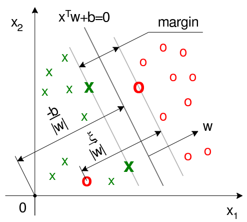
SURVIVAL ANALYSIS WITH SUPPORT VECTOR MACHINES

Wolfgang HÄRDLE

Rouslan MORO

Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin



Applications in Medicine

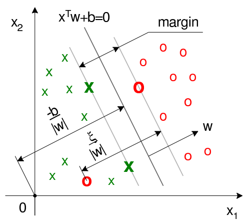
- estimation of survival chances
- classification of patients with respect to their sensitivity to treatment
- reproduction of test results without using invasive methods

Other Applications

- company rating based on survival probability
- insurance

General Approach

- estimate the probability of death in period t given that the patient has survived up to period $t-1$
- What statistical methods are suitable?



Standard Methodology

- Cox proportional hazard regression (1972)

A semi-parametric model based on a generalised linear model

$$\ln h_i(t) = a(t) + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_d x_{id}$$

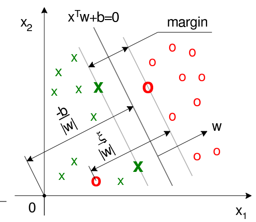
or explicitly for the *hazard* $h_i(t)$

$$h_i(t) = h_0(t) \exp(b_1 x_{i1} + b_2 x_{i2} + \ldots + b_d x_{id})$$

The hazard ratio for any two observations is independent of time t:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) e^{\eta_i}}{h_0(t) e^{\eta_j}} = \frac{e^{\eta_i}}{e^{\eta_j}}$$

where $\eta_i = b_1 x_{i1} + b_2 x_{i2} + \ldots + b_d x_{id}$

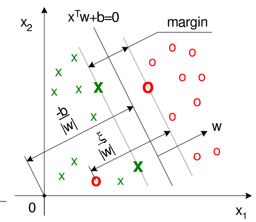


Proposed Methodology

- at time t break all surviving patients into two groups:
  1. those who will die in period $t + 1$
  2. the rest patients who will survive in period $t + 1$
- train a classification machine on these two groups
- repeat the procedure for all $t \in \{0, 1, ..., T - 1\}$

Alltogether we will get T differently trained classification machines

What *classification* method to apply?



Multivariate Discriminant Analysis

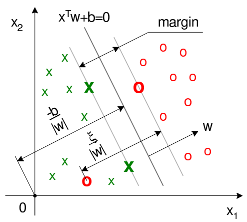
- Fisher (1931)

The score:

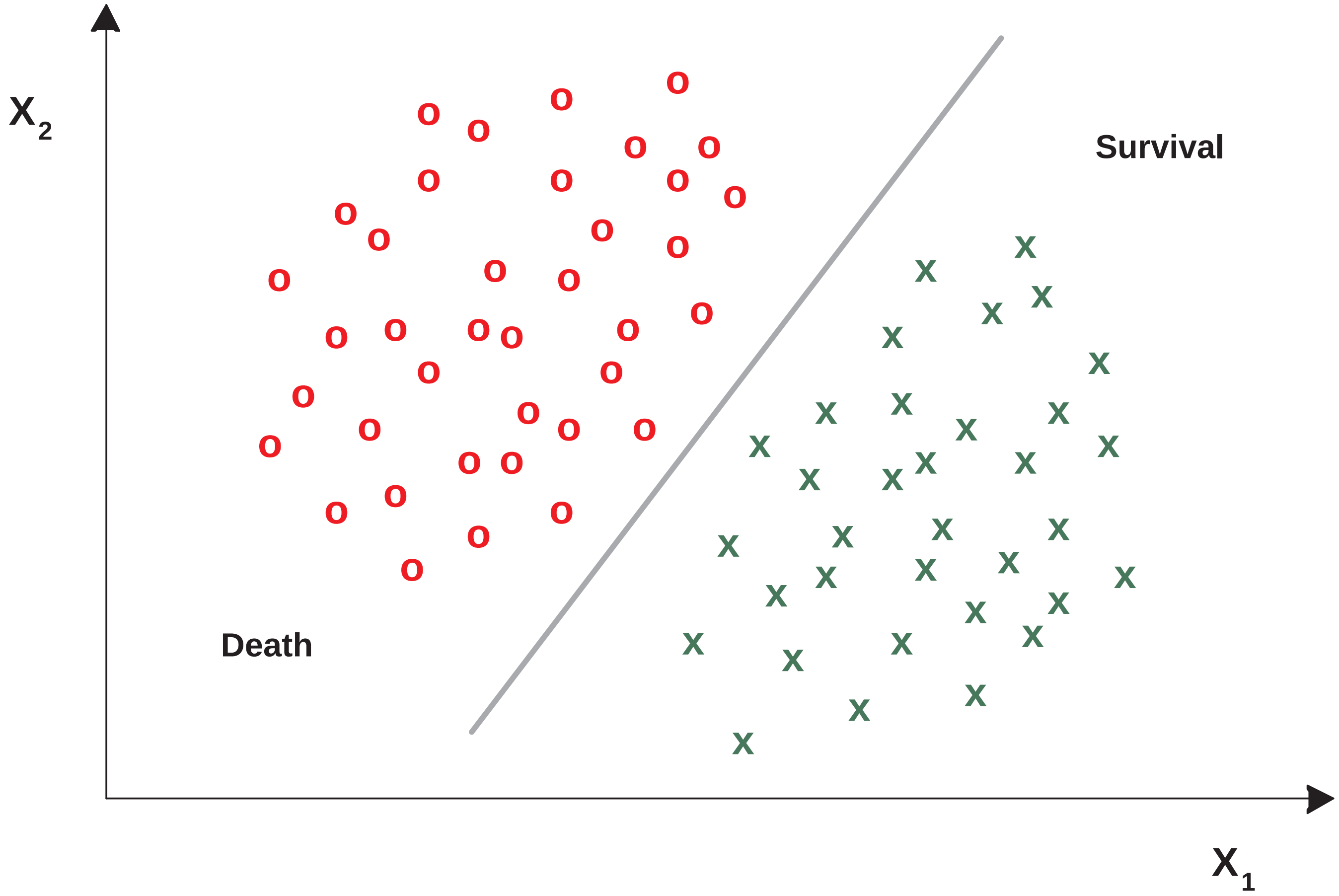
$$S_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_d x_{id} = a^\top x_i$$

