PREDICTING CORPORATE BANKRUPTCY WITH SUPPORT VECTOR MACHINES

Wolfgang HÄRDLE [2,3]

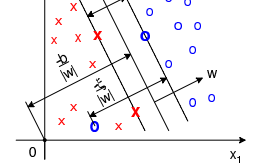
Rouslan MORO [1,2]

Dorothea SCHÄFER [1]

[1] Deutsches Institut für Wirtschaftsforschung (DIW)

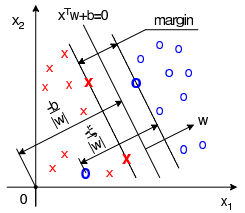
[2] Humboldt-Universität zu Berlin

[3] Center for Applied Statistics and Economics (CASE)



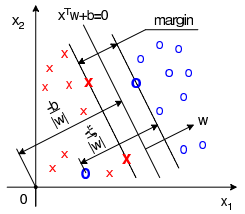
Motivation
Corporate Bankruptcy

- Does a company survive or go bankrupt within the prediction period?
- Are there any changes in the dynamics of its indicators?



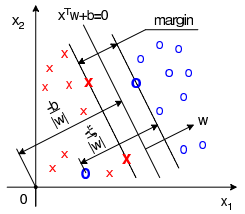
Available Information

- fundamental indicators
- option and stock trading data
- announcements
- macroeconomic indicators
- corporate governance principles
- employee profiles
- expert assessments



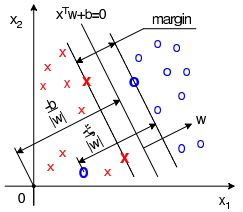
Applications

- estimating bankruptcy risks
- company bond rating (e.g. AAA, C, BB) based on the default probability
- loss given default (LGD) estimation (Basel II)
- pricing of non-traded companies (IPOs, private companies)



General Questions

- What structural changes are typical for failing companies?
- What indicators are most useful for predicting default?
- What methods should be used to extract maximum information contained in the performance indicators?
  - stock and option markets
  - expert assessment
  - statistical tools



Bankruptcy Prediction Methods

- Multivariate discriminant analysis Beaver (1966), Altman (1968)

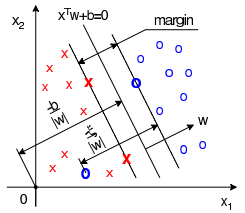
  Z-score:

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

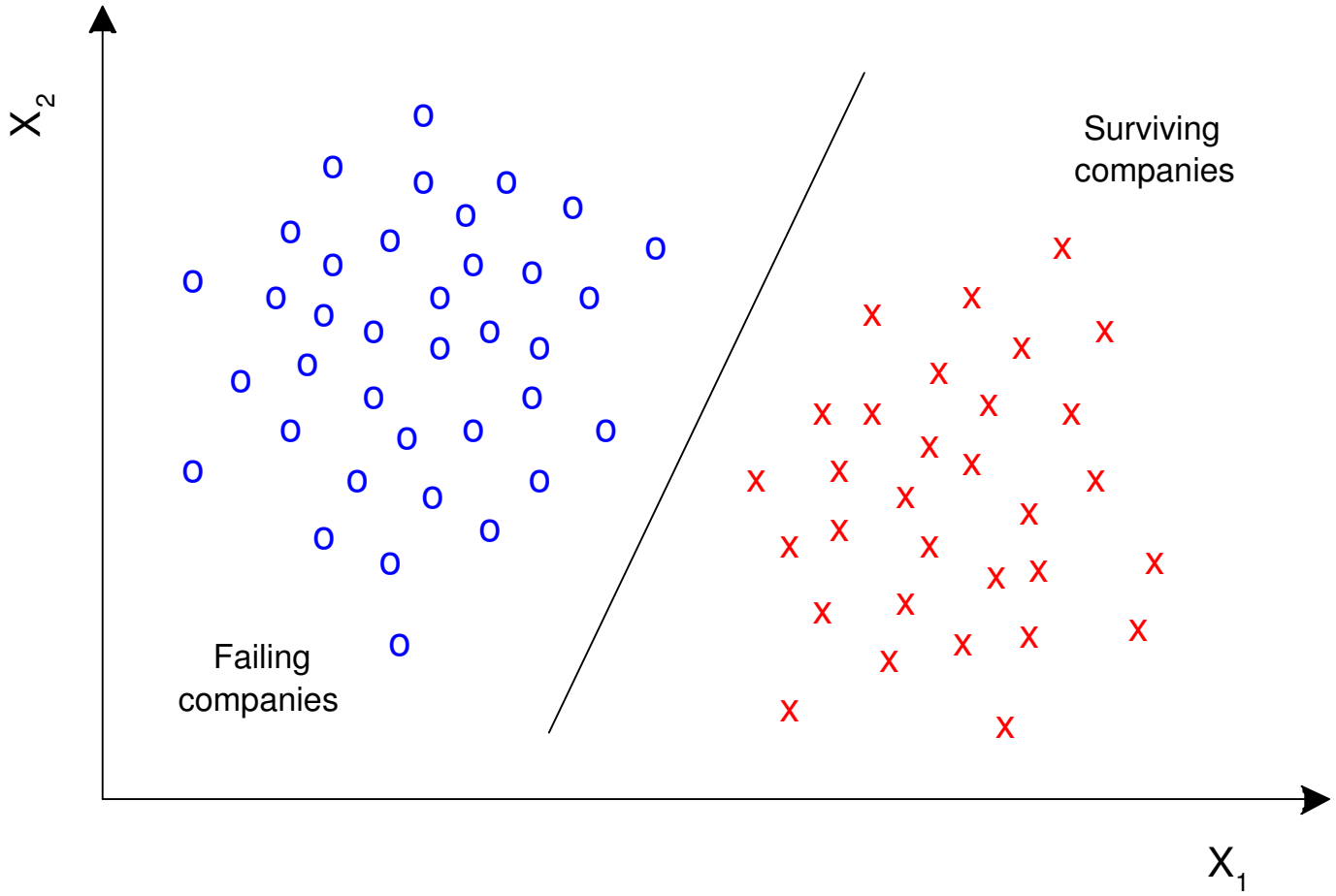
  where $x_i = (x_{i1}, ..., x_{id})^\top \in \mathbb{R}^d$ are financial ratios for the i-th company.

