

Support Vector Machines

Miguel A. Veganzones

Grupo Inteligencia Computacional
Universidad del País Vasco

Outline

- 1 Introduction
 - A brief history of the Learning Problem
 - Vapnik-Chervonenkis (VC) dimension
 - Structural Risk Minimization (SRM) Inductive Principle
- 2 Support Vector Machines (SVM)
 - Optimal hyperplane for linearly separable patterns
 - Optimal hyperplane for non-separable patterns
 - SVMs for pattern recognition



Outline

- 1 Introduction
 - A brief history of the Learning Problem
 - Vapnik-Chervonenkis (VC) dimension
 - Structural Risk Minimization (SRM) Inductive Principle
- 2 Support Vector Machines (SVM)
 - Optimal hyperplane for linearly separable patterns
 - Optimal hyperplane for non-separable patterns
 - SVMs for pattern recognition



Rosenblatt's Perceptron (the 1960s)

- F. Rosenblatt suggested the first model of a learning machine, the Perceptron.
- He described the model as a program for computers and demonstrated with simple experiments that this model could generalize.
- The Perceptron was constructed for solving pattern recognition problems.
 - Simplest case: construct a rule for separating data of two different classes using given examples.



Novikoff's theorem (1962)

- In 1962, Novikoff proved the first theorem about the Perceptron, starting learning theory.
- It somehow connected the cause of generalization ability with the principle of minimizing the number of errors on the training set.
- Novikoff proved that Perceptron can separate training data, and that if the data are separable, then after a finite number of corrections, the Perceptron separates any infinite sequence of data.



Applied and Theoretical Analysis of Learning Processes

- Many researchers thought that minimizing the error on the training set was the only cause of generalization. Two branches:
 - Applied analysis: to find methods for constructing the coefficients simultaneously for all neurons such that the separating surface provides the minimal number of errors on the training data.
 - Theoretical analysis: to find the inductive principle with the highest level of generalization ability and to construct algorithms that realize this inductive principle.

Construction of the fundamentals of learning theory

- 1968: a philosophy of statistical learning theory was developed.
 - Essentials concepts of emerging theory, VC entropy and VC dimension for indicator functions (pattern recognition problem).
 - Law of large numbers.
 - Main non-asymptotic bounds for the rate of convergence.
- 1976-1981: previous results generalized to the set of real functions.
- 1989: necessary and sufficient conditions for consistency of the empirical risk minimization inductive principle and maximum likelihood method.
- 1990: Theory of the Empirical Risk Minimization Principle.



Neural Networks (1980s)

- 1986: several authors discover the Back Propagation method for simultaneously constructing the vector coefficients for all neurons of the Perceptron.
- Introduction of the neural network concept.
- Researchers in AI became the main players in the computational learning game.
- Statistical analysis keeps apart from the attention of the AI community, focused in constructing “simple algorithms” for the problems where the theory is very complicated.

Alternatives to NN (1990s)

- Study of the Radial Basis Functions methods.
- Structural Risk Minimization principle: SVM.
- Minimum description length principle.
- Small sample size theory.
- Synthesis of optimal algorithms which possess the highest level of generalization ability for any number of observations.

Support Vector Machines

- Originated from the statistical learning theory developed by Vapnik and Chervonenkis.
- SVMs represent novel techniques introduced in the framework of structural risk minimization (SRM) and in the theory of VC bounds.
- Instead of minimizing the absolute value of an error or a squared error, SVMs perform SRM, minimizing VC dimension.
- Vapnik showed that when the VC dimension of the model is low, the expected probability of error is also low (good generalization).
- Remark: good performance on training data is a necessary but insufficient condition for a good model.



Outline

1 Introduction

- A brief history of the Learning Problem
- **Vapnik-Chervonenkis (VC) dimension**
- Structural Risk Minimization (SRM) Inductive Principle

2 Support Vector Machines (SVM)

- Optimal hyperplane for linearly separable patterns
- Optimal hyperplane for non-separable patterns
- SVMs for pattern recognition



Introduction

- The VC dimension is a property of a set of approximating functions of a learning machine that is used in all important results of statistical learning theory.
- Unfortunately its analytic estimations can be used only for the simplest sets of functions.



Two-class pattern recognition case

Indicator functions

- An indicator function, $i_F(\mathbf{x}, \mathbf{w})$, is a function that can assume only two values, say, $i_F(\mathbf{x}, \mathbf{w}) \in \{0, 1\}$ or $i_F(\mathbf{x}, \mathbf{w}) \in \{-1, 1\}$.
- The VC dimension of a set of indicator functions $i_F(\mathbf{x}, \mathbf{w})$ is defined as the largest number h of points that can be separated (shattered) in all possible ways.
- For two-class pattern recognition, a set of l points can be labeled in 2^l possible ways.