x_i are screening and test results for the i-th patient

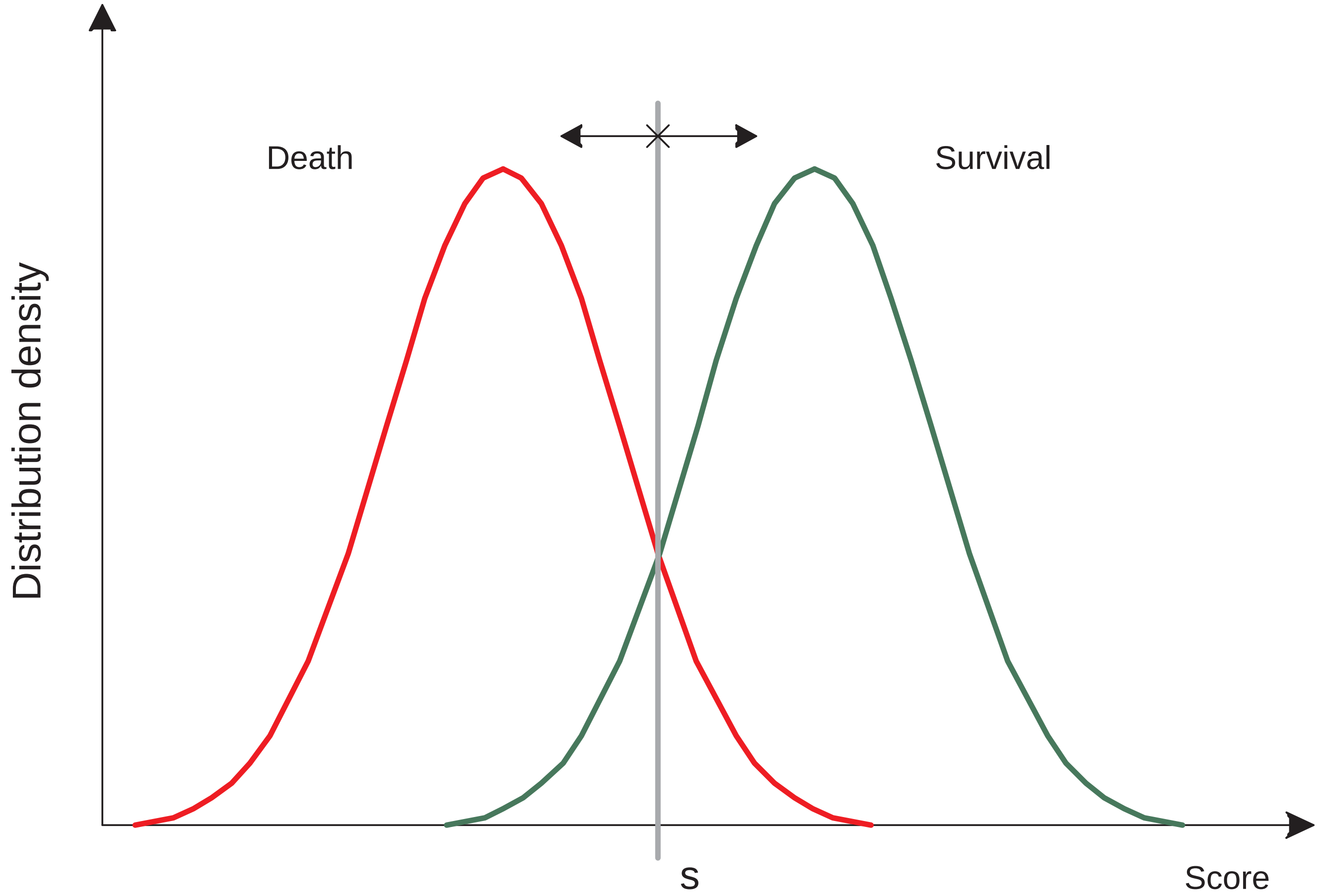
$$\begin{array}{ll} \text{survival:} & S_i \geq s \\ \text{death:} & S_i < s \end{array}$$



Linear Discriminant Analysis



Linear Discriminant Analysis



Other Models

- Logit

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

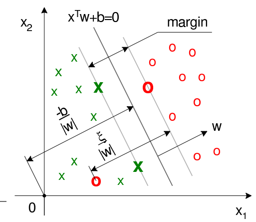
  $y_i = \{0, 1\}$ denotes the class, e.g. 'surviving' or 'dead'

- Probit

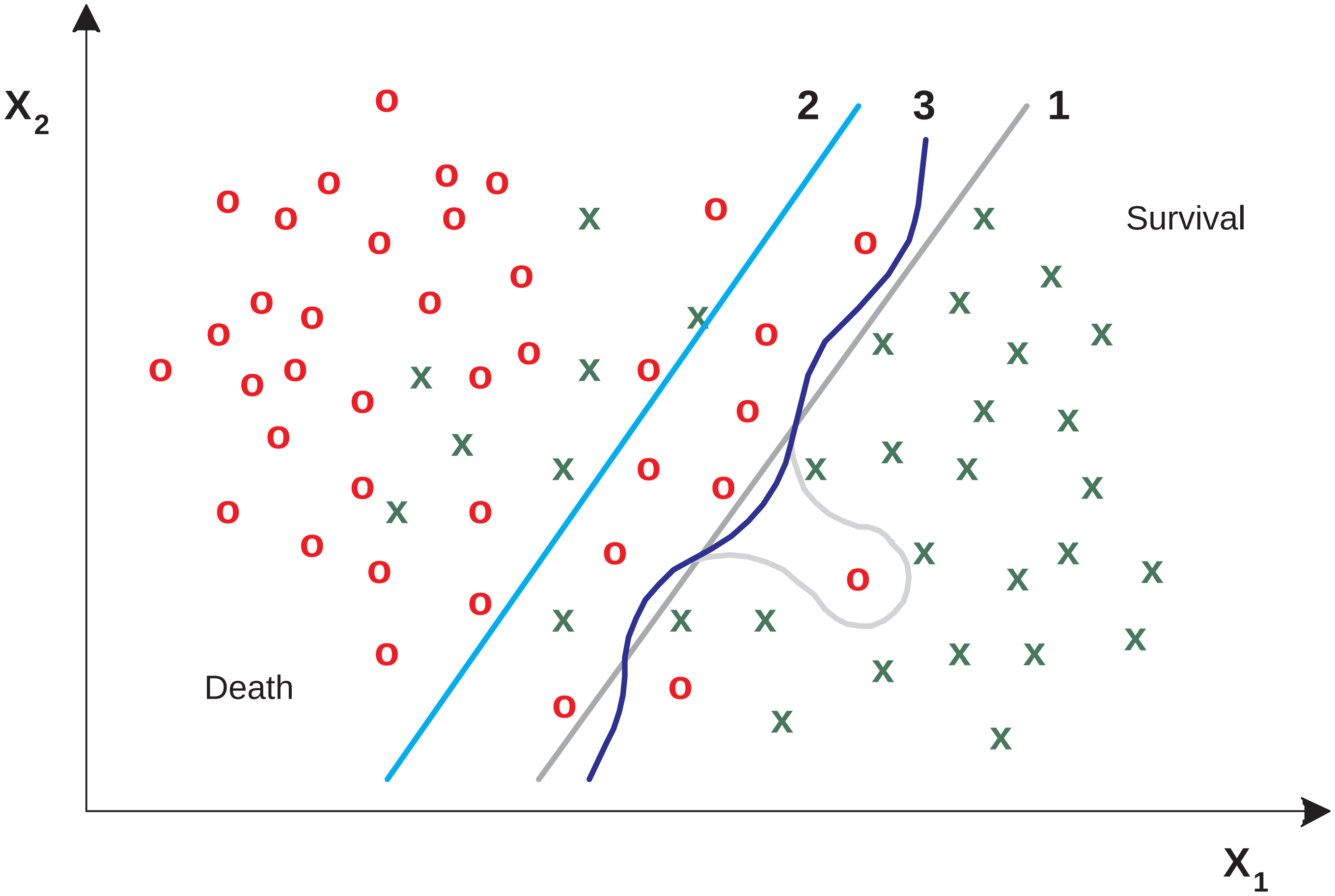
$$E[y_i|x_i] = \Phi\left(a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_d x_{id}\right)$$

- CART

- Neural networks



Linearly Non-separable Classification Problem



Outline of the Talk

√ 1. Motivation

2. Support Vector Machines and Their Properties

3. Expected Risk vs. Empirical Risk Minimisation

4. Realisation of a SVM

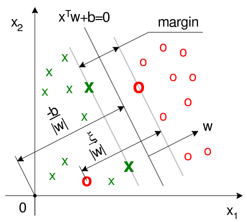
5. Non-linear Case

6. Survival Estimation 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 high, i.e. in most practically significant cases
- SVMs give a unique solution



Classification Problem

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

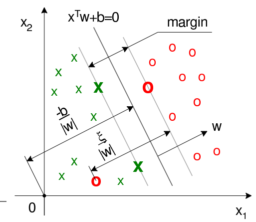
Find the class y of a new object x using the classifier $f : \mathcal{X} \mapsto \{+1; -1\}$, such that the expected risk $R(f)$ is minimal.