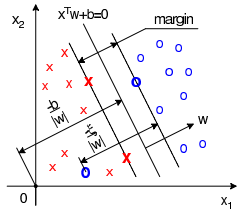
  successful company: $Z_i \geq z$

  failure: $Z_i < z$

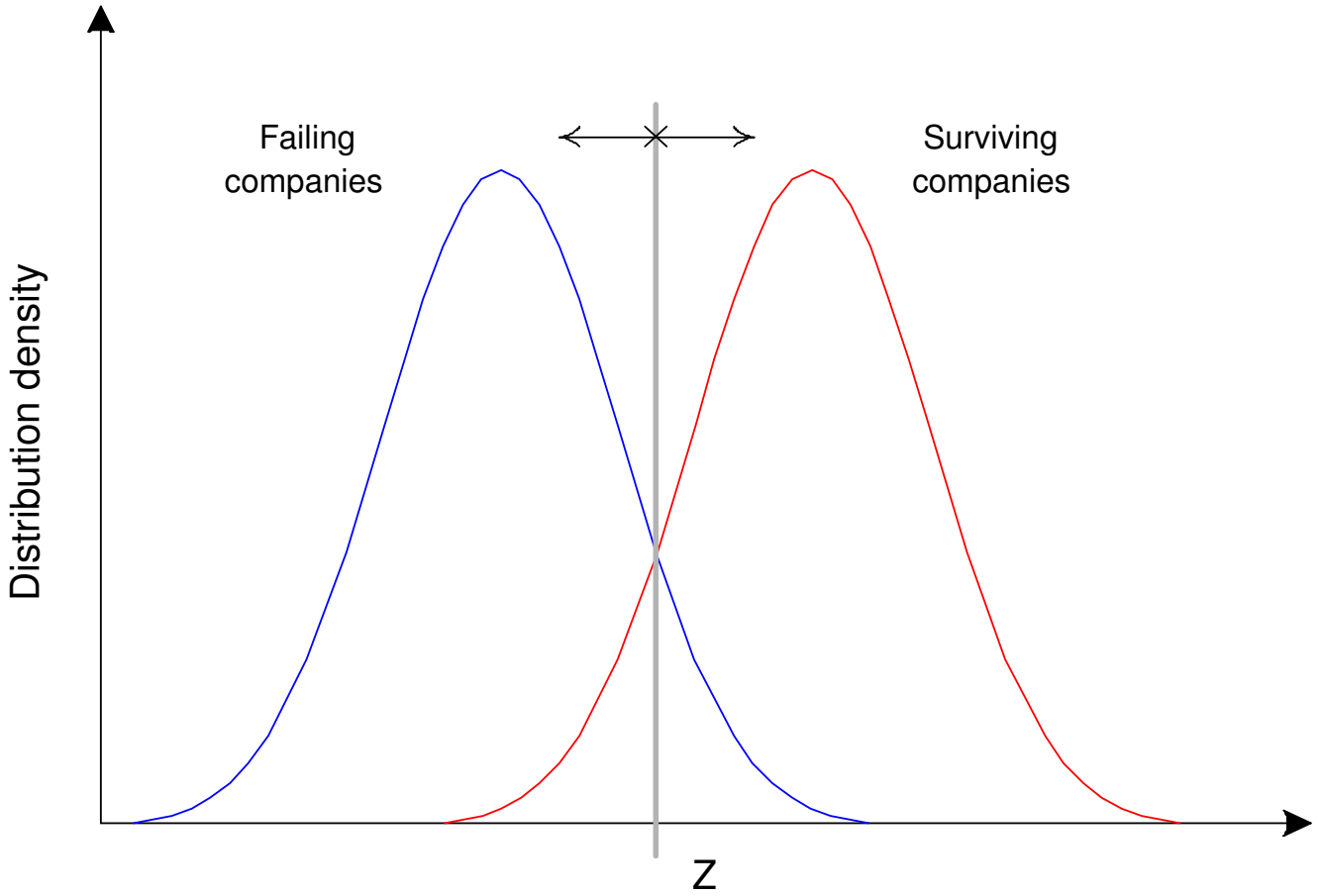


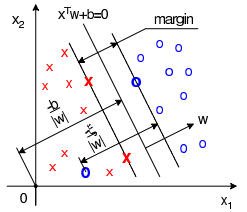
Linear Discriminant Analysis





Linear Discriminant Analysis





Bankruptcy Prediction Methods (cont.)

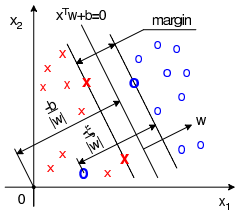
- Probit model Martin (1977), Ohlson (1980)

$$\mathrm{E}[y_i|x_i] = \Phi\left(a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_d x_{id}\right), \quad y_i = \{0, 1\}$$

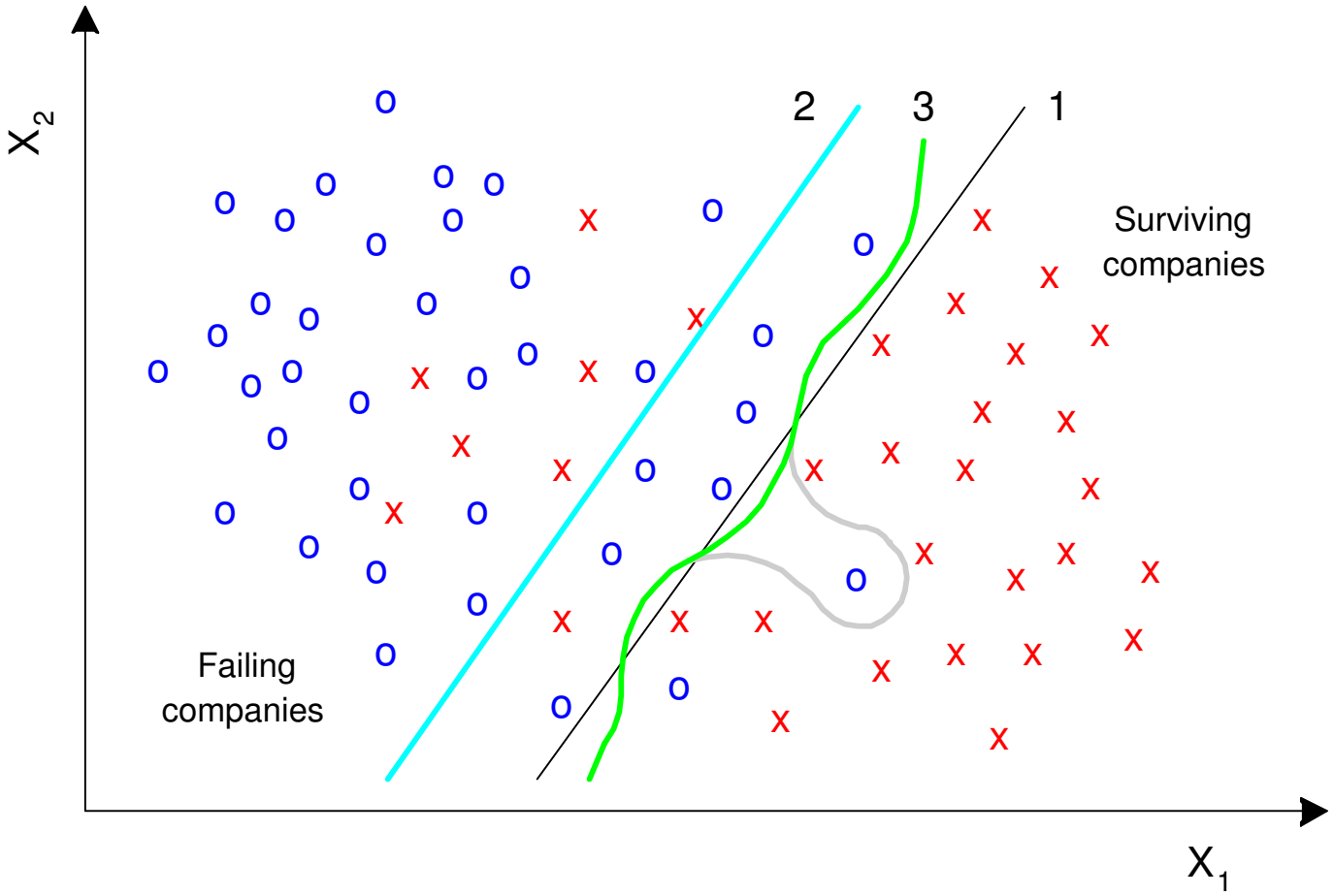
- Logit model

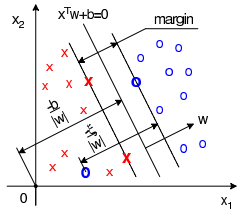
$$\mathrm{E}[y_i|x_i] = \frac{\exp(a_0 + a_1 x_{i1} + ... + a_d x_{id})}{1 + \exp(a_0 + a_1 x_{i1} + ... + a_d x_{id})}$$

