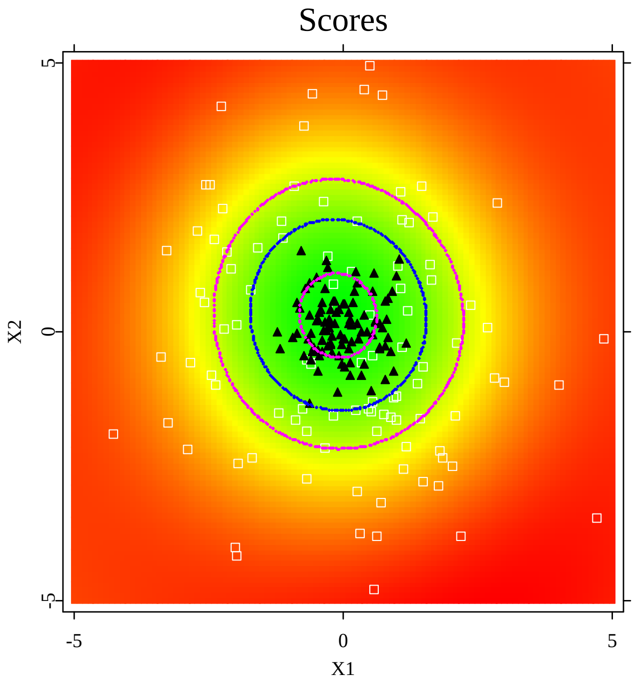




spr04c1p2.xpl

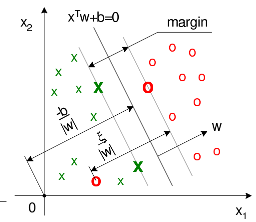
Figure 1: SVM classification results for slightly noisy spiral data ($RB = 0.4\hat{\Sigma}^{1/2}$, $C = 1.2/n$). The spirals spread over 3π radian; the distance between the spirals equals 1. The noise was injected with the parameters $\varepsilon_i \sim \mathrm{N}(\mathbf{0}, 0.1^2\mathcal{I})$. The separation is perfect.

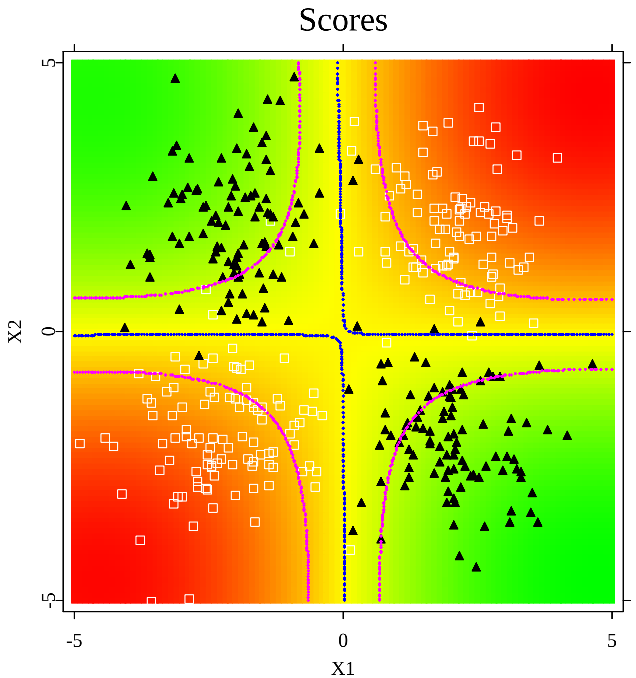




opr2c1.xpl

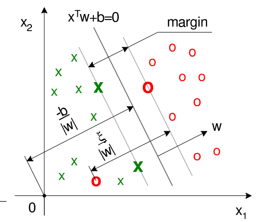
Figure 2: SVM classification results for the “orange peel” data ($RB = 2\hat{\Sigma}^{1/2}$, $C = 1/n$). $d = 2$, $n_{-1} = n_{+1} = 100$, $x_{+1,i} \sim \mathrm{N}(\mathbf{0}, 2^2\mathcal{I})$, $x_{-1,i} \sim \mathrm{N}(\mathbf{0}, 0.5^2\mathcal{I})$. Accuracy: 84% correctly cross-validated observations