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 Minimisation

Expected risk

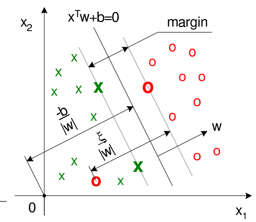
$$R(f) = \int \frac{1}{2}|f(x) - y| dP(x,y) \qquad = \mathrm{E}_{P(x,y)}[L]$$

can be minimised directly with respect to f

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

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

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



Empirical Risk Minimisation

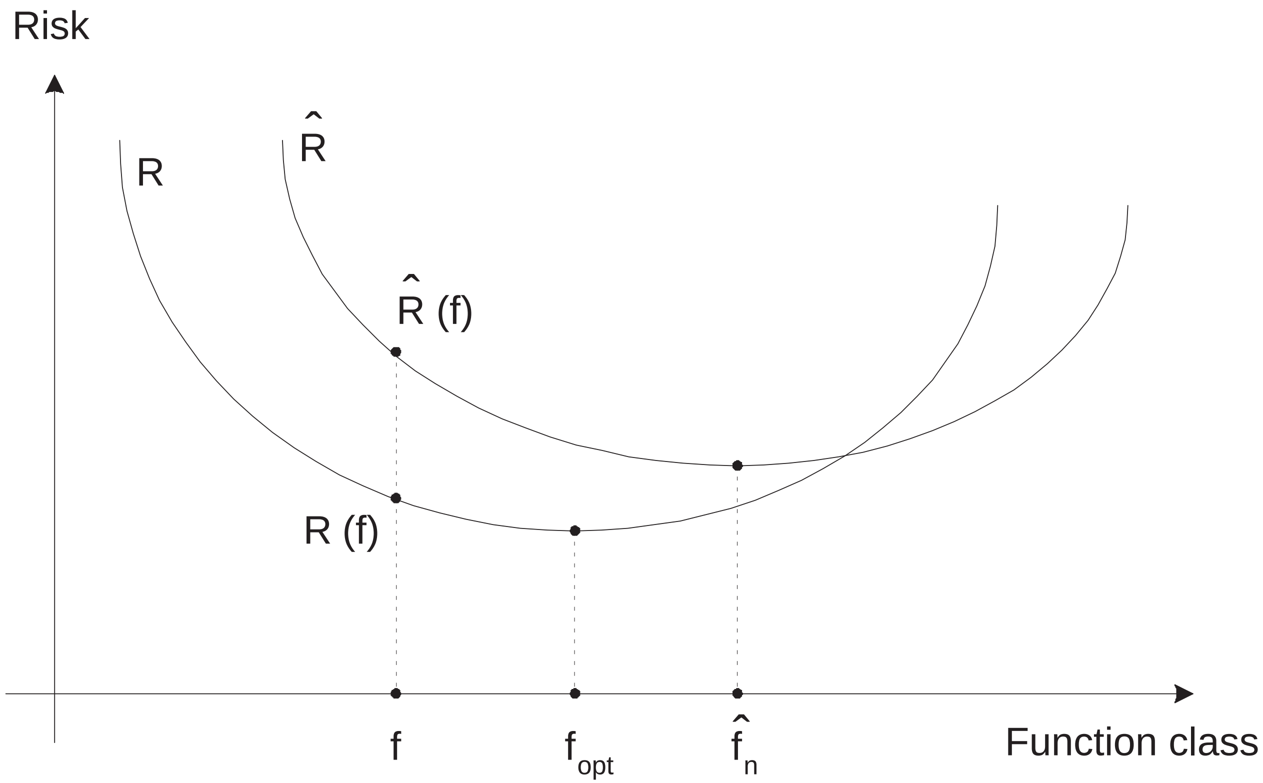
In practice $P(x, y)$ is usually **unknown**: use *Empirical Risk*

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

Minimisation (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

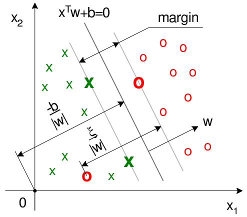
From the law of large numbers

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

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

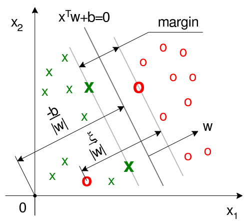
A basic result of Statistical Learning Theory (for linear classifier functions):

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

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

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

Structural Risk Minimisation – search for the optimal model structure described by $\mathcal{S}_h \subseteq \mathcal{F}$ such that the VC bound is minimised; $f \in \mathcal{S}_h$ (h is VC dimension)