- Gambler's ruin Wilcox (1971)
- Recursive partitioning Frydman et al. (1985)
- Neural networks (1990's)



Linearly Non-separable Classification Problem

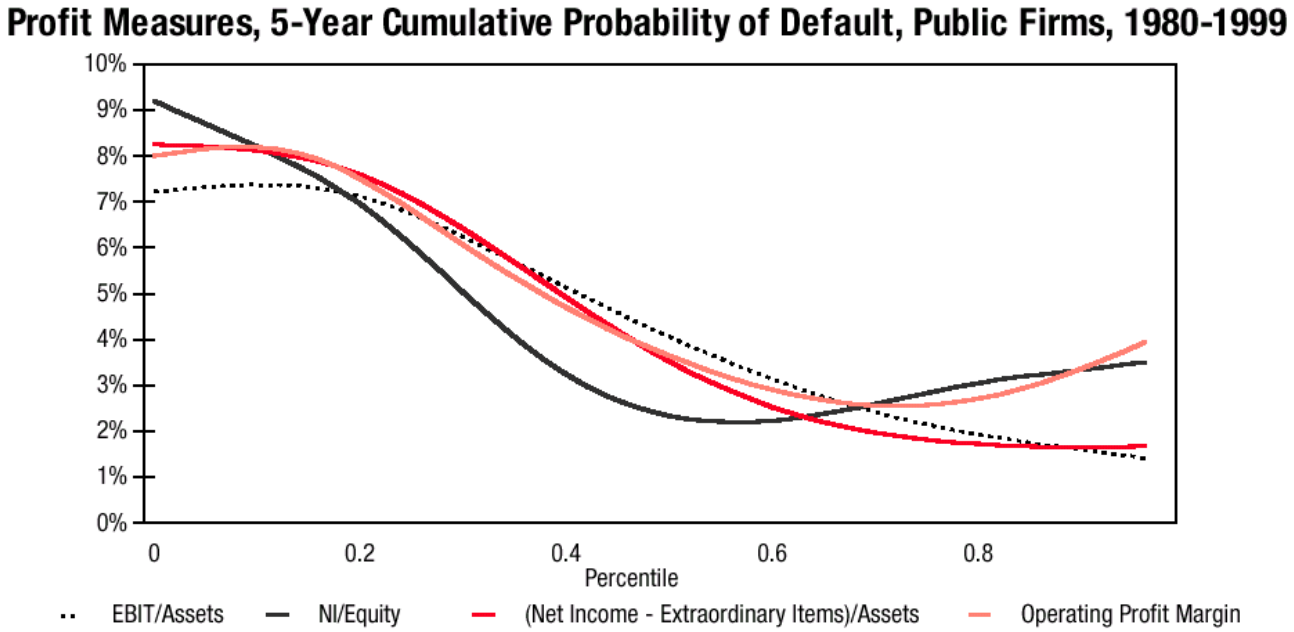


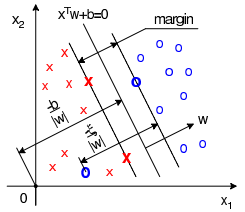


The Benchmark Moody's Model

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

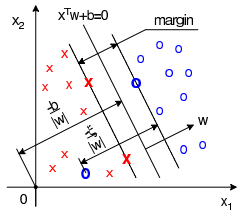
f_j are estimated non-parametrically on univariate models





Outline of the Talk

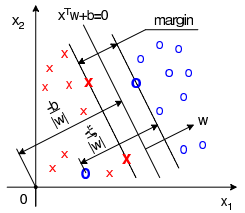
1. ✓ Motivation
2. Support Vector Machines and Their Optimal Properties
3. Expected Risk vs. Empirical Risk Minimization
4. Realization of SVMs
5. Non-linear Case
6. Company Classification with SVMs



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
- Optimality of SVMs is given by the statistical learning theory
- SVMs do not rely on asymptotic properties; they are optimal for small samples, i.e. in most practically significant cases
- SVMs always give a unique solution



Classification Problem

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

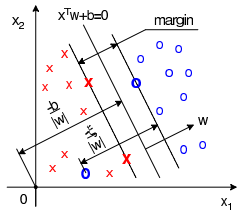
Find the class y of a new object x using the classifier $f : \mathcal{X} \mapsto \{\pm 1\}$, such that the expected risk $R(f)$ is minimal.

x_i 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.

Regression Problem

Setup as for the classification problem but: $y \in \mathbb{R}$



Expected Risk Minimization

If $P(x, y)$ is known, then the expected risk

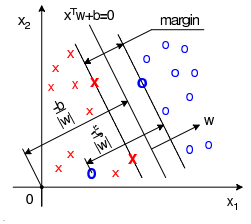
$$R(f) = \int \frac{1}{2}|f(x) - y| dP(x, y) \quad = \mathrm{E}_{P(x,y)}[L]$$

can be minimized directly over $P(x, y)$

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

$$\begin{aligned} \text{The loss } L = \tfrac{1}{2}|f(x) - y| \quad &= \quad 0 \text{ if classification is correct} \\ &= \quad 1 \text{ if classification is wrong} \end{aligned}$$

