SUPPORT VECTOR MACHINES

Václav Hlaváč

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

hlavac@fel.cvut.cz, http://cmp.felk.cvut.cz

LECTURE PLAN

- Discriminative approach. Maximal margin classifier.
- Minimization of the structural risk.
- SVM, task formulation, solution: quadratic programming.
- Linearly separable case.
- Linearly non-separable case.

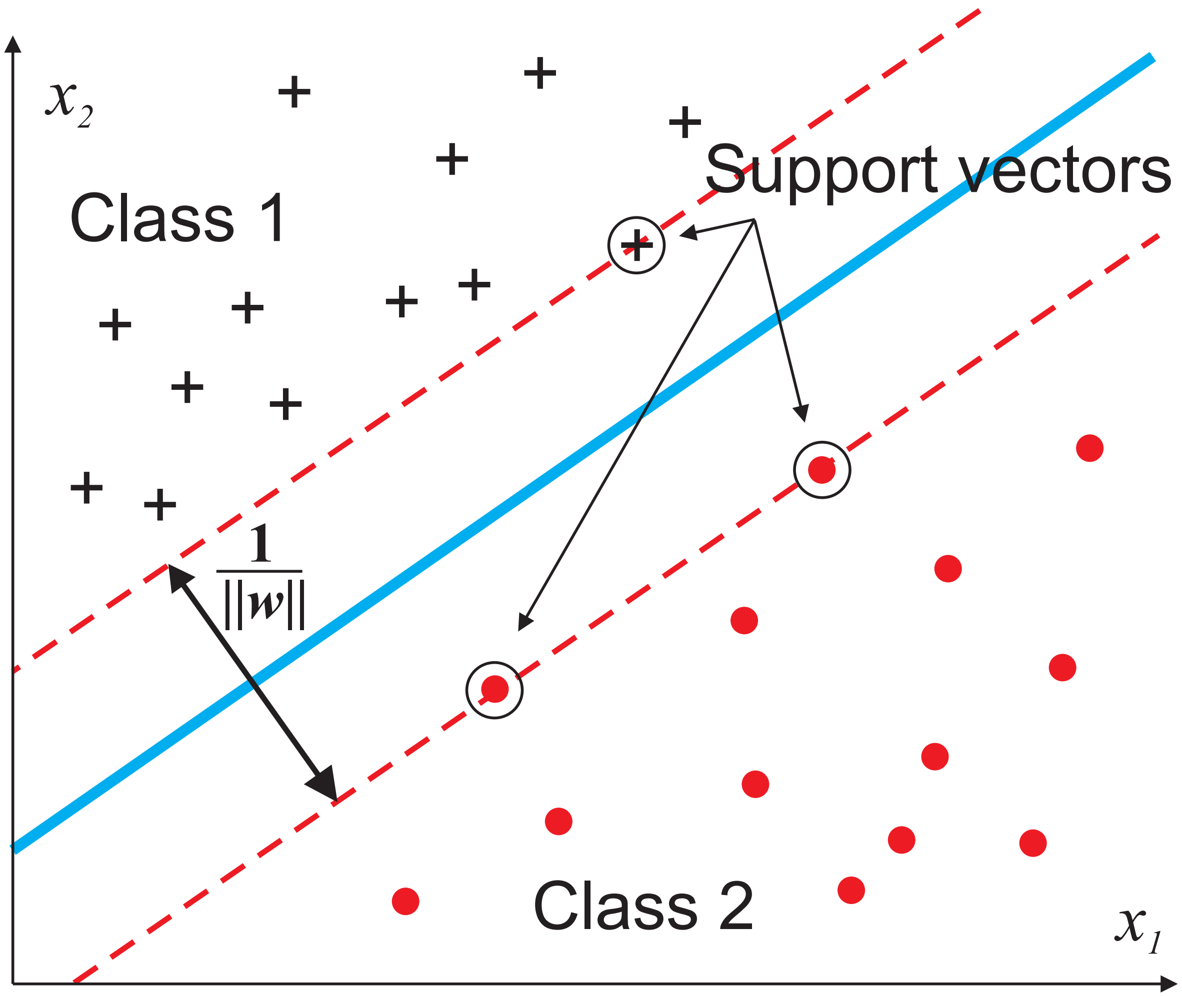
INTRODUCTION

- There are two principal approaches to design a classifier:
  - Generative.
  - Discriminative.
- So far, the generative methods were used. A known statistical model was assumed $\Rightarrow$ decision rule.
- Now, we will assume that class of decision rules is known.

  V. Vapnik: Learning is the selection of one decision rule from the class of rules.

MAXIMAL MARGIN CLASSIFIER 1

- Maximizes margin between classes which increases generalization ability.
- The Vapnik's Support Vector Machine is based on the same idea.



SUPPORT VECTOR MACHINES, TASK

- Two hidden states (classes) only, k_1, k_2.
- A separable hyperplane is sought which maximizes a distance (margin) between classes.
- The task is converted into a quadratic programming task

$$(w^*, b^*) = \operatorname*{argmin}_{w,b} \frac{1}{2} \|w\|^2$$

under constraints

$$\begin{aligned} \langle w, x_j \rangle + b &> 1 \quad \text{for} \quad k_j = 1 \\ \langle w, x_j \rangle + b &< 1 \quad \text{for} \quad k_j = 2 \end{aligned}$$

SUPPORT VECTOR MACHINES, A ROAD MAP

Minimization of the structural risk
Theory how to learn classifier

↓

Linear classifier, a maximal margin classifier
Seek for a maximal margin, respectively. a soft margin for nonseparable data

↓

Learning expressed as quadratic optimization
Primal task

↓

Transformation of the primal task to the dual task