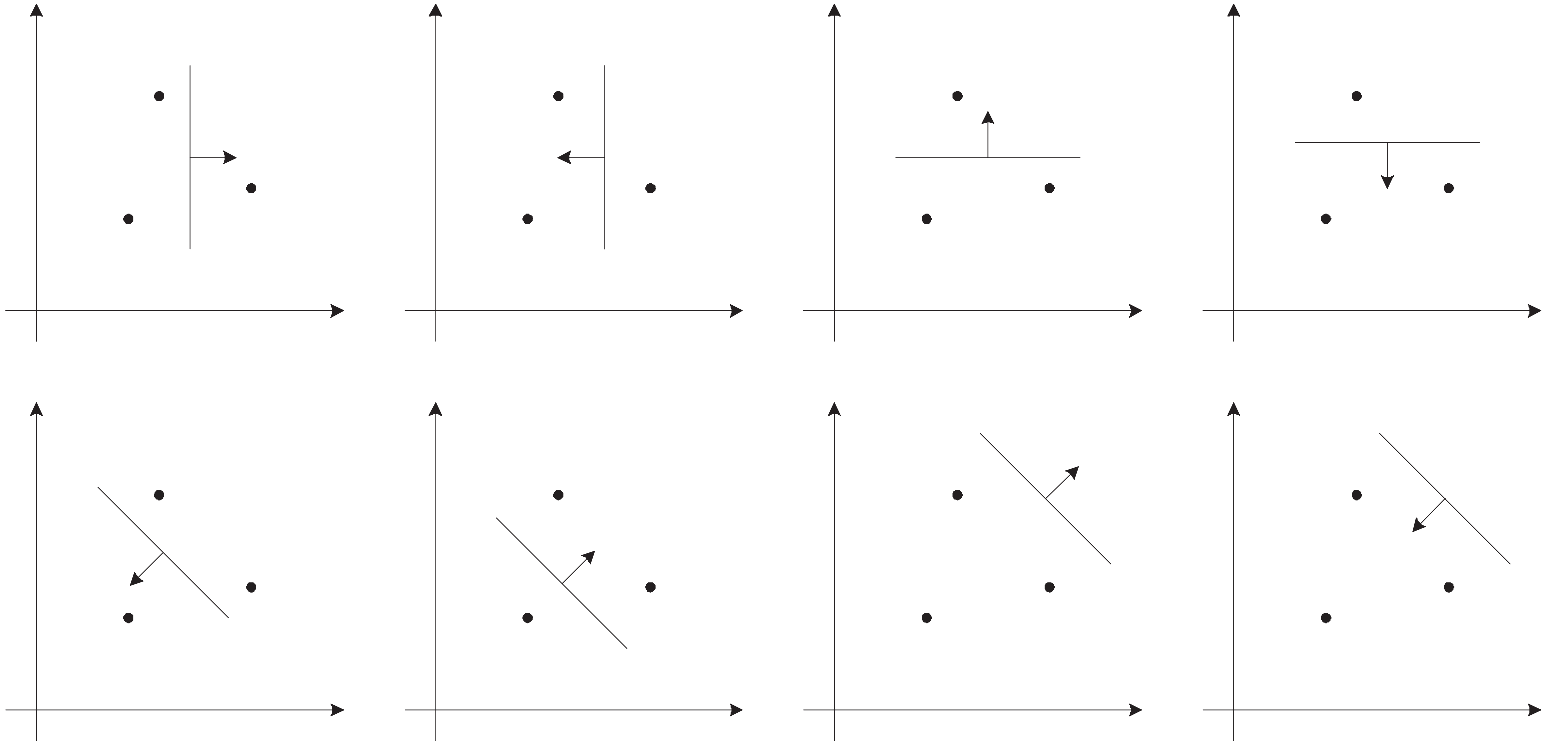
Vapnik-Chervonenkis (VC) Dimension

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

VC Dimension. Example



Regularised LS Estimation and VC Bound

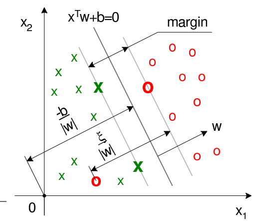
Problem solved:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \{f(x_i) - y_i\}^2 + \lambda \Omega(f)$$

The regularised functional: a specific type of the VC bound with a quadratic empirical loss function

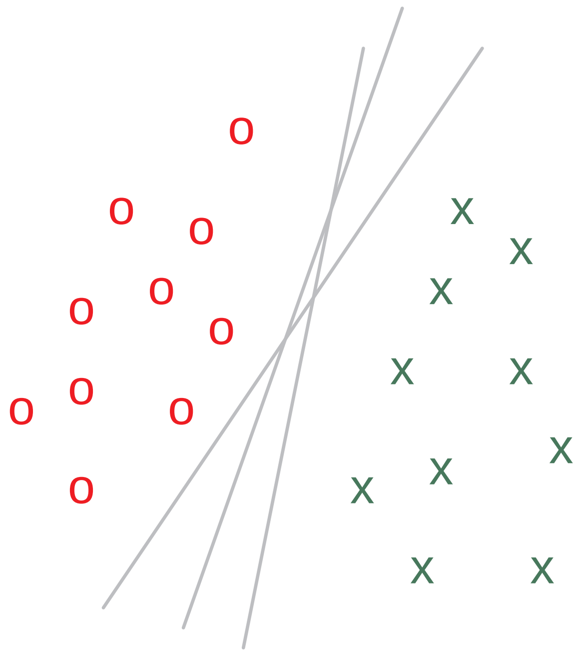
The Classifier Function Class of an SVM

$$\mathcal{F}_\Lambda = \{f : \mathbb{R}^n \mapsto \mathbb{R} | f(x) = w^\top x + b, \|w\| \leq \Lambda\}$$



Linearly 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" – the gap between the parallel hyperplanes separating two classes where the vectors of neither class can lie. Maximisation of the margin minimises the VC dimension.



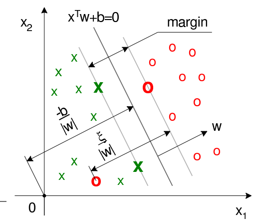
Let $x^\top w + b = 0$ be a separating hyperplane. Then d_+ (d_-) will be the shortest distance to the closest objects of the classes $+1$ (-1).

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

combine them into one constraint

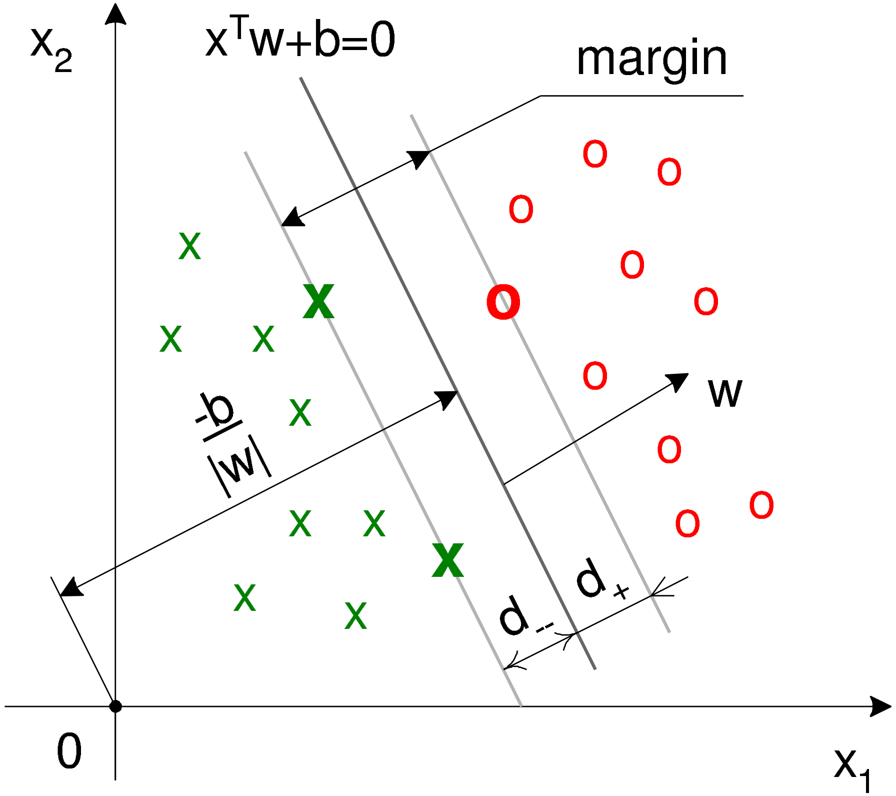
$$y_i(x_i^\top w + b) - 1 \geq 0 \qquad i = 1, 2, ..., n \tag{1}$$

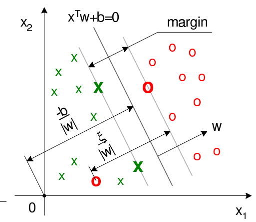
The canonical hyperplanes $x_i^\top w + b = \pm 1$ are parallel and the distance between each of them and the separating hyperplane is $d_\pm = 1/\|w\|$.