qur2c2.xpl

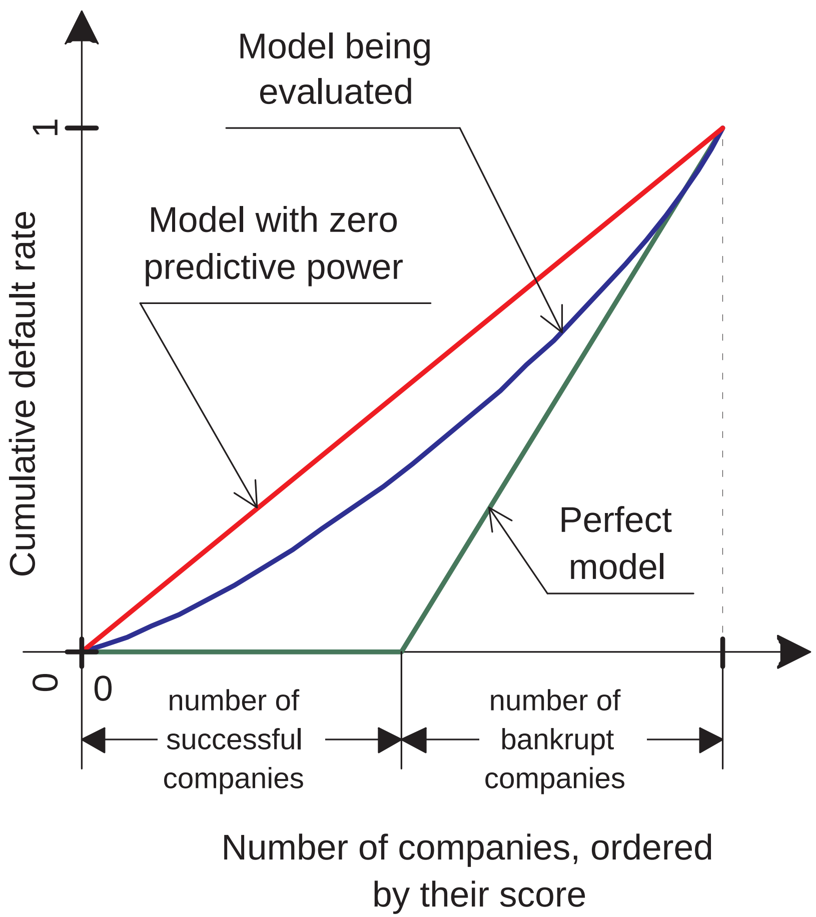
Figure 3: SVM classification results ($RB = 2\hat{\Sigma}^{1/2}$, $C = 2/n$). $d = 2$, $n_{-1} = n_{+1} = 200$; $x_{+1,i} \sim \frac{1}{2}\mathsf{N}((2,2)^\top, \mathcal{I}) + \frac{1}{2}\mathsf{N}((-2,-2)^\top, \mathcal{I})$, $x_{-1,i} \sim \frac{1}{2}\mathsf{N}((2,-2)^\top, \mathcal{I}) + \frac{1}{2}\mathsf{N}((-2,2)^\top, \mathcal{I})$. Accuracy: 96.3% correctly cross-validated observations