In dual task, data is expressed as scalar products

↓

Support Vector Machines

Extensions

Straightening of the feature space

Use of kernel functions.

INTRODUCTION

- Learning the classifier from the finite training set.
- There is an estimate – upper bound of the mean classification error.
- Solves problem of generalization, i.e. choice of a statistical model.

ASSUMPTIONS

- $x \in \mathbb{R}^n$ . . . observation of the object (vector of measurements).
- $y \in \{-1, 1\}$ . . . hidden states
- There is a training set available $\{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)\}$, which is drawn randomly and generated by an unknown probability distribution $p(x, y)$.

MINIMIZATION OF THE STRUCTURAL RISK

THE AIM

is to find a classifier

$$f(x, a) \,,$$

where a is a parameter with the minimal expected classification error (risk)

$$R(f(x,a)) = \int \frac{1}{2} \left| y - f(x,a) \right| \, \mathrm{d}\, p(x,y) \,.$$

Note: a 1/0 loss (penalty) function was used, i.e.,

$$\frac{1}{2} \left| y - f(x,a) \right| = \begin{cases} 0 & \text{if } y = f(x,a) \,, \\ 1 & \text{if } y \neq f(x,a) \,. \end{cases}$$

MINIMIZATION OF THE STRUCTURAL RISK

COMPLICATIONS

$R(f(x,a))$ cannot be calculated because the probability distribution $p(x,y)$ is unknown.

SOLUTION

Use the upper bound for R by Vapnik-Červoněnkis.

$$R(f(x,y)) \leq R_{emp} + \underbrace{\sqrt{\frac{h\left(\log\frac{2L}{h}+1\right) - \log\frac{\eta}{4}}{L}}}_{\text{structural risk}}$$

MINIMIZATION OF THE STRUCTURAL RISK

Empirical risk $$R_{emp} = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{2} |f(x_i, a) - y_i|$$

h is a VC dimension characterizing the class of decision functions $f(x, a) \in F$.

L is the length of the training set.