Linear SVMs. Separable Case

The **margin** is $d_+ + d_- = 2/\|w\|$. To maximise it minimise 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

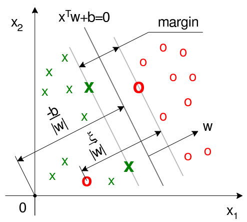
$$\frac{\partial L_P}{\partial w_k} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i x_{ik} = w_k \qquad k = 1, ..., d$$

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

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

$$\alpha_i \geq 0$$

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



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

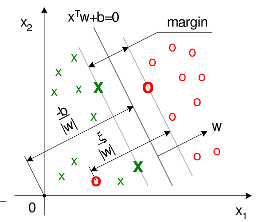
$$\min_{w_k, b} \max_{\alpha_i} L_P$$

$$\max_{\alpha_i} L_D$$

s.t.

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

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



The Classification Stage

The classification rule is:

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

where

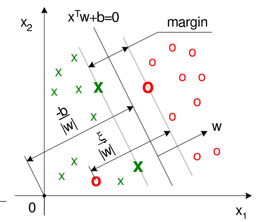
$w = \sum_{i=1}^{n} \alpha_i y_i x_i$

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

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

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

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



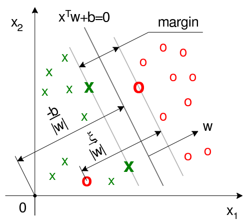
Adaption of an SVM to Hazard Estimation

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

$$f \mapsto hazard$$

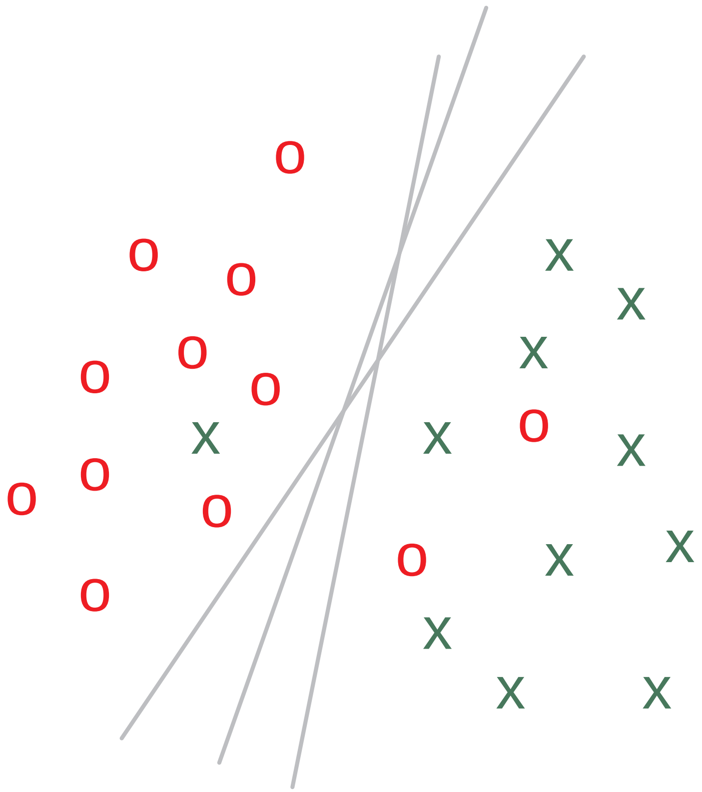
Suggestion:

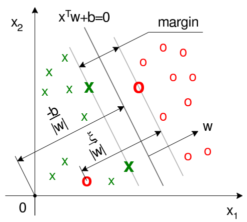
- select an area $f \pm \Delta f$
- count the number of deaths and survivals in the area
- if the data is representative of the whole population
  $\hat{hazard} = \#\text{deaths}/\#\text{survivals}$
- estimate the mapping $f \mapsto \hat{hazard}$ for several $f \pm \Delta f$



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

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

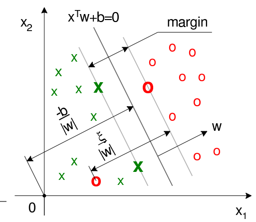
$$x_i^\top w + b \leq -1 + \xi_i \qquad for \quad y_i = -1$$

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

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



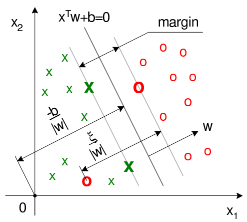
The Lagrangian Formulation

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

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

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

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

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

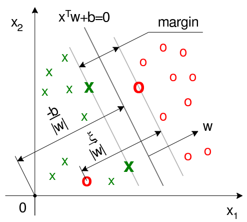
$$\xi_i \geq 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0$$

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

$$\mu_i \xi_i = 0$$



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

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

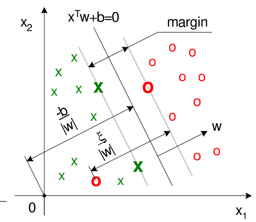
The dual problem is

$$\max_{\alpha_i} L_D$$

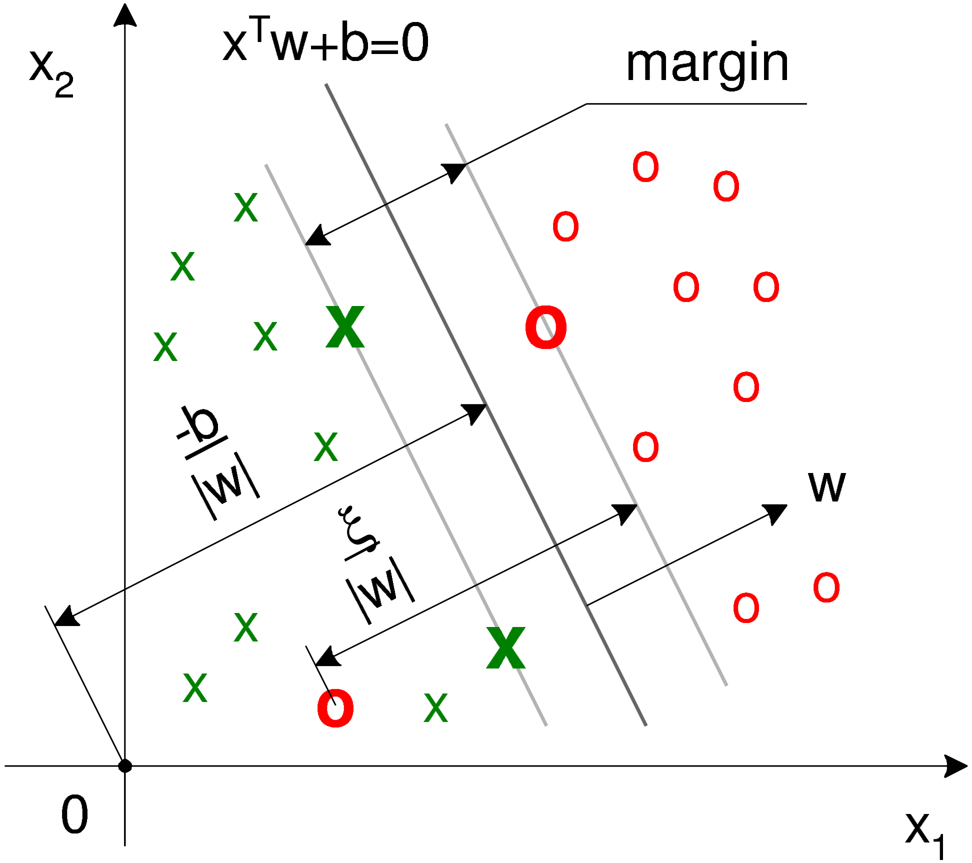
subject to

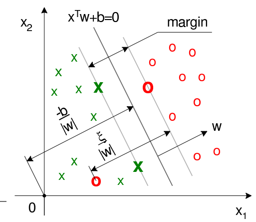
$$0 \leq \alpha_i \leq C$$

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



Linear SVM. Non-separable Case





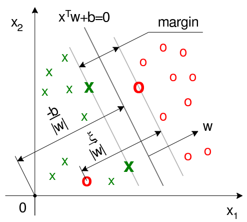
Non-linear SVMs

Map the data to the 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 $\Psi(x_i)^\top \Psi(x_j)$.