$\mathcal{F}$ is the set of (non)linear classifier functions



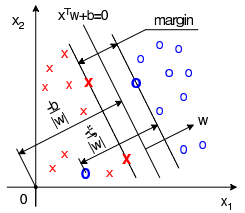
Empirical Risk Minimization

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

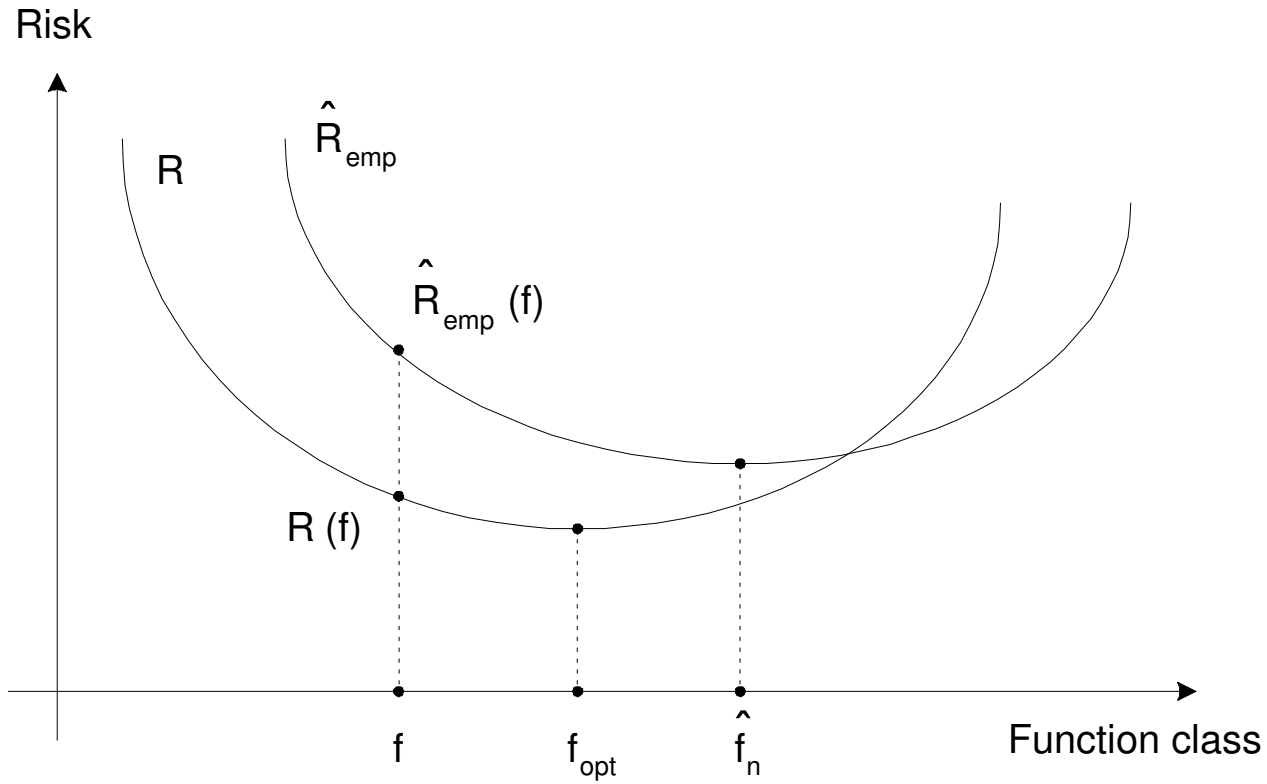
$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} |f(x_i) - y_i|$$

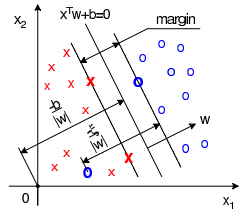
Minimization (ERM) over the training set $\{(x_i, y_i)\}_{i=1}^{n}$

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



Empirical Risk vs. Expected Risk





Convergence

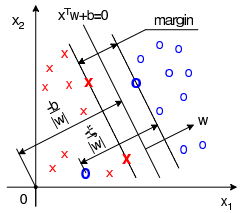
From the law of large numbers

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

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

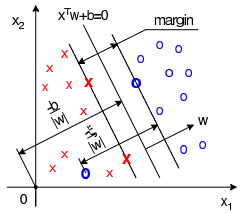
This is a basic result of the Statistical Learning Theory that already started in the 1960s:

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

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

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

Minimize VC bound – **Structural Risk Minimization** – search for the optimal model structure described by $\mathcal{S}_h \subseteq \mathcal{F}$; $f \in \mathcal{S}_h$.

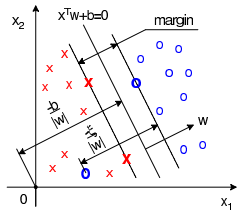


Vapnik-Chervonenkis (VC) Dimension

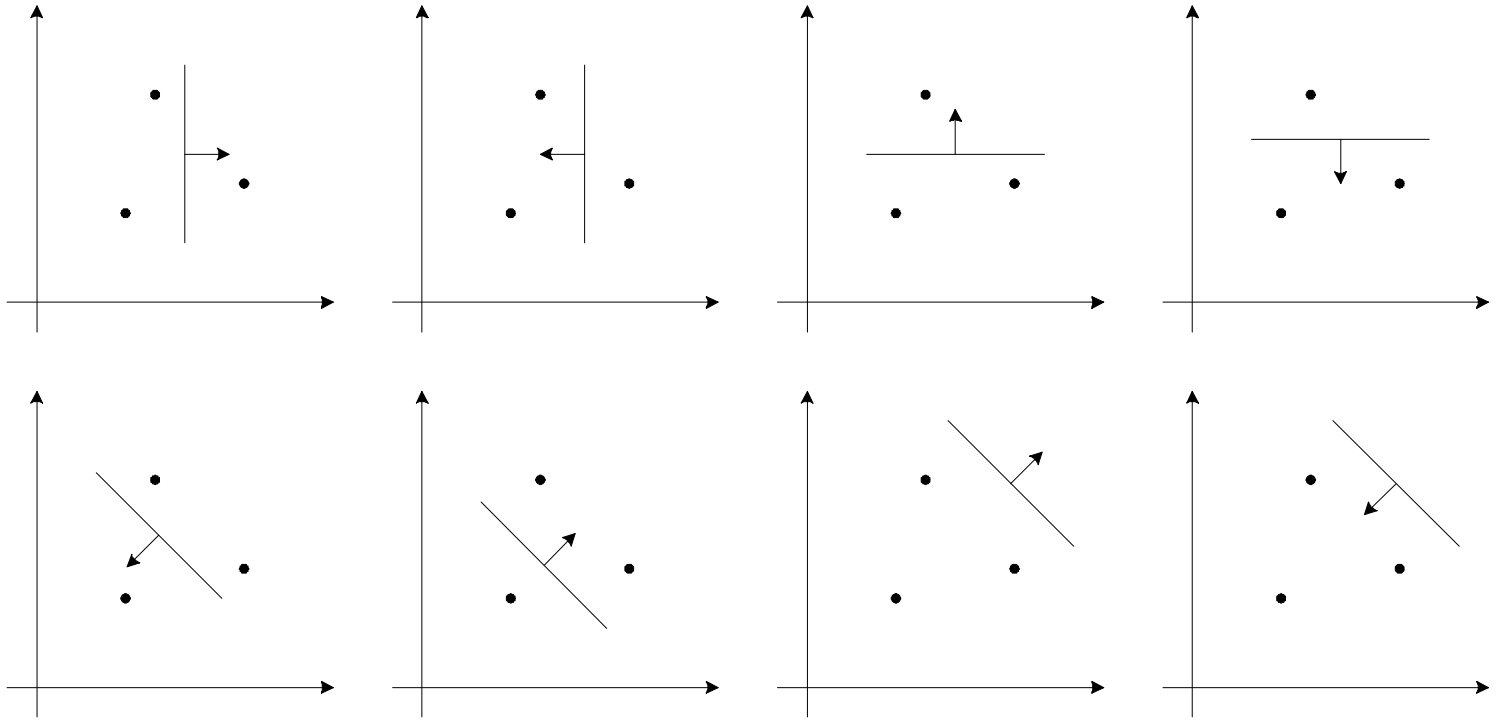
Definition. h is VC dimension of a set of functions if there exists a set of points $\{x_i\}_{i=1}^{h}$ such that these points can be separated in all 2^h possible configurations, and no set $\{x_i\}_{i=1}^{q}$ exists where $q > h$ satisfies this property.

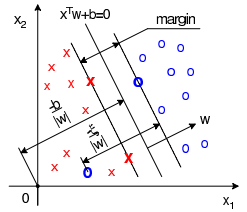
Example 1. The functions $A \sin \theta x$ has an infinite VC dimension.

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 $2^h = 2^4 = 16$ ways). The VC dimension equals $h = 3$.