Two-class pattern recognition case

Labelings that cannot be shattered in \mathcal{R}^2

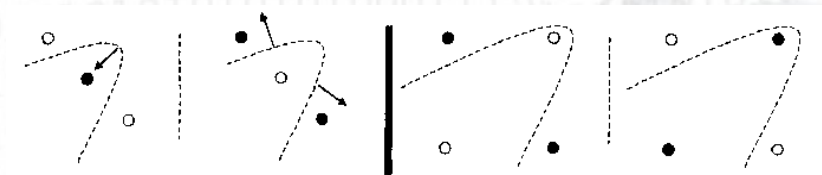


Figure: Left: two labelings of a three co-linear points that cannot be shattered by $i_F(\mathbf{x}, \mathbf{w}) = \text{sign}(u)$. Right: $i_F(\mathbf{x}, \mathbf{w}) = \text{sign}(u)$ cannot shatter the depicted two out of sixteen labelings of four points. A quadratic indicator function (dashed line) can easily shatter both sets of points.



Two-class pattern recognition case

VC Dimension

- In an n -dimensional input space, the VC dimension of the oriented hyperplane indicator function, $i_F(\mathbf{x}, \mathbf{w}) = \text{sign}(u)$, is equal to $h = n + 1$.
 - In a two-dimensional space of inputs, $h = 3$.
- If the VC dimension is h , then there exists at least one set of h points in input space that can be shattered. This does not mean that every set of h points in input space can be shattered by a given set of indicator functions.
 - In a two-dimensional set of inputs at least one set of three points in input space can be shattered by $i_F(\mathbf{x}, \mathbf{w}) = \text{sign}(u)$.
 - In a two-dimensional set of inputs no set of four points can be shattered by $i_F(\mathbf{x}, \mathbf{w}) = \text{sign}(u)$.

Two-class pattern recognition case

VC Dimension and the space of features

- In a n -dimensional input space, the VC dimension of the oriented hyperplane indicator function, $i_F(\mathbf{x}, \mathbf{w}) = \text{sign}(u)$, is equal to the number of unknown parameters that are elements of the weight vector $w = [w_0 w_1 \dots w_n]$.
- It's a coincidence and the VC dimension does not necessarily increase with the number of weights vector parameters.
 - Example: the indicator function $i_F(\mathbf{x}, \mathbf{w}) = \text{sign}(\sin(wx))$, $w, x \in \mathfrak{R}$, has an infinite VC dimension.



VC Dimension of a Loss Function

- The VC dimension of an specific loss function

$$L[y, f_a(\mathbf{x}, \mathbf{w})]$$

is equal to the VC dimension of the approximating function $f_a(\mathbf{x}, \mathbf{w})$ for both, classification and regression tasks.

VC Dimension for Radial Basis Functions (RBFs)

- For regression, the VC dimension of a set of RBFs as given by

$$f_a(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^N w_i \varphi_i(\mathbf{x}) + w_0$$

is equal to $h = N + 1$, where N is the number of hidden layer neurons.



VC Dimension for other functions

- For nonlinear functions, calculate the VC dimension is a very difficult task, if possible at all.
- Even, in the simple case of the sum of two basis functions, each having a finite VC dimension, the VC dimension of the sum can be infinite.



Outline

1 Introduction

- A brief history of the Learning Problem
- Vapnik-Chervonenkis (VC) dimension
- **Structural Risk Minimization (SRM) Inductive Principle**

2 Support Vector Machines (SVM)

- Optimal hyperplane for linearly separable patterns
- Optimal hyperplane for non-separable patterns
- SVMs for pattern recognition

Controlling the generalization ability of learning processes

- Construct an inductive principle for minimizing the risk functional using a small sample of training instances.
- The sample size l is considered to be small if the ratio l/h is small, say $l/h < 20$, where h is the VC dimension of functions of a learning machine.



Bounds for the generalization ability of LM

- To construct small sample size methods, two bounds can be used that hold with probability $1 - \eta$, $0 \leq \eta \leq 1$:
 - With sets of totally bounded non-negative functions:

$$R(\alpha_l) \leq R_{emp}(\alpha_l) + \frac{B\varepsilon}{2} \left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha_l)}{B\varepsilon}} \right)$$

- With sets of unbounded functions:

$$R(\alpha_l) \leq \frac{R_{emp}(\alpha_l)}{(1 - a(p)\tau\sqrt{\varepsilon})_+}, \quad a(p) = \sqrt{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}$$

where

$$\varepsilon = 4 \frac{h \left(\ln \left(\frac{2l}{h} \right) + 1 \right) - \ln(\eta/4)}{l}$$

Generalization bound for binary classification

- For binary classification, the bound above simplifies to:

$$R(\alpha_l) \leq R_{emp}(\alpha_l) + \sqrt{\frac{h(\log(\frac{2l}{h}) + 1) - \log(\eta/4)}{l}}$$

- The left part of the inequality is the actual risk, the right part is the risk bound.
- The risk bound is composed of the sum of the empirical risk and a VC confidence.