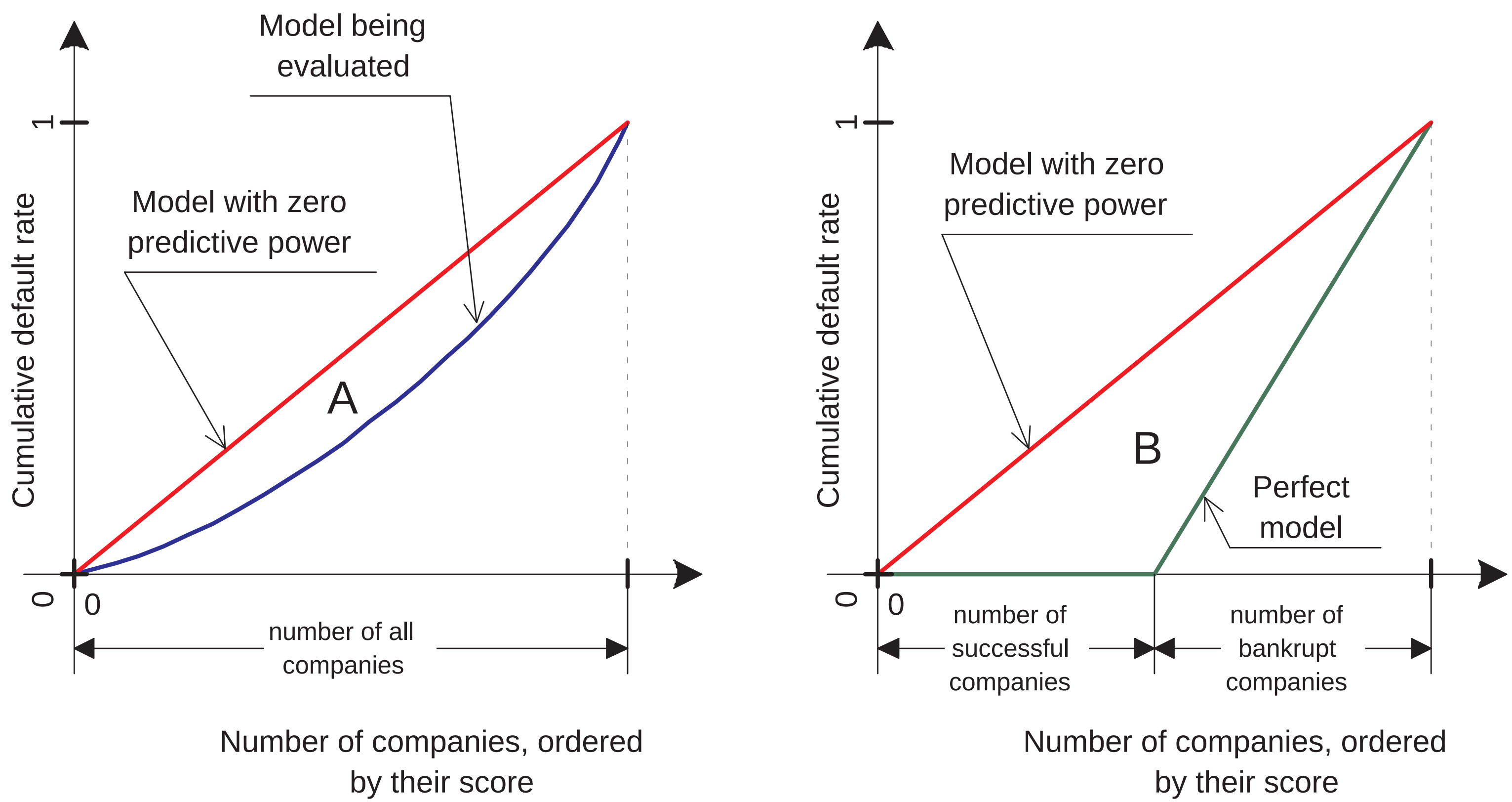
Out-of-Sample Accuracy Measures

- Percentage of correctly cross-validated observations
- Percentage of correctly validated out-of-sample observations, α- and β-errors
- Power curve (PC) aka Lorenz curve or cumulative accuracy profile. PC for a real model lies between PCs for the perfect and zero predictive power models
- Accuracy ratio (AR)

Cumulative Accuracy Profile Curve



Accuracy Ratio



Accuracy Ratio (AR) $=$ A/B

Data Selection

Source: Bundesbank's Central Corporate Database

Around 350 000 balance sheets, 100 000 companies

Selected were manufacturing, private companies with turnover >36 000 EUR a year

About 20 000 operating and 1100 bankrupt companies satisfied these and other minor criteria

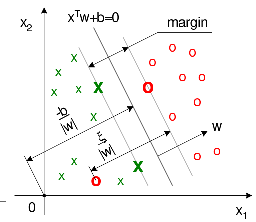
All bankruptcies took place in 1997-2004 no later than three years and no sooner than three months after the last report was submitted

Data Selection (Cont.)

- included were 1028 bankrupt companies that satisfied the criteria. The same number (1028) of operating companies was randomly chosen from 20 000 to avoid different penalties for the two groups
- there is only one observation for each company in the data set
- the whole data set was randomly split into training (509 operating and failed firms) and validation (519 operating and failed firms) sets
- altogether, the 2056 companies are described by 26 financial ratios in % ($d = 26$, $n = 2056$)

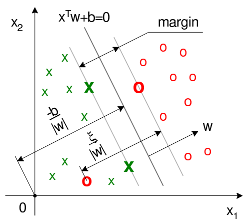
Variables and Their Predictive Power

No.	Name (Eng.)	Name (Ger.)	max AR
K1	Pre-tax profit margin	Umsatzrendite	0.496
K2	Operating profit margin	Betriebsrendite	0.410
K3	Cash flow ratio	Einnahmenüberschussquote	0.456
K4	Capital recovery ratio	Kapitalrückflussquote	0.519
K5	Debt cover	Schuldentilgungsfähigkeit	0.553
K6	Days receivable	Debitorenumschlag	0.161
K7	Days payable	Kreditorenumschlag	0.471
K8	Equity ratio	Eigenkapitalquote	0.408



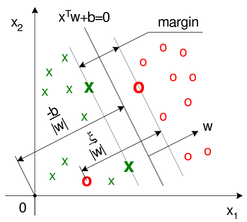
No.	Name (Eng.)	Name (Ger.)	max AR
K9	Equity ratio (adj.)	Eigenmittelquote	0.386
K11	Net income ratio	Umsatzrendite ohne a.E.	0.549
K13	Debt ratio	Finanzbedarfsquote	0.287
K15	Liquidity 1	Liquiditätsgrad 1	0.269
K16	Liquidity 2	Liquiditätsgrad 2	0.335
K17	Liquidity 3	Liquiditätsgrad 3	0.264
K18	Short term debt ratio	kurzfr. Fremdkapitalquote	0.309
K19	Inventories ratio	Vorratsquote	0.146
K21	Net income change	Umsatzveränderungen	0.193

No.	Name (Eng.)	Name (Ger.)	max AR
K22	Own funds yield	Eigenkapitalrendite	0.266
K23	Capital yield	Gesamtkapitalrendite	0.454
K24	Net interest ratio	Nettozinsquote	0.464
K25	Own funds/pension prov. r.	Pensionsrückstellungsquote	0.432
K27	Own funds/provisions ratio	Eigenkapitalrückstellungsq.	0.470
K29	Interest coverage ratio	Zinsdeckung	0.573
K30	Cash flow ratio	Einnahmenüberschußquote	0.395
K31	Days of inventories	Lagedauer	0.292
K32	Current liabilities ratio	Fremdkapitalstruktur	0.405

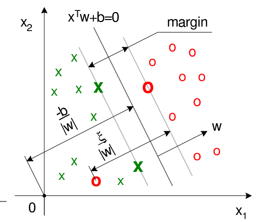


Summary Statistics

Predictor	Group	Min	Max	Mean	STD
K1	Profitability	-123.6	146.0	0.7	11.7
K2	Profitability	-178.4	165.5	2.0	10.8
K3	Liquidity	-442.6	177.7	3.4	16.7
K4	Liquidity	-327.8	603.3	8.0	23.5
K5	Liquidity	-42.0	3277.1	20.6	125.1
K6	Activity	0.0	826.0	45.6	36.2
K7	Activity	0.0	1822.8	41.5	59.2
K8	Financing	-92.0	82.8	10.5	21.1