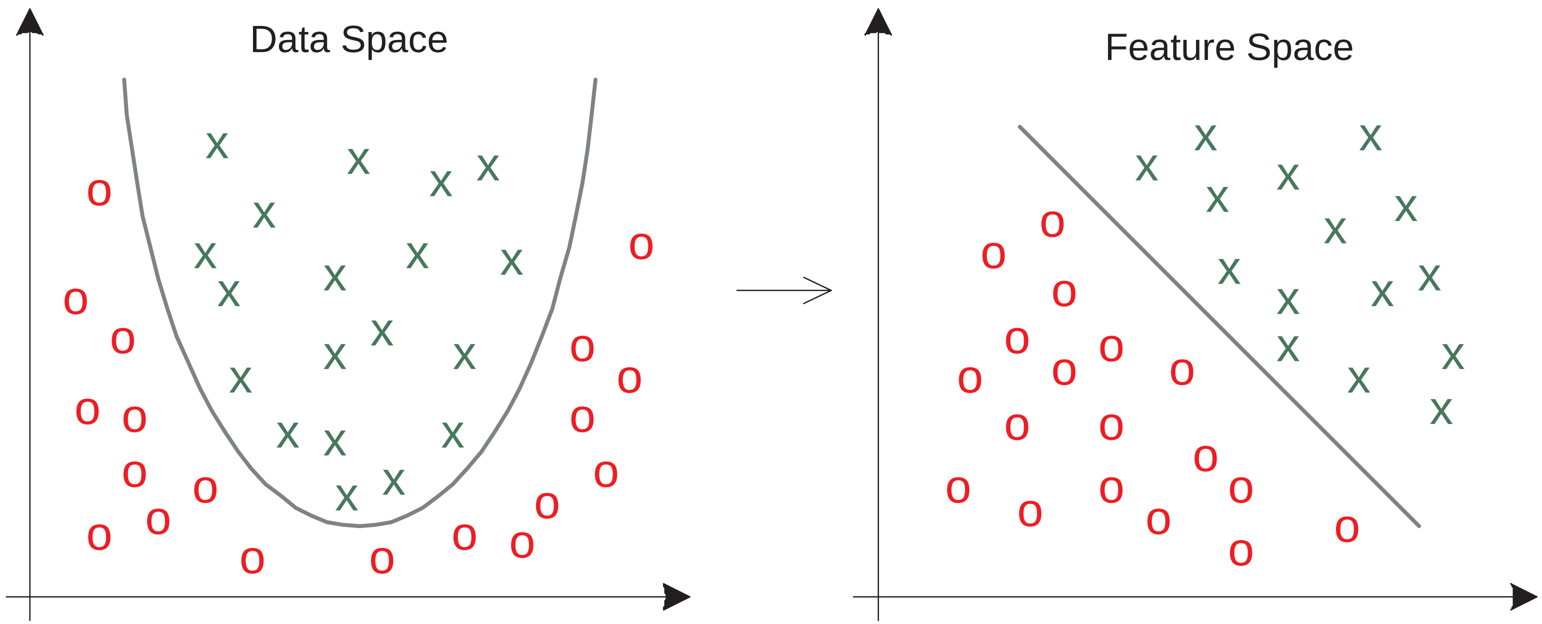
If a *kernel function* K exists such that $K(x_i, x_j) = \Psi(x_i)^\top \Psi(x_j)$, then we can use K without knowing Ψ explicitly

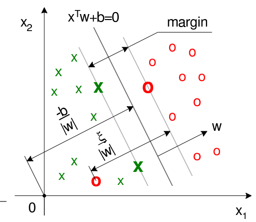


Mapping into the Feature Space. Example

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

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





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

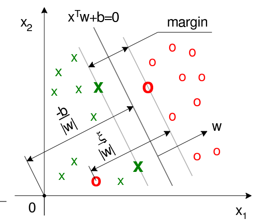
$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j K(x_i, x_j) \geq 0$$

Some examples of kernel functions:

$K(x_i, x_j) = e^{-(x_i - x_j)^\top \Sigma^{-1} (x_i - x_j)/2}$ – 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

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

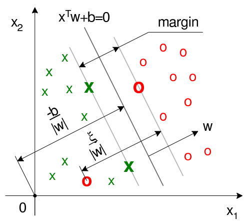
An **isotropic** (homogeneous) kernel is one which depends only on the norm of the lag vector (distance) between two data points

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

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

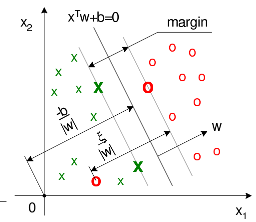
$$\frac{K_I(\|x_i - x_j\|)}{K_I(0)} = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\sqrt{\nu}\|x_i - x_j\|}{\theta}\right)^{\nu} H_\nu\left(\frac{2\sqrt{\nu}\|x_i - x_j\|}{\theta}\right)$$

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 $\nu \to \infty$.

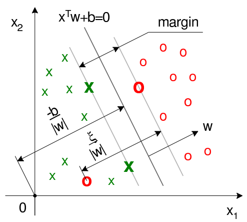
Estimation of Survival Chances for Breast Cancer Patients

- Data source: the "Breast cancer survival.sav" file supplied with SPSS and the database used in Lee et al. (2001)
- 325 cases selected and merged in one database (112 deaths, 223 censored cases)
- Predictors: 2 variables that are contained in both databases – the pathology size and the number of methastased lymph nodes
- an SVM with an anisotropic Gaussian kernel with the radial basis $3\Sigma^{1/2}$ and capacity $C = 1$ was applied (here $\Sigma =$ cov. matrix)