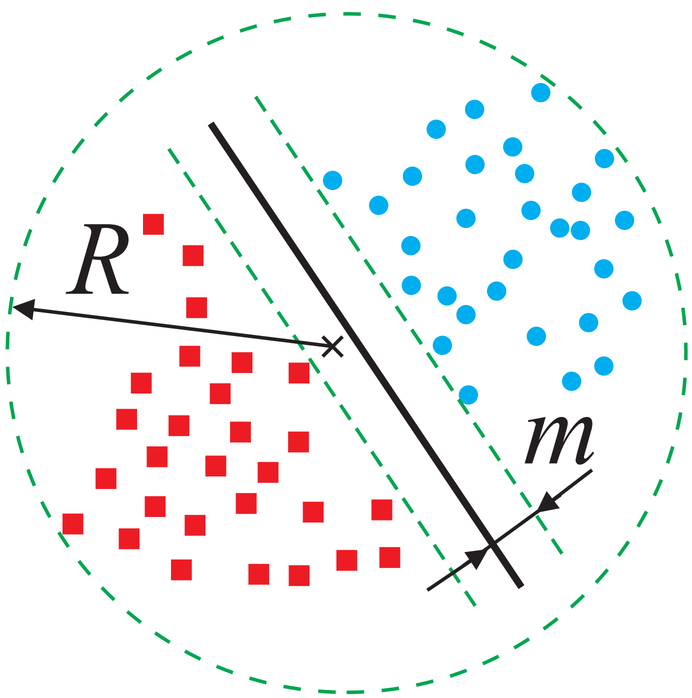
η is the degree of belief into the bound $R(f(x, a))$, i.e., $0 \leq \eta \leq 1$.

Support Vector Machines implement structural risk minimization principle.

LINEARLY SEPARABLE SVM

The aim is to find linear discriminant function

$$f(x, w, b) = \mathrm{sign}(\langle w, x\rangle + b) = \mathrm{sign}\left(w^\top x + b\right)$$