Predictor	Group	Min	Max	Mean	STD
K9	Financing	-92.0	87.8	15.5	22.3
K11	Profitability	-119.6	215.8	0.2	12.1
K13	Liquidity	-89.7	588.3	-1.9	25.0
K15	Liquidity	0.0	316.1	12.2	31.1
K16	Liquidity	1.4	669.1	72.2	61.4
K17	Liquidity	1.4	1739.2	132.9	97.4
K18	Financing	2.8	554.2	51.2	26.3
K19	Investment	0.0	95.5	30.2	18.7
K21	Growth	-91.8	7780.3	11.9	191.6



Predictor	Group	Min	Max	Mean	STD
K22	Profitability	-87200.0	11440.0	-41.1	2269.4
K23	Profitability	-64.0	580.7	6.3	18.3
K24	Cost structure	-4.8	50.8	2.1	2.9
K25	Financing	-92.0	100.1	15.7	23.9
K27	Financing	-84.9	121.0	22.6	26.3
K29	Cost structure	-203780.0	1037200.0	2073.7	29559.0
K30	Liquidity	-136.7	128.3	4.6	10.9
K31	Activity	0.0	933.5	71.6	64.0
K32	Financing	100.0	20983.0	421.1	1099.8

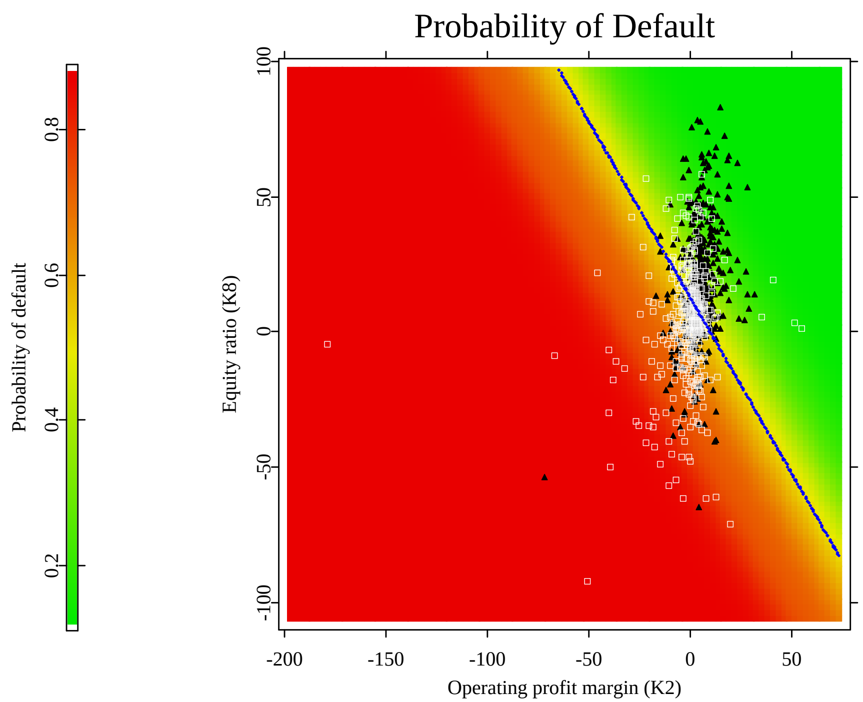


Figure 4: The data wtih a DA and logit classifying functions. For DA the percentage of correctly classified out-of-sample observations: 68.0%; accuracy ratio: 0.497; α-error: 33.3%.

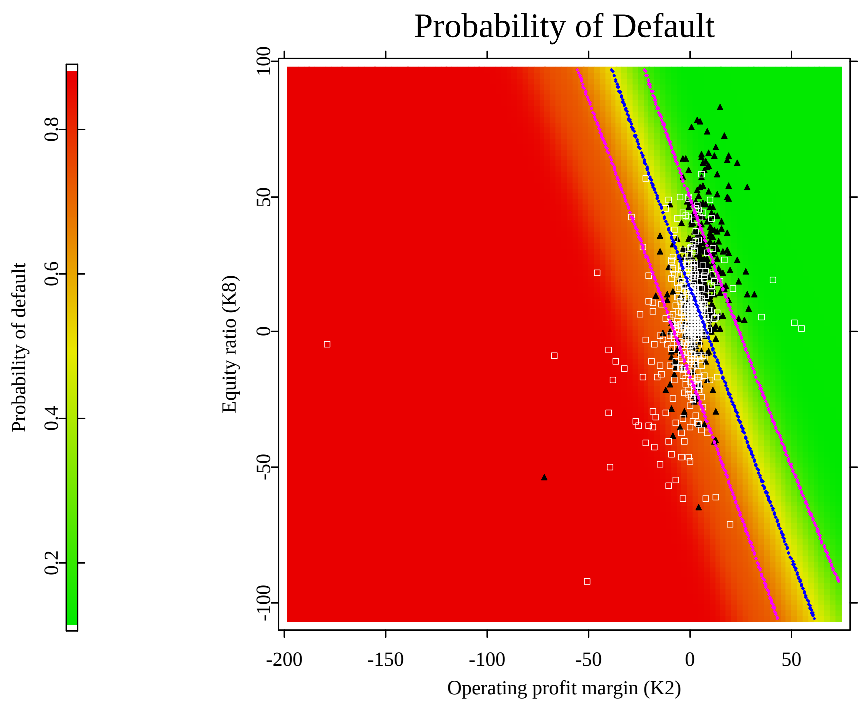
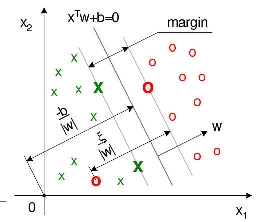


Figure 5: The case of low complexity (radial basis: $20\hat{\Sigma}^{1/2}$; capacity: $C = 8000/n$). The percentage of correctly classified out-of-sample observations: 67.4%; accuracy ratio: 0.513; α-error: 29.1%.