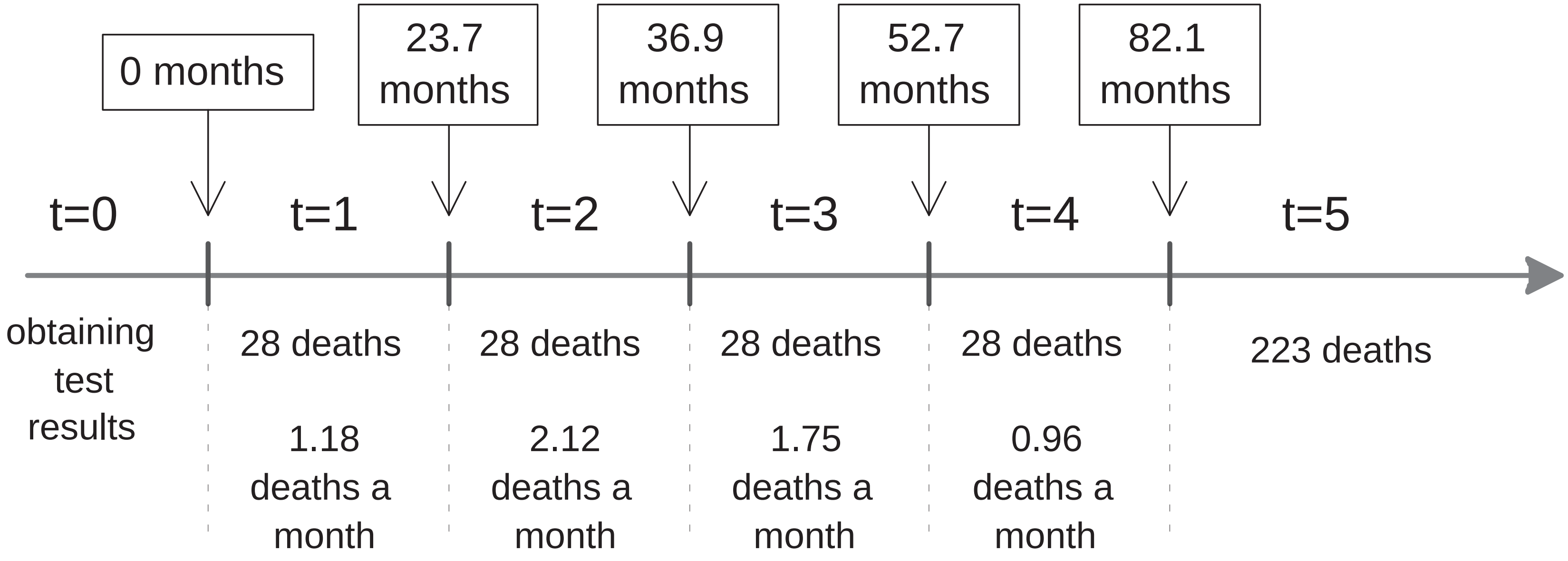


Methodology

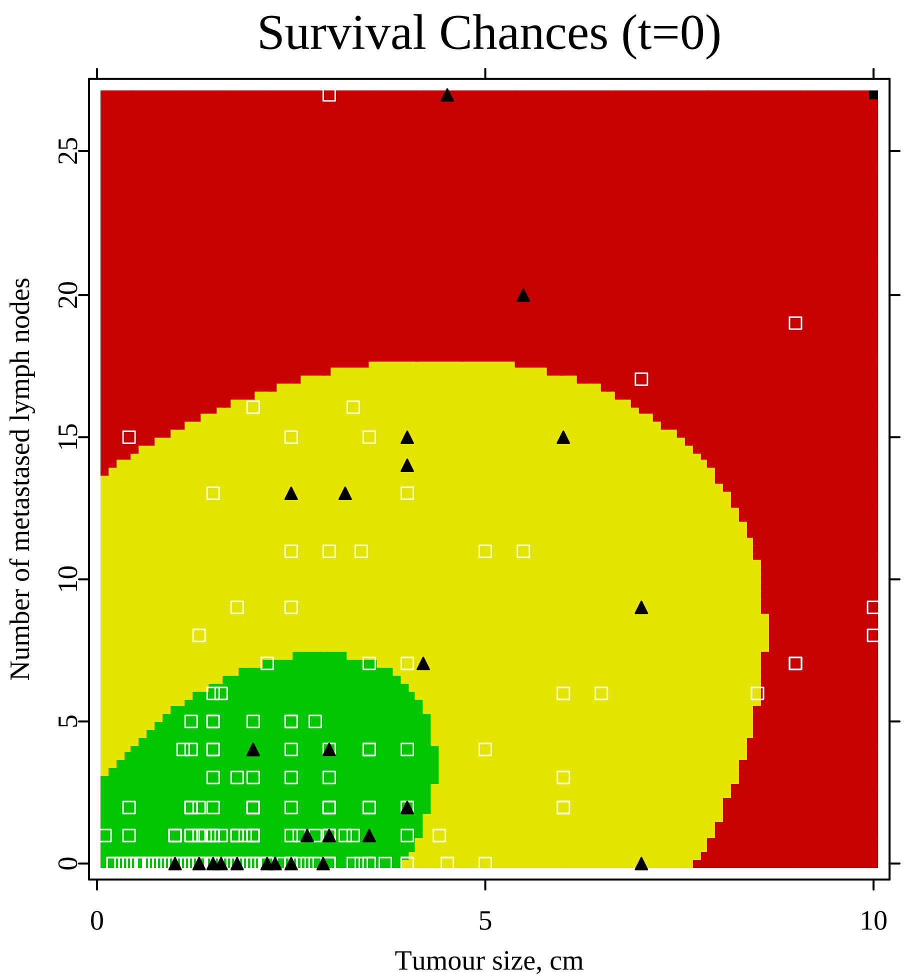
- the cases were sorted in ascending order by survival time or time to censoring
- 5 groups $(t = 1, ..., 5)$ were selected; all 112 death cases are in groups $t = 1, ..., 4$; all 213 censored cases are in group $t = 5$
- an SVM was trained at time t $(t = 0, ..., 3)$; the patients who would die in period $t + 1$ were given the label $y_i = 1$, those who would survive: $y_i = -1$