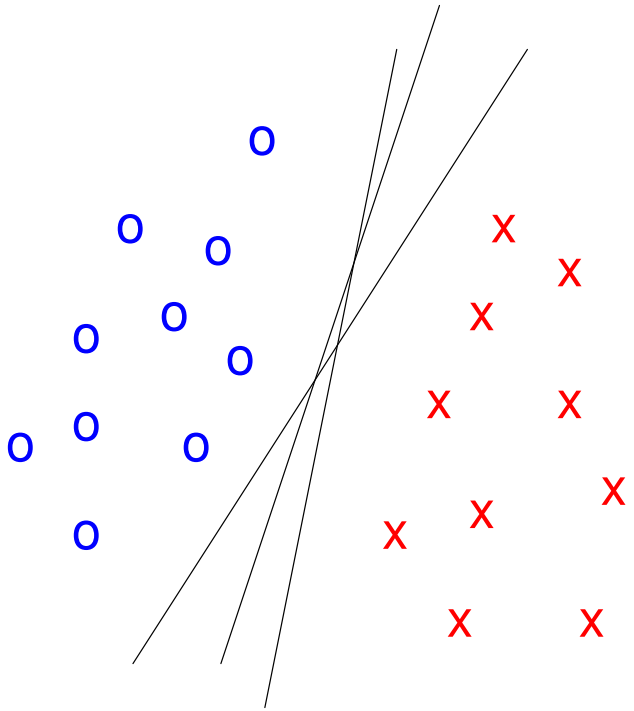
VC Dimension (cont.)

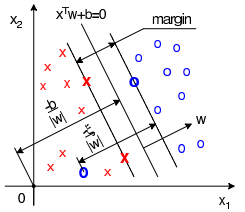




Linear SVMs. Separable Case

The training set: $\{(x_i, y_i)\}_{i=1}^{n}$, $y_i = \{\pm 1\}$, $x_i \in \mathbb{R}^d$. Find the classifier with the highest margin.





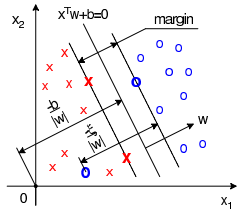
Let $x^\top w + b = 0$ be a separating hyperplane. Then d_+ (d_-) will be the shortest distance to the closest objects from the classes $+1$ (-1).

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

combine them into one constraint

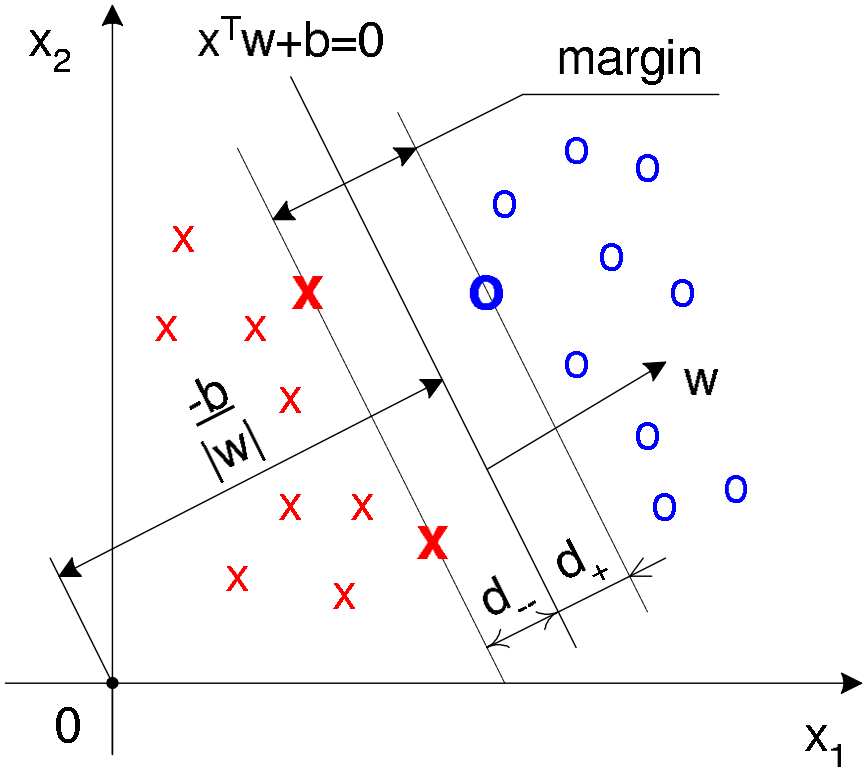
$$y_i(x_i^\top w + b) - 1 \geq 0 \qquad \forall i \tag{1}$$

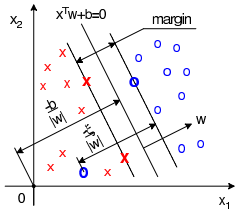
The canonical hyperplanes $x_i^\top w + b = \pm 1$ are parallel and the distance between each of the them and the separating hyperplane is $d_\pm = 1/\|w\|$.



Linear SVMs. Separable Case

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





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

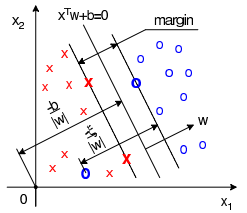
$\frac{\partial L_P}{\partial w_k} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i x_{ik} = 0 \qquad k = 1, ..., d$

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

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

$\alpha_i \geq 0$

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



Substitute the KKT conditions into L 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

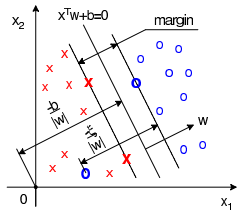
$$\min_{w_k, b} \max_{\alpha_i} L_P$$

$$\max_{\alpha_i} L_D$$

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 classifier function is:

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

where

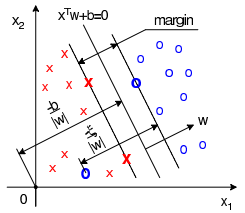
$w = \sum_{i=1}^{n} \alpha_i y_i x_i$

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

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

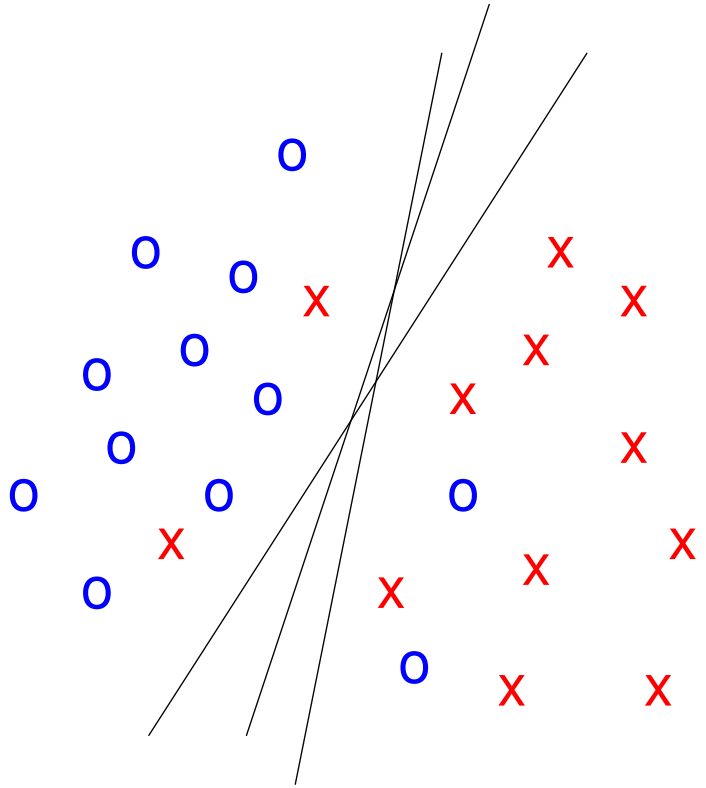
$$\alpha_i = \arg\max_{\alpha_i} L_D$$

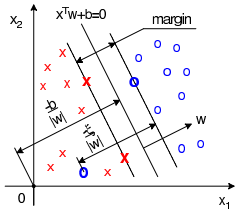
subject to the constraints.