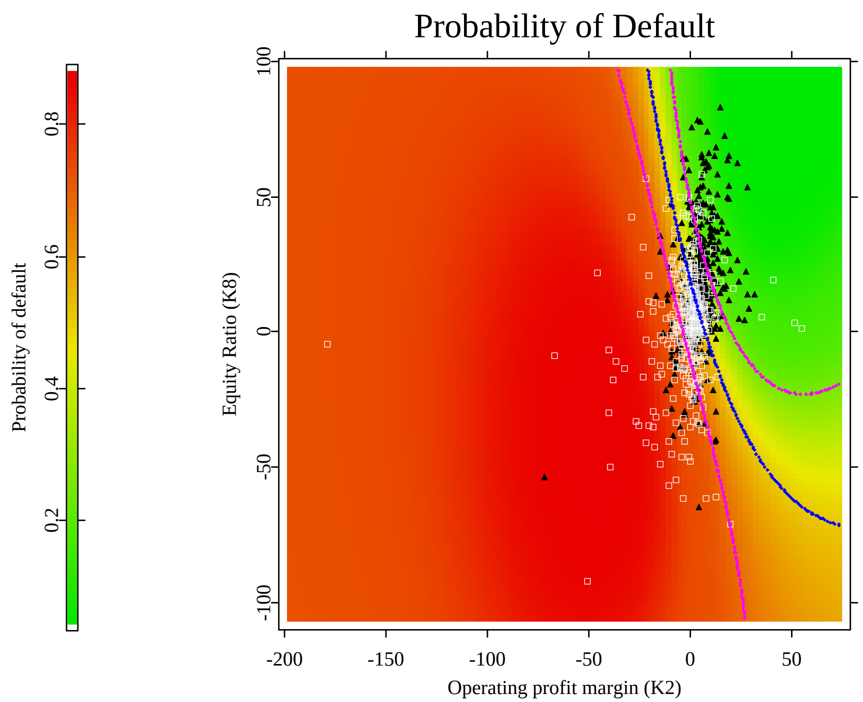
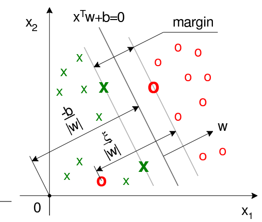


Figure 6: Medium complexity of classifier functions (radial basis: $5\hat{\Sigma}^{1/2}$; capacity: $C = 2000/n$). The percentage of correctly classified out-of-sample observations: 67.9%; accuracy ratio: 0.520; α-error: 27.6%.



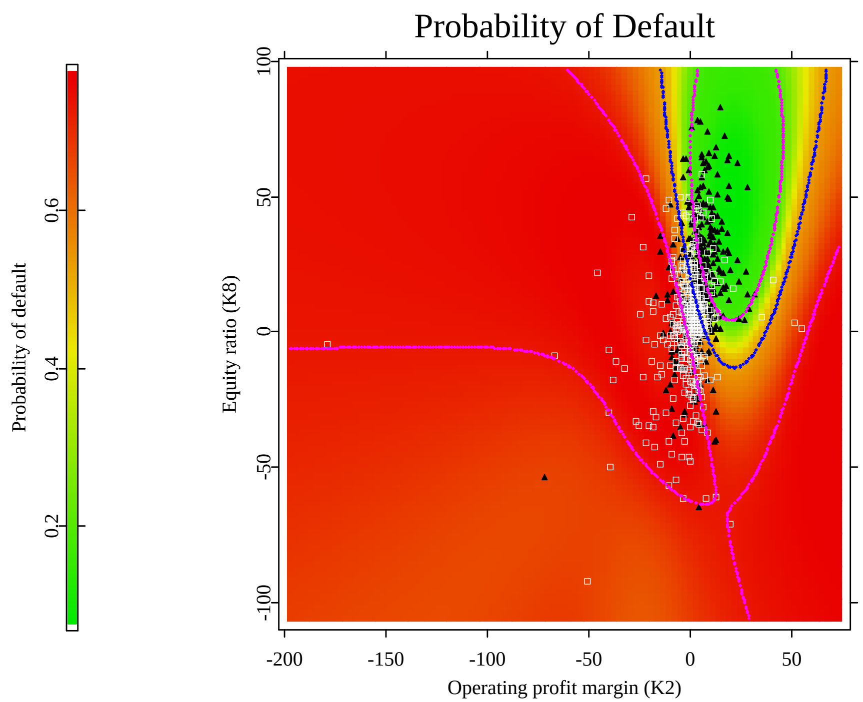
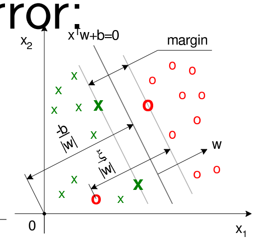


Figure 7: Optimal complexity of classifier functions (radial basis:$2\hat{\Sigma}^{1/2}$); capacity: $C = 800/n$). The percentage of correctly classified out-of-sample observations: 68.6% (maximum); accuracy ratio: 0.528; α-error: 25.8%.