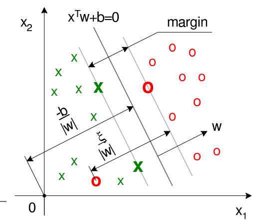


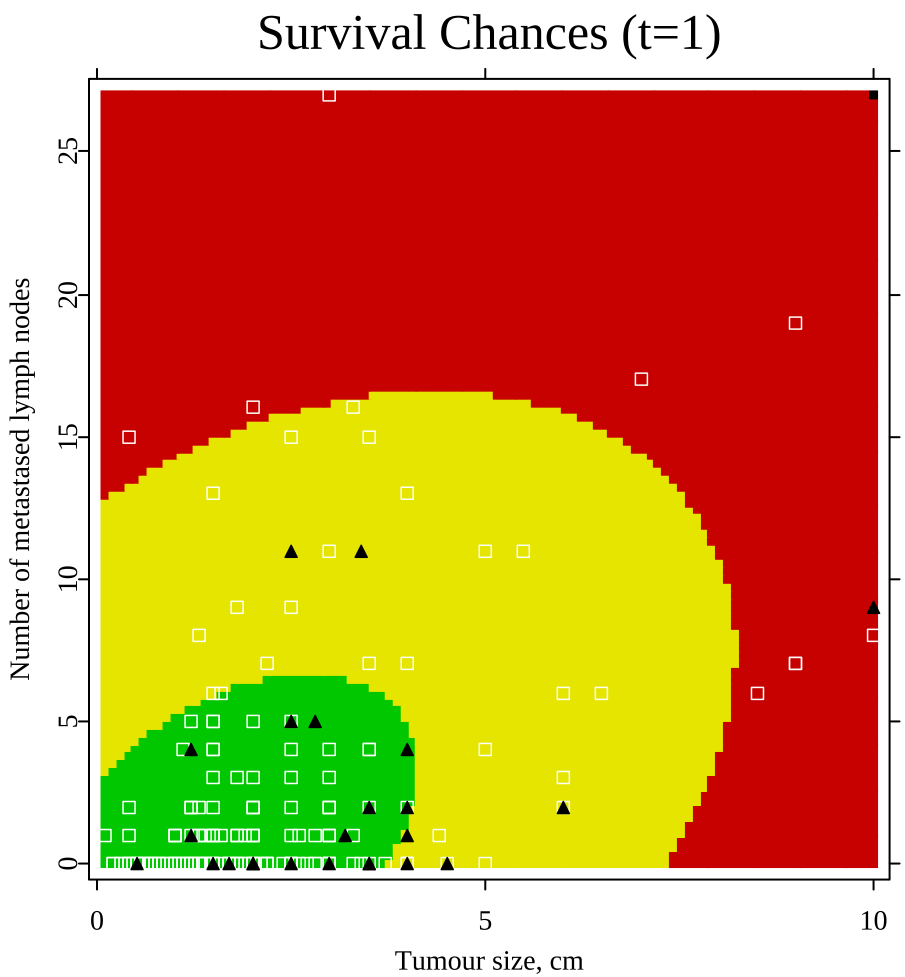
The Timeline

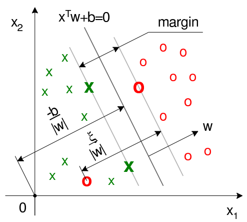


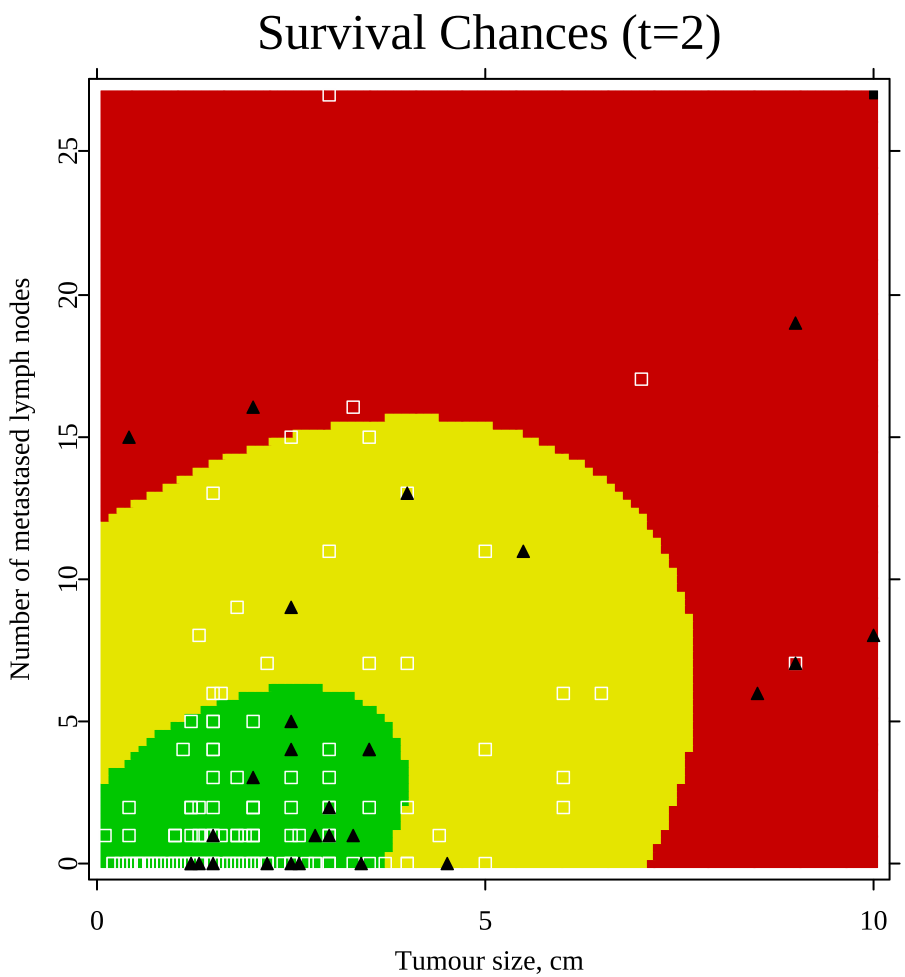
Survival Estimation

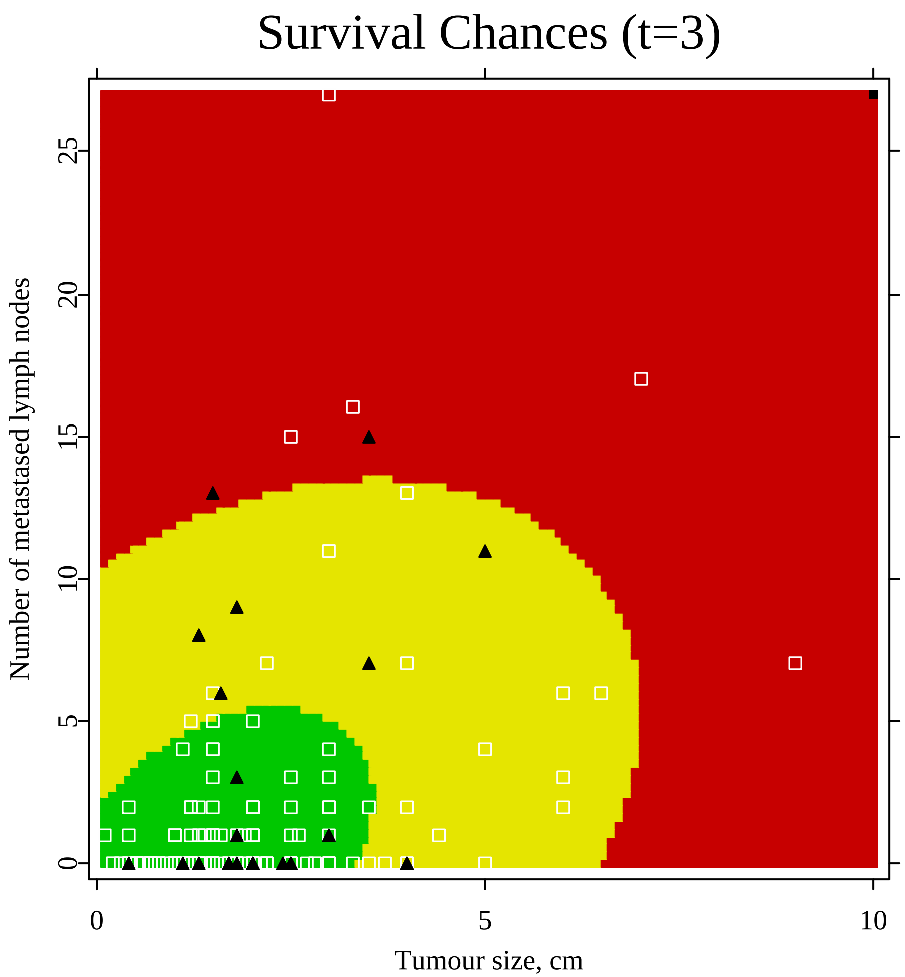

Survival Chances (t=0)
Number of metastased lymph nodes
Tumour size, cm
0
5
10
15
20
25
0
5
10




Survival Chances (t=1)
Number of metastased lymph nodes
Tumour size, cm
0
5
10
15
20
25


x2
x1
margin
0
w


Survival Chances (t=2)
Number of metastased lymph nodes
0
5
10
15
20
25
0
5
10
Tumour size, cm


Survival Chances (t=3)
Number of metastased lymph nodes
0
5
10
15
20
25
0
5
10
Tumour size, cm

References

Cox, D. R. (1972). Regression Models and Life Tables, *Journal of the Royal Statistical Society* **B34**: 187-220.

Lee, Y.-J., Mangasarian, O. L., Wolberg, W. H. (2001). Survival Time Classification of Breast Cancer Patients (technical report): ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-03.ps.

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