Linear SVMs. Non-separable Case

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





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

$$x_i^\top w + b \geq 1 - \xi_i \qquad for \quad y_i = 1$$

$$x_i^\top w + b \leq -1 + \xi_i \qquad for \quad y_i = -1$$

$$\xi_i \geq 0 \qquad \forall i$$

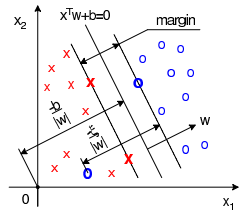
If $\xi_i > 1$, an error occurs. The objective function in this case is

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

where ν is a positive integer controlling sensitivity to outliers

C controls the generalization power

Under such a formulation the problem is convex.



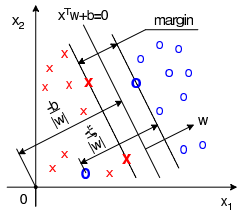
The Lagrangian Formulation

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

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

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

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

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

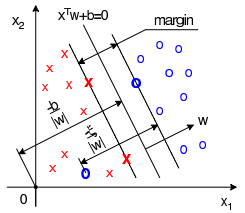
$$\xi_i \geq 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0$$

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

$$\mu_i \xi_i = 0$$



For $\nu = 1$ the dual Lagrangian will not contain ξ_i or their Lagrange multipliers

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

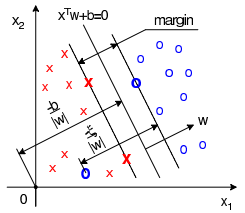
The dual problem is

$$\max_{\alpha_i} L_D$$

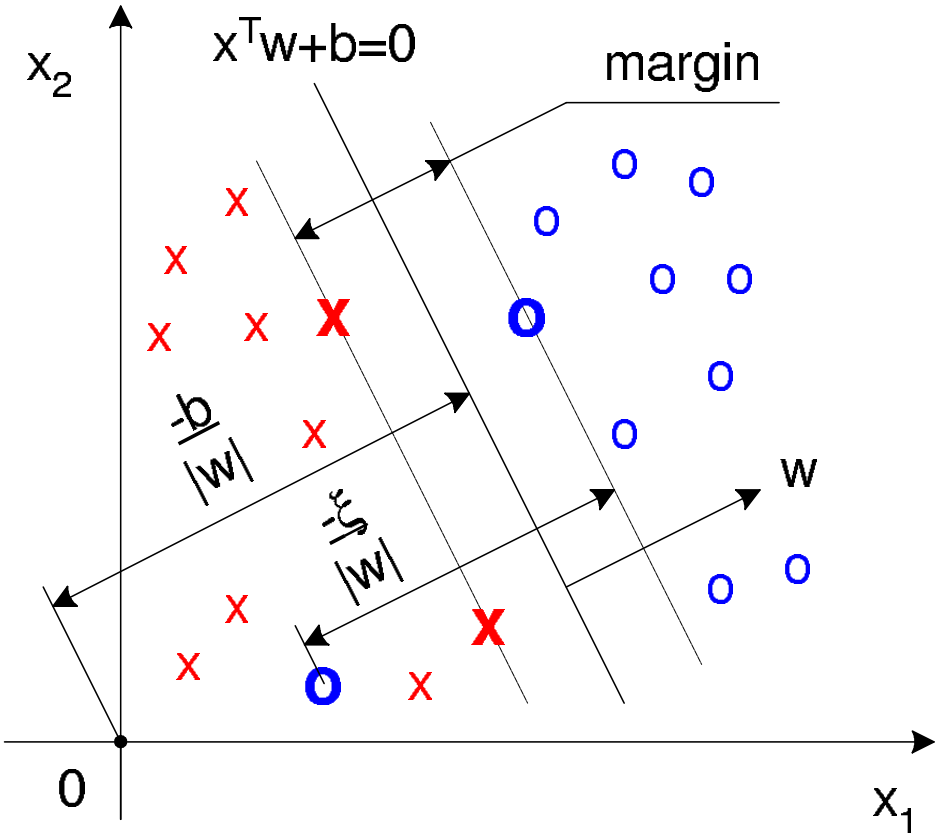
subject to

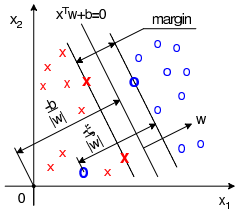
$$0 \leq \alpha_i \leq C$$

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



Linear SVM. Non-separable Case





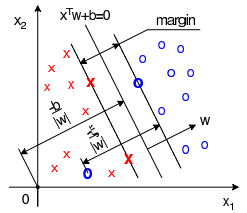
Non-linear SVMs

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

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

Notice, 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 into $\Psi(x_i)^\top \Psi(x_j)$.

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.



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 data set $x_1, ..., x_n$ and any real numbers $\lambda_1, ..., \lambda_n$ the function K must satisfy

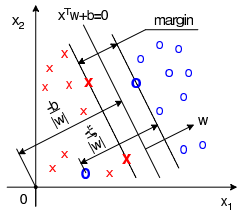
$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j K(x_i, x_j) \geq 0$$

Some examples of kernel functions:

$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2}$ – Gaussian kernel

$K(x_i, x_j) = (x_i^\top x_j + 1)^p$ – polynomial kernel

$K(x_i, x_j) = \tanh(k x_i^\top x_j - \delta)$ – hyperbolic tangent kernel



Classes of Kernels

A **stationary** kernel is the kernel which is translation invariant

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

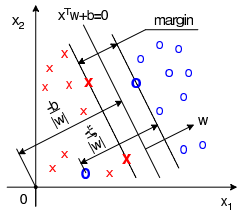
An **isotropic** (homogeneous) kernel is one which depends only on the norm of the lag vector (distance) between two data points

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

A **local stationary** kernel is the 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.

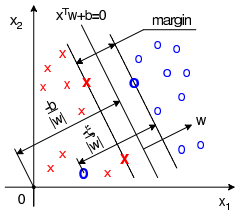


Matérn kernel

$$\frac{K_I(\|x_i - x_j\|)}{K_I(0)} = \frac{1}{2^{\nu-1}\Gamma(\nu)}(\frac{2\sqrt{\nu}\|x_i - x_j\|}{\theta})^{\nu} H_{\nu}(\frac{2\sqrt{\nu}\|x_i - x_j\|}{\theta})$$

where Γ is the Gamma function and H_{ν} is the modified Bessel function of the second kind of order ν.

The parameter ν allows to control the smoothness. The Matérn kernel reduces to the Gaussian kernel for $\nu \to \infty$.