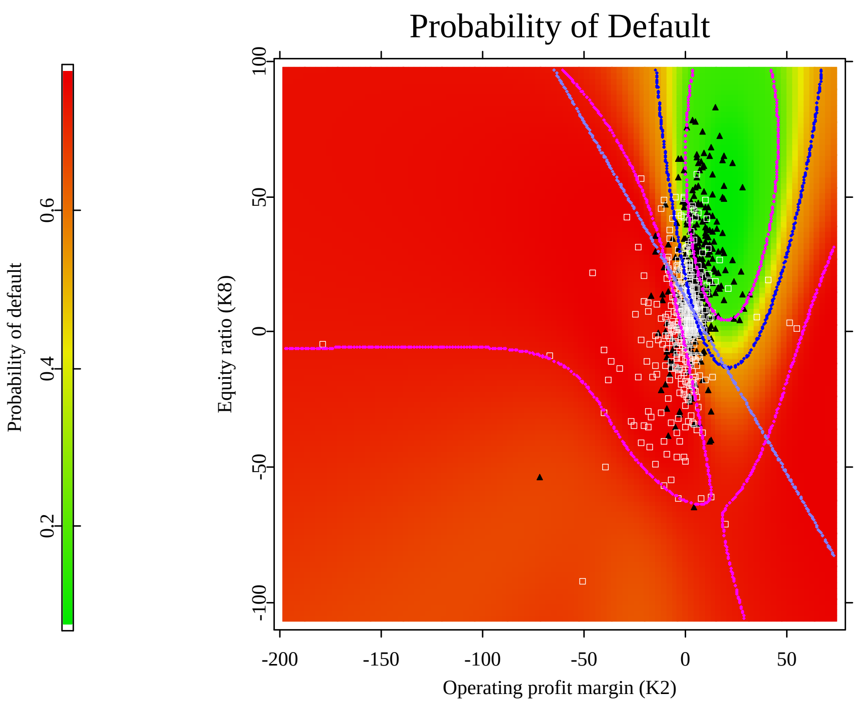
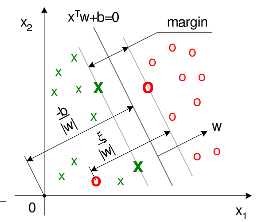


Figure 8: Comparison of an SVM with DA. (Radial basis is $2\hat{\Sigma}^{1/2}$, $C = 800/n$). The accuracy ratios are 0.528 for the SVM and 0.497 for DA.



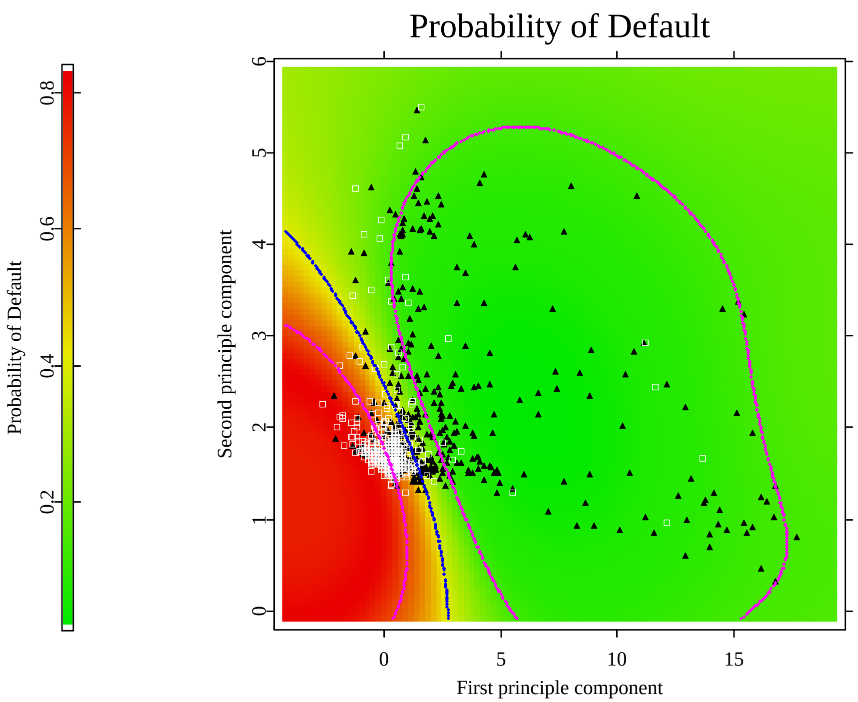


Figure 9: Two principal components. An SVM is used with the radial basis equal $2\hat{\Sigma}^{1/2}$ and $C = 800/n$. 2.5% of the smallest and 2.5% of the largest outliers were taken as equal to 2.5% and 97.5% quantiles.