- VC dimension (capacity) depends on the margin m

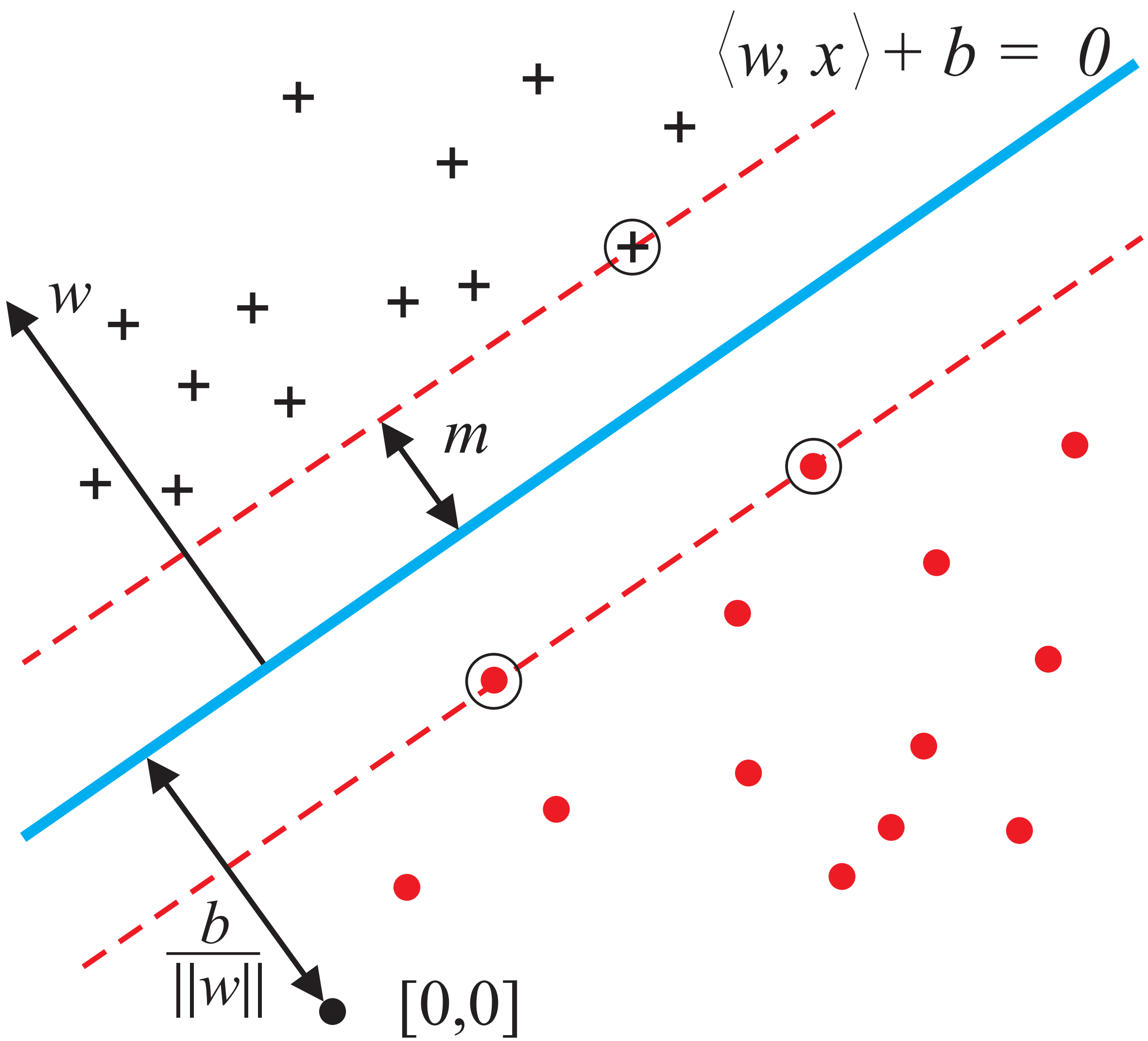
$$h \leq \frac{R^2}{m^2} + 1$$

- R is given by the data itself.
- Margin m can be optimized in the classifier design.

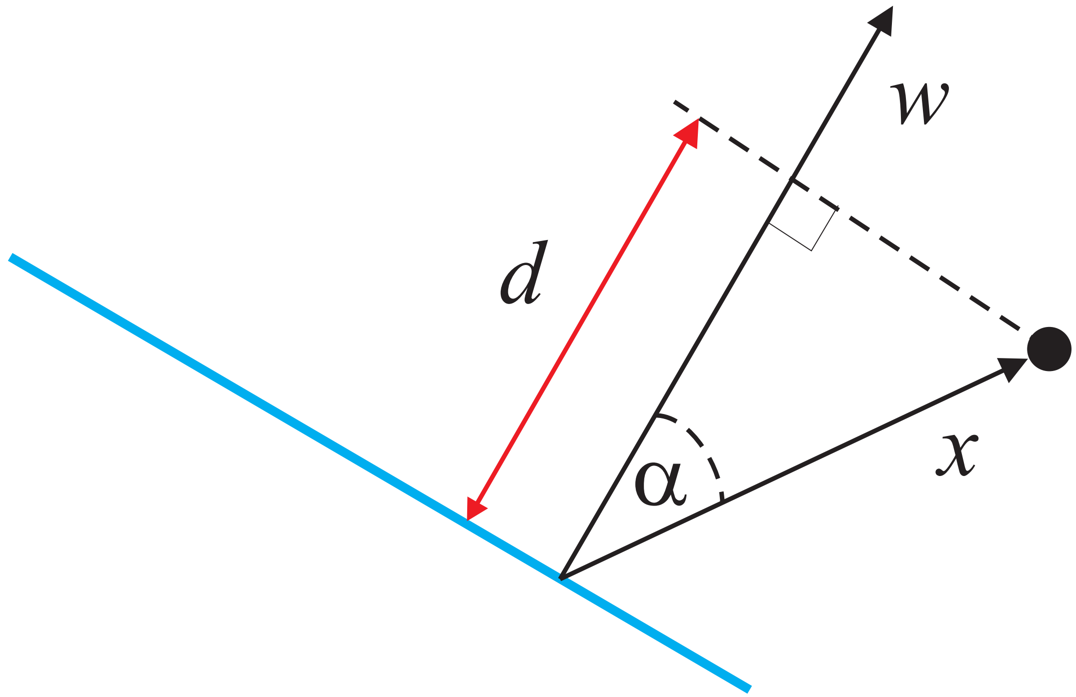
Conclusion: separation hyperplanes with larger margin have lower VC dimension $\Leftrightarrow$ lower value of the upper bound.