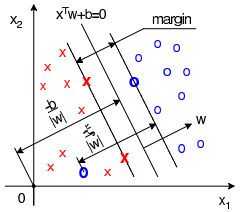


Company Analysis

Source: annual reports of the companies from 1998-1999 available through the Securities and Exchange Commision (SEC) (www.sec.gov)

- n=84.
- 42 companies survived and 42 companies went bankrupt by 2001-2002.

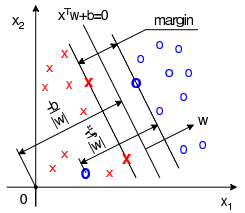
The failing and surviving companies were matched in size and industry. The bankruptcy was declared by filing Chapter 11 of the Bankruptcy Code.



The companies were characterized by 14 variables from which the following financial ratios were calculated:

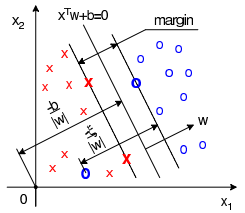
1. Profit measures: EBIT/TA, NI/TA, EBIT/Sales;
2. Leverage ratios: EBIT/Interest, TD/TA, TL/TA;
3. Liquidity ratios: QA/CL, Cash/TA, WC/TA, CA/CL, STD/TD.
4. Activity or turnover ratios: Sales/TA, Inventories/COGS.

The average capitalization of a company: \$8.12 bln. $d = 14$



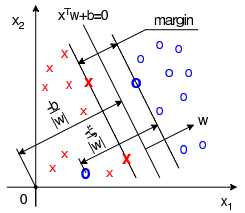
Cluster Analysis of the Companies

	Operating	Bankrupt
EBIT/TA	0.263	0.015
NI/TA	0.078	-0.027
EBIT/Sales	0.313	-0.040
EBIT/INT	13.223	1.012
TD/TA	0.200	0.379
TL/TA	0.549	0.752
SIZE	15.104	15.059



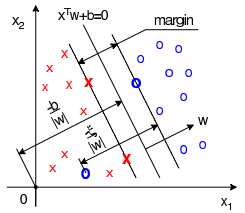
	Operating	Bankrupt
QA/CL	1.108	1.361
CASH/TA	0.047	0.030
WC/TA	0.126	0.083
CA/CL	1.879	1.813
STD/TD	0.144	0.061
Sales/TA	1.178	0.959
INV/COGS	0.173	0.155

There are 19 members in the cluster of survived companies and 65 members in cluster of failed companies. The result significantly changes in the presence of outliers.



Company Classification with SVMs.

The Influence of Different Classifier Complexities



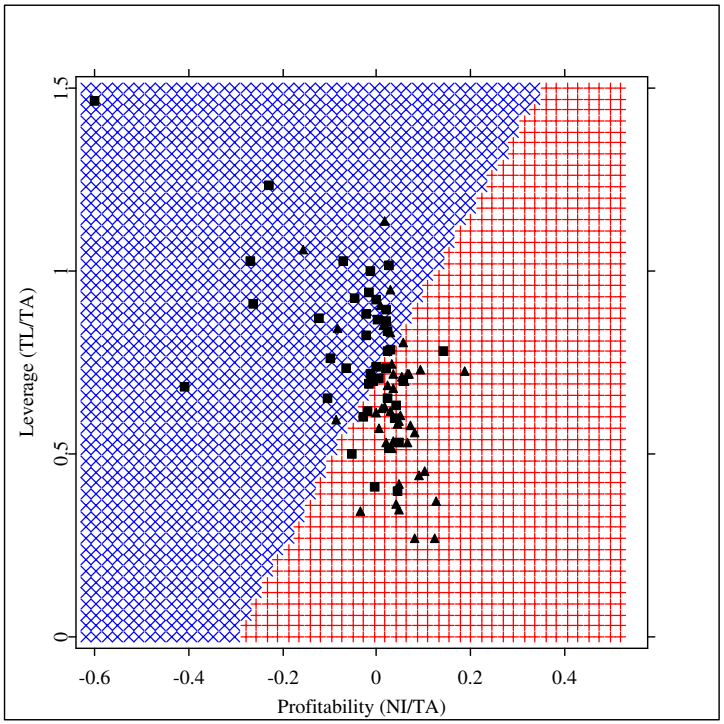
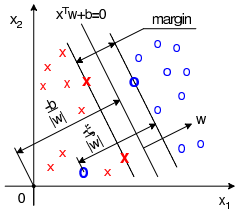


Figure 1: The case of a low complexity of classification functions (near linear functions are used, the radial basis is $20\Sigma^{1/2}$). The generalization ability is fixed ($C = 1$).



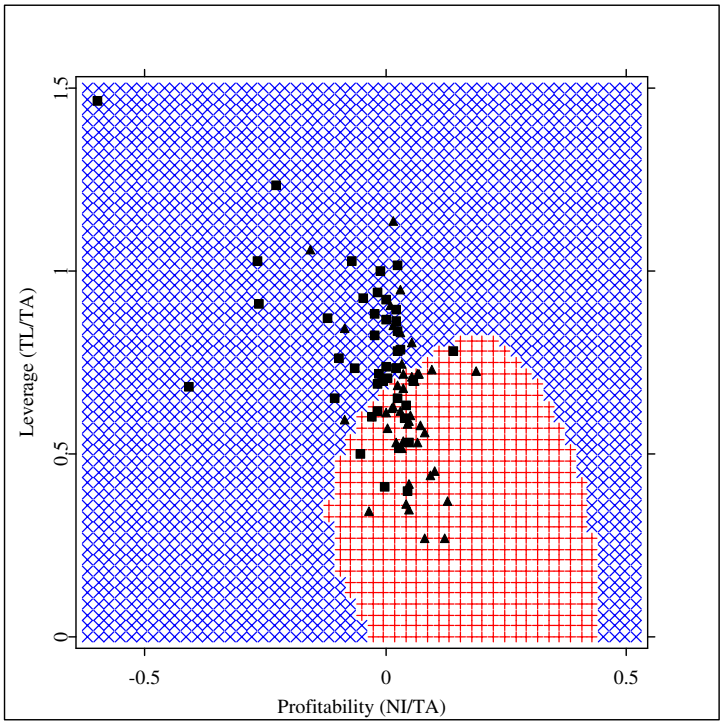
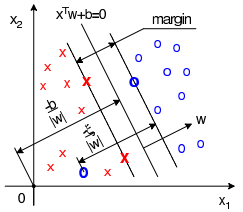


Figure 2: The case of an average complexity of the classifier functions (the radial basis is $2\Sigma^{1/2}$). The generalization ability is fixed ($C = 1$)