Out-of-sample Classification Results 1

Six variables are selected one from each six groups except Growth such that their combination provides the highest AR as estimated with the logit model:

K4: capital recovery ratio
K7: days payable
K11: pre-tax profit margin
K19: inventories ratio
K24: net interest ratio
K25: own funds/pension provision ratio

Discriminant Analysis

		Estimated	
		Bankrupt	Non-bankrupt
Data	Bankrupt	404 (0.742)	115 (0.258)
	Non-bankrupt	152 (0.370)	367 (0.630)

Accuracy Ratio: 0.625

Logit

Applied to the same six variables

The logit's threshhold is chosen so that the number of correctly classified observations is maximized

		Estimated	
		Bankrupt	Non-bankrupt
Data	Bankrupt	408 (0.786)	111 (0.214)
	Non-bankrupt	125 (0.241)	394 (0.759)

Accuracy Ratio: 0.670

Support Vector Machine

Applied to the same six variables

The radial basis is $5\hat{\Sigma}^{1/2}$, $C = 2000/n$

		Estimated	
		Bankrupt	Non-bankrupt
Data	Bankrupt	418 (0.805)	101 (0.195)
	Non-bankrupt	138 (0.266)	381 (0.734)

Accuracy Ratio: 0.675

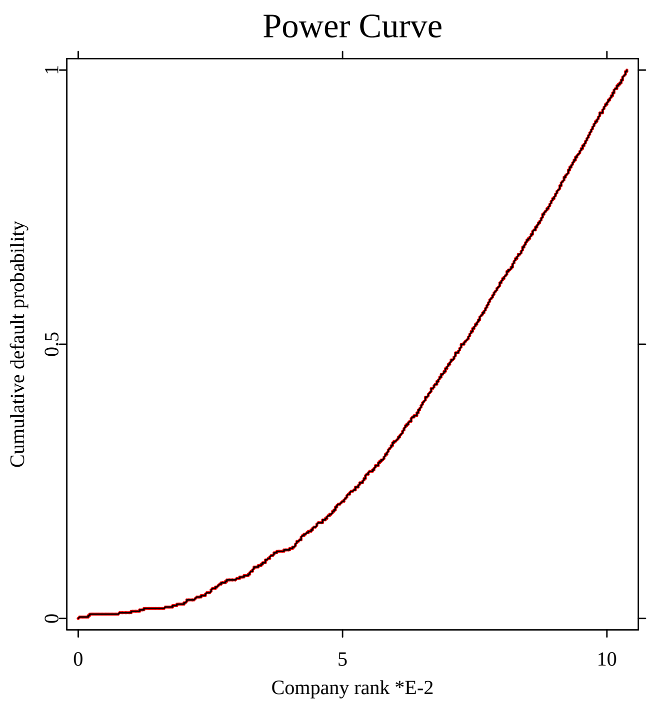
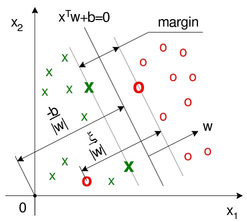


Figure 10: Power (Lorenz) curve for the SVM with six variables. (The radial basis is $5\hat{\Sigma}^{1/2}$, $C = 2000/n$). The accuracy ratio is 0.672.



Out-of-sample Classification Results 2

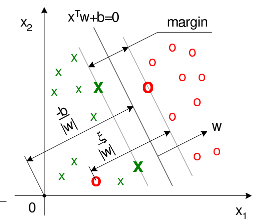
The four ratios are selected that are the basis ratios used in the Bundesbank rating system:

K1: pre-tax profit margin

K3: cash flow ratio

K6: days receivable

K9: equity ratio (adjusted)



Discriminant Analysis

		Estimated	
		Bankrupt	Non-bankrupt
Data	Bankrupt	382 (0.736)	137 (0.264)
	Non-bankrupt	155 (0.299)	364 (0.701)

Accuracy Ratio: 0.564

Logit

Estimated with the same four variables

The logit's threshhold is chosen so that the number of correctly classified observations is maximized

		Estimated	
		Bankrupt	Non-bankrupt
Data	Bankrupt	366 (0.705)	153 (0.295)
	Non-bankrupt	136 (0.262)	383 (0.738)