LINEARLY SEPARABLE SVM (2)

The separating hyperplane is sought which maximizes distance to data (margin).



LINEARLY SEPARABLE SVM (3)



The distance between the observation x_i and the separating hyperplane $w^\top x_i + b = 0$ is

$$\cos\alpha = \frac{w^\top x_i}{\|w\| \|x_i\|}, \quad \cos\alpha = \frac{d}{\|x_i\|} \quad \Rightarrow \quad d = \frac{w^\top x_i + b}{\|w\|}$$

LINEARLY SEPARABLE SVM, PRIMAL TASK

The optimization task

$$(w^*, b^*) = \operatorname*{argmax}_{w,b} \min_{i=1,\ldots,L} \frac{w^\top x_i + b}{\|w\|} y_i$$

can be converted in to a standard quadratic programming problem (primal task)

$$(w^*, b^*) = \operatorname{argmin} \frac{1}{2}\|w\|^2$$

$$w^\top x_i + b \geq +1\,, \quad y_i = +1$$

$$w^\top x_i + b \leq -1\,, \quad y_i = -1$$

TOWARDS THE DUAL TASK

The aim is to convert the problem into a formulation without constraints.

Lagrange function L is introduced, α_i are Lagrange multipliers,

$$L(w, b, \alpha_i) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i \left(w^\top x_i\right) \, y_i + \sum_{i=1}^{L} \alpha_i \, . \quad \text{(Eq. 1)}$$

Now we have formulated the dual task, i.e., the problem without constraints

$$(w^*, b^*) = \operatorname*{argmin}_{w,b} \, \max_{\alpha_i > 0} \, L(w, b, \alpha_i) \, .$$

SOLUTION TO THE DUAL TASK

$$\min_{w,b} \max_{\alpha_i > 0} L(w, b, \alpha_i) = \max_{\alpha_i > 0} \min_{w,b} L(w, b, \alpha_i)$$

Seek optimum, i.e., 1st partial derivatives $= 0$,

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{L} \alpha_i y_i x_i \, , \qquad \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0 \, .$$

Substitute to (Eq. 1), get rid off w, b and get

$$\alpha_i = \operatorname*{argmax}_{\alpha_i} \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j x_i^{\top} x_j \, ,$$

$$\alpha_i \geq 0 \, , \qquad \sum_{i=1}^{L} \alpha_i y_i = 0 \, .$$

SVM – PRIMAL AND DUAL TASKS

Primal task

- Optimized according to vector $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$.
- Number of variables is $n + 1$.
- Number of linear constraints is n.

Dual task

- Optimized according to $\alpha_1, \alpha_2, \ldots, \alpha_L$, $\alpha_i \in \mathbb{R}$.
- Number of variables is L.
- Number of linear constraints is $L + 1$.
- Data appear as scalar products only, i.e., $x_i^\top x_j$.