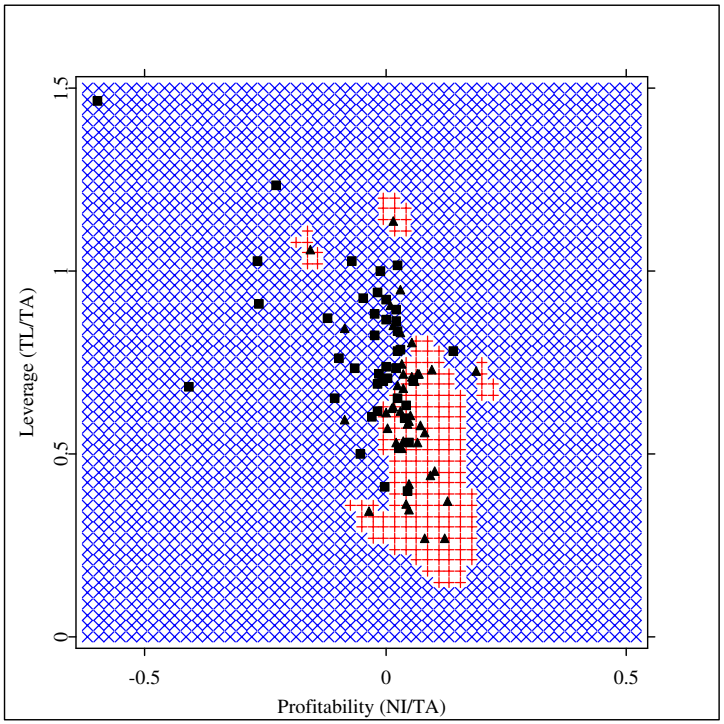
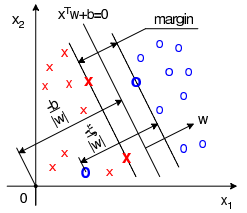
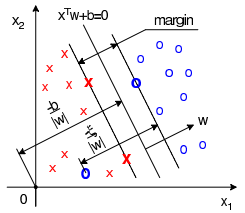


Figure 3: The case of a highly complex classification functions (the radial basis is $0.53\Sigma^{1/2}$). The generalization ability is fixed ($C = 1$)



Company Classification with SVMs.

The Influence of Different Generalization Abilities



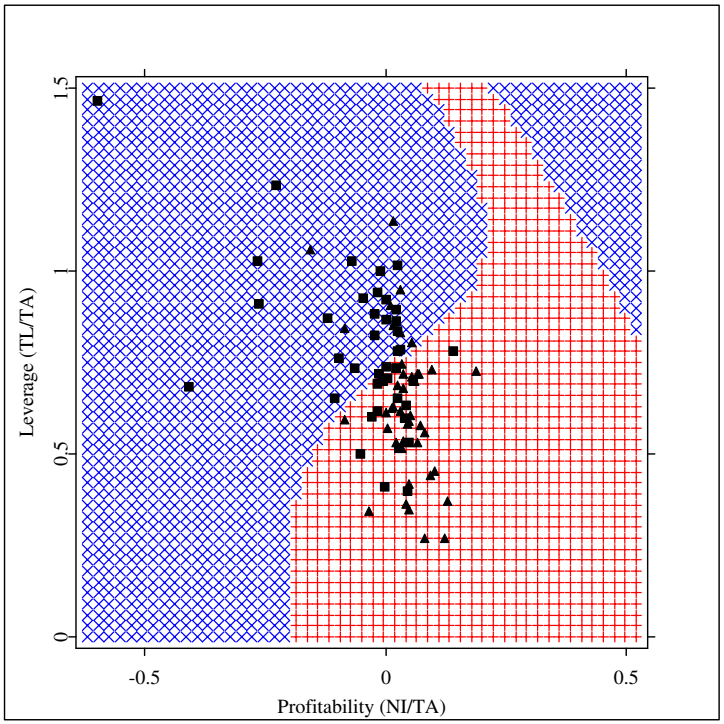
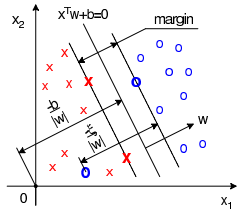


Figure 4: The case of a very high generalization ability ($C = 0.01$). The radial basis is fixed at $2\Sigma^{1/2}$



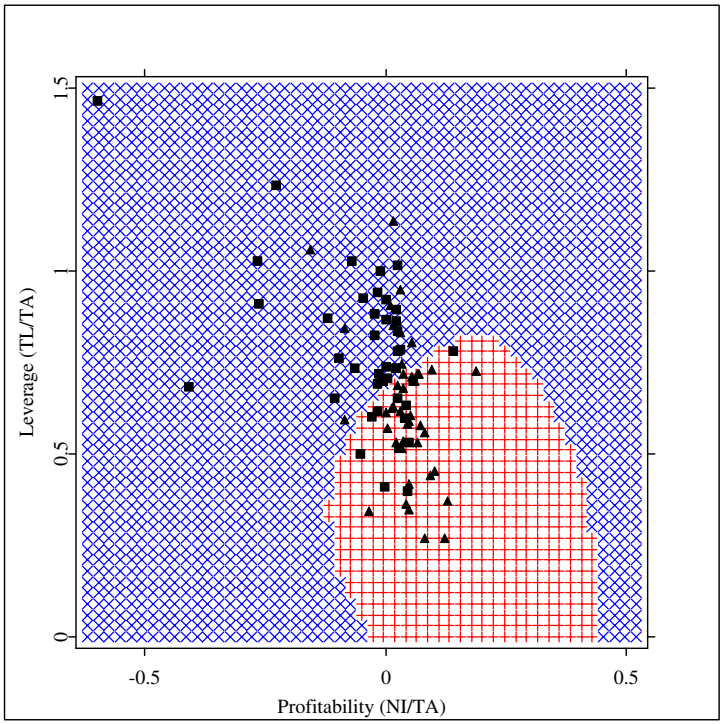
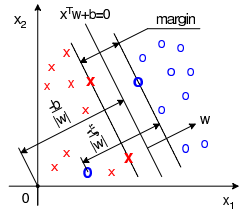


Figure 5: The case of an average generalization ability ($C = 1$). The radial basis is fixed at $2\Sigma^{1/2}$



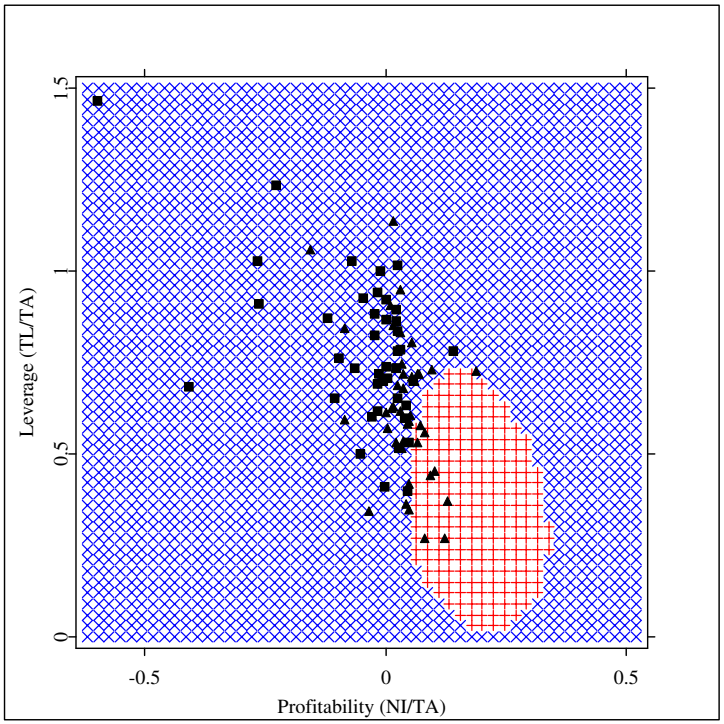


Figure 6: The case of low generalization ability ($C = 100$). The radial basis is fixed at $2\Sigma^{1/2}$