Accuracy Ratio: 0.587

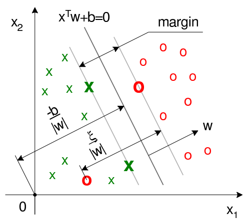
Support Vector Machine

Estimated with the same four variables

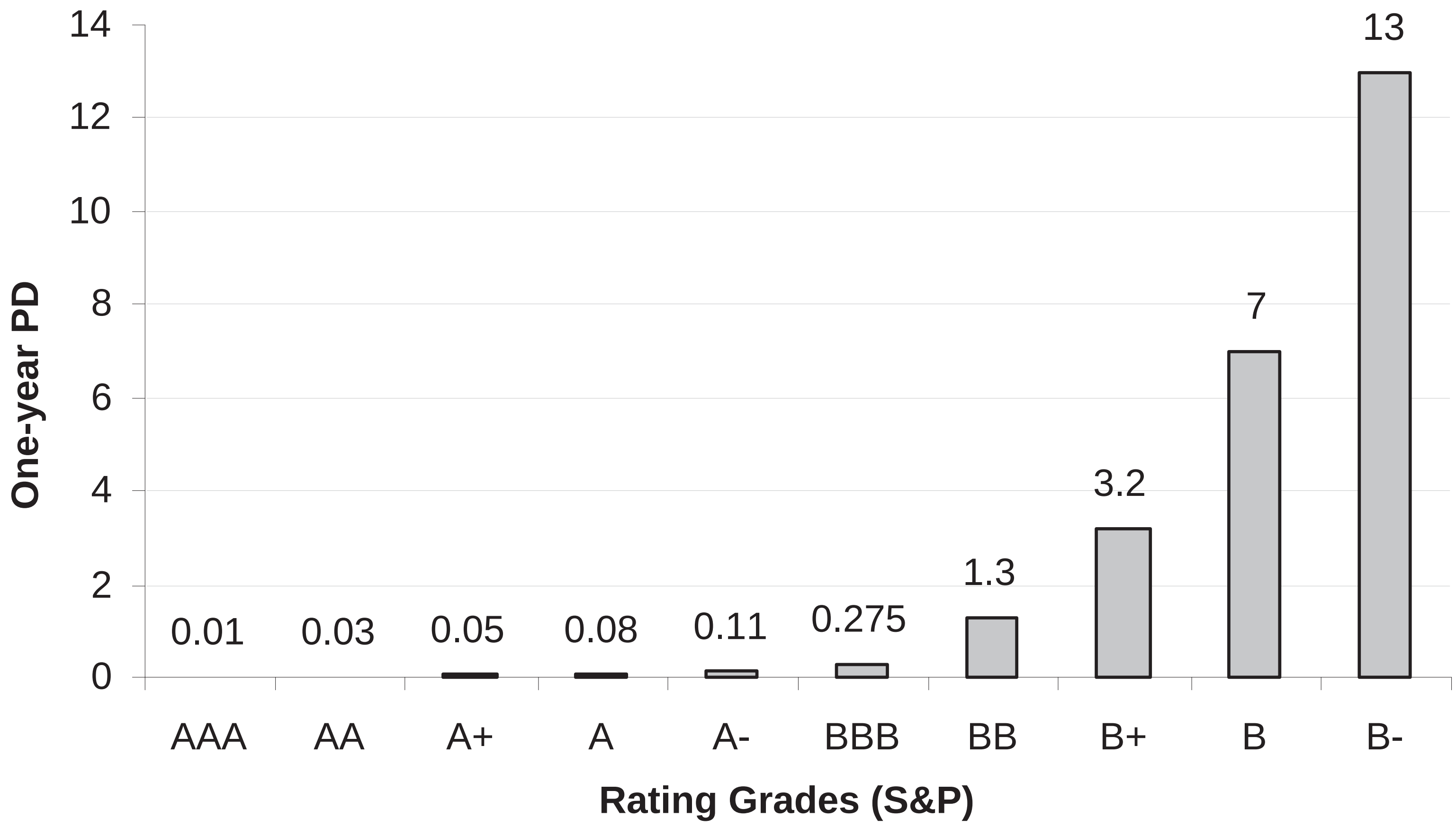
The radial basis is $2\hat{\Sigma}^{1/2}$, $C = 800/n$

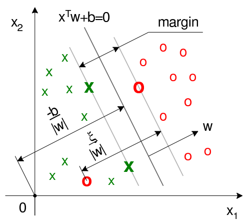
		Estimated	
		Bankrupt	Non-bankrupt
Data	Bankrupt	421 (0.811)	98 (0.189)
	Non-bankrupt	182 (0.351)	337 (0.649)

Accuracy Ratio: 0.598



Rating Grades and Probabilities of Default



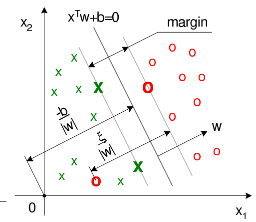


Adaption of an SVM to Company Rating

The score values $f = x^\top w + b$ estimated by an SVM correspond to default probabilities:

$$f \mapsto PD$$

- select a sliding window $f \pm \Delta f$
- count the bankrupt and all companies inside the window
- if the data is representative of the whole population, $\widehat{PD}(f) = \#_{\text{bankrupt}} / \#$
- repeat the procedure for another value of f



Estimation of PDs

- select a window around each observation $f_i \pm \Delta f$ such that it contains 41 observations and estimate the default probability for that observation. Repeat for each observation $i = 21, 22, \ldots, 2036$ ordered with respect to its score
- estimate the default probabilities for the grid points using the Nadaraya-Watson estimator with the Gaussian kernel and $\sigma = 0.2$
- plot the grid points coding the PD values with colour

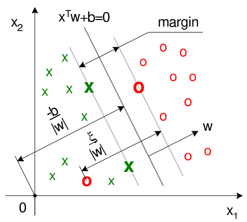
References

Altman, E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance*, September: 589-609.

Basel Committee on Banking Supervision (2003). The New Basel Capital Accord, third consultative paper, *http://www.bis.org/bcbs/cp3full.pdf*.

Beaver, W. (1966). Financial Ratios as Predictors of Failures. Empirical Research in Accounting: Selected Studies, *Journal of Accounting Research*, supplement to vol. 5: 71-111.

Falkenstein, E. (2000). *RiskCalc for Private Companies: Moody's Default Model*, Moody's Investors Service.



Füser, K. (2002). Basel II – was muß der Mittelstand tun?, *http://www.ey.com/global/download.nsf/Germany/Mittelstandsrating/$file/Mittelstandsrating.pdf*.

Härdle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*, Springer Verlag.

Merton, R. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, *The Journal of Finance*, **29**: 449-470.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, Spring: 109-131.

Platt, J.C. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, *Technical Report MSR-TR-98-14*, April.

Division of Corporate Finance of the Securities and Exchange Commission (2004). Standard industrial classification (SIC) code list, *http://www.sec.gov/info/edgar/siccodes.htm*.

Securities and Exchange Commission (2004). Archive of Historical Documents, *http://www.sec.gov/cgi-bin/srch-edgar*.

Tikhonov, A.N. and Arsenin, V.Y. (1977). *Solution of Ill-posed Problems*, W.H. Winston, Washington, DC.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer Verlag, New York, NY.