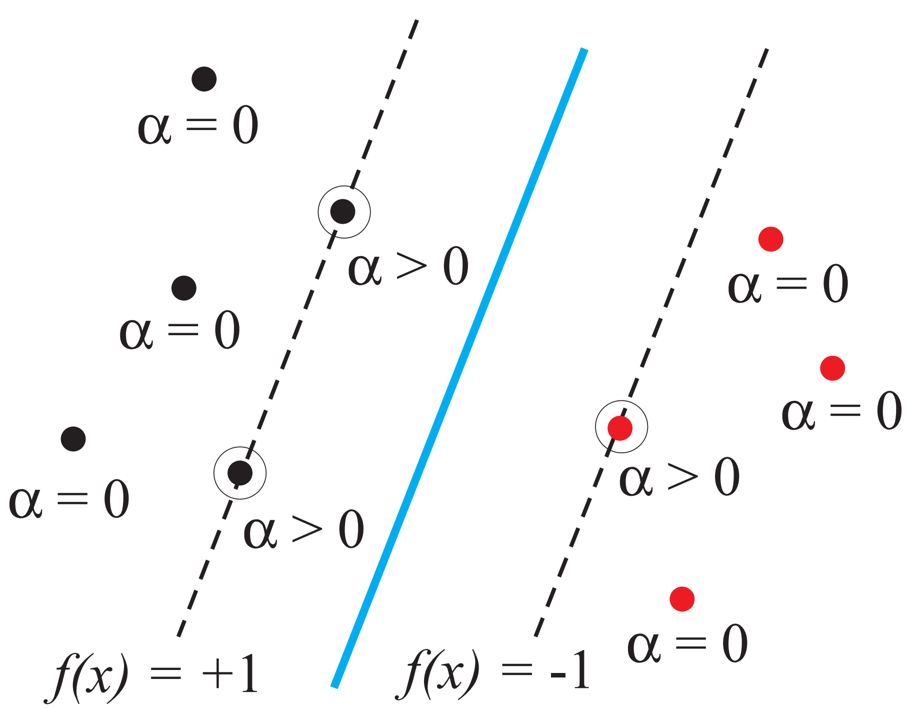
DUAL TASK PROPERTIES, cont.

- The solution is sparse. Many α_i equal to 0.

  $\alpha_i = 0 \Rightarrow y_i(w^\top x_i + b) > 1.$

  $\alpha_i > 0 \Rightarrow y_i(w^\top x_i + b) = 1.$

- Data x_i for which $\alpha_i > 0$ are called Support Vectors.



$$w = \sum_{i=1}^{L} \alpha_i y_i x_i = \sum_{i \in \mathrm{SV}} \alpha_i y_i x_i$$

Calculation of b for $i \in \mathrm{SV}$:

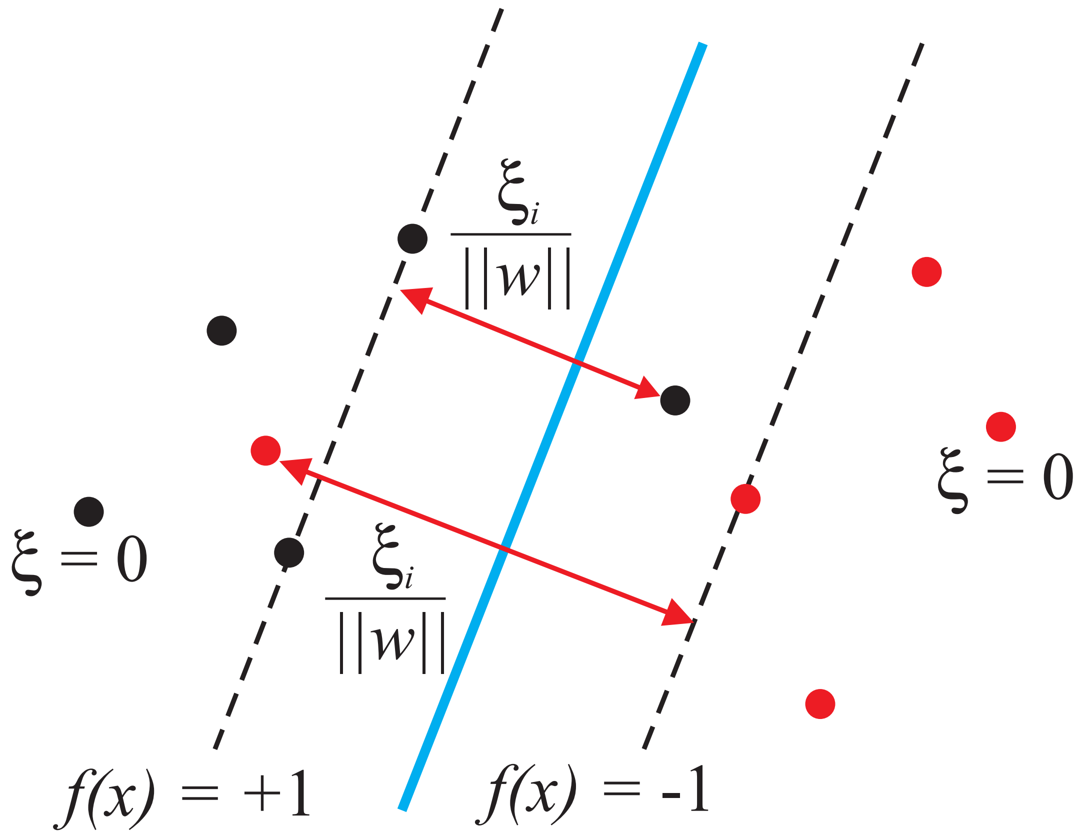
$$y_i(w^\top x_i + b) = 1 \Rightarrow b = \frac{1 - y_i w^\top x_i}{y_i}$$

One SV should be enough.

Practically, many SVs, mean of b.

SVM LINEARLY NON-SEPARABLE

Nonseparable data. $\Leftrightarrow$ It is not possible to find separable hyperplane without errors.



Solution: Regularization, i.e., introduction of slack variables $\xi \geq 0 \Rightarrow$ Soft Margin SVM.

SOFT MARGIN SVM

$$(w^*, b^*, \xi^*) = \underset{w,b,xi_i}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{L} \xi_i^k$$

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

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

Optimization criterion, marginal behavior

- $\min \|w\|^2$ – maximization of the margin.
- $\min \sum_{i=1}^{L} \xi_i^k$ – minimal number misclassified training points (upper bound of the empirical error).

Quadratic programming for $k = 1, 2$.

SVM LINEARLY NON-SEPARABLE, cont.

How to choose regularization constant C? Common solutions:

- Design the classifier for several values of $C = \{C_1, \ldots, C_n\}$. Follow by 1D optimization.
- Use some other criterion to choose C, e.g., cross validation.
- Transform to dual task, analogically to separable case.

$$\alpha_i = \underset{\alpha_i}{\operatorname{argmax}} \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j x_i^\top x_j \ ,$$

$$0 \le \alpha_i \le C \ , \qquad \sum_{i=1}^{L} \alpha_i y_i = 0 \ .$$

Note: $\le C$ above is the only difference when comparing to the linearly separable case.

SOFT MARGIN SVM, THEORETIC BACKING

$$\mathrm{Risk} = \frac{C}{L}\left(\frac{R^2 + \left(\sum_{i=1}^{L}\xi_i\right)\log\left(\frac{1}{L}\right)}{m^2}\log^2 L + \log\left(\frac{1}{\eta}\right)\right)$$

is minimized when

$$\|w\|^2 R + \left(\sum_{i=1}^{L}\xi_i\right)\log\left(\frac{1}{\sqrt{(\|w\|)}}\right)$$

This matches to Soft Margin SVM criterion with exception to the last term on the right side.