Empirical risk minimization principle

- The ERM principle is intended for dealing with large sample sizes.
- When l/h is large, ε is small. Therefore, the VC confidence is small.
- The actual risk is then close to the value of empirical risk.
- A small value of the empirical risk guarantees a small value of the expected risk.

Small sample size

- If l/h is small, a small $R_{emp}(\alpha_l)$ does not guarantee a small value of the actual risk.
- To minimize the actual risk, minimization have to be done simultaneously over both terms: empirical risk and VC confidence.
- The Structural Risk Minimization (SRM) principle, is intended to minimize the risk functional with respect to both terms, making the VC dimension a controlling variable.

Structures

- Let S be a set of functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, which is provided with an structure consisting of nested subsets of functions $S_k = \{Q(\mathbf{z}, \alpha), \alpha \in \Lambda_k\}$ such that:

$$S_1 \subset S_2 \subset \dots \subset S_n \dots$$

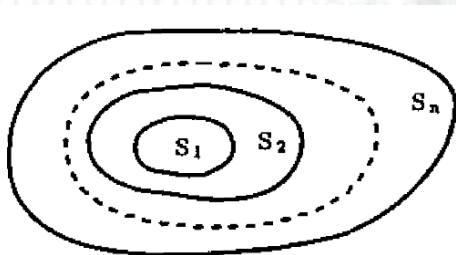


Figure: An structure on the set of functions is determined by the nested subsets of functions.

Admissible structure

- An structure is *admissible* if satisfies the following two conditions:

- 1 The VC dimension h_k of each set S_k of functions is finite:

$$h_1 \leq h_2 \leq \dots \leq h_n \dots$$

- 2 Any element S_k of the structure contains either

- a set of totally bounded functions:

$$0 \leq Q(\mathbf{z}, \alpha) \leq B_k, \quad \alpha \in \Lambda_k$$

- or a set of functions satisfying for some pair (p, τ_k) the inequality:

$$\sup_{\alpha \in \Lambda_k} \frac{(\int Q^p(\mathbf{z}, \alpha) dF(\mathbf{z}))^{\frac{1}{p}}}{\int Q(\mathbf{z}, \alpha) dF(\mathbf{z})} \leq \tau_k, \quad p > 2$$



SRM induction principle

- For a given set of observations $\mathbf{z}_1, \dots, \mathbf{z}_l$ the SRM principle chooses the functions $Q(\mathbf{z}, \alpha_l^k)$ minimizing the empirical risk in the subset S_k for which the guaranteed risk is minimal.
- The SRM principle defines a trade-off between the quality of the approximation of the given data and the complexity of the approximation function.
- As the subset index n increases, the minima of the empirical risks decreases, however, the term responsible for the confidence interval increases.



SRM principle illustration

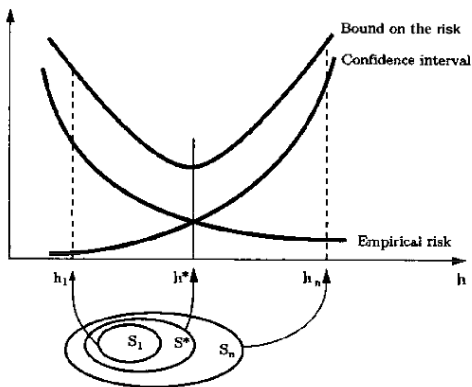


Figure: The bound on the risk is the sum of the empirical risk and the confidence interval. The smallest bound of the risk is achieved on some appropriate element of the structure.

Outline

- 1 Introduction
 - A brief history of the Learning Problem
 - Vapnik-Chervonenkis (VC) dimension
 - Structural Risk Minimization (SRM) Inductive Principle
- 2 Support Vector Machines (SVM)
 - **Optimal hyperplane for linearly separable patterns**
 - Optimal hyperplane for non-separable patterns
 - SVMs for pattern recognition

Binary classification problem definition

- Given a training data set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, $\mathbf{x} \in \mathcal{X}^n$, $y \in \{+1, -1\}$.
- It's assumed that the data are linearly separable.
- The equation of a decision surface in the form of an hyperplane that does the separation is

$$\mathbf{w}^T \mathbf{x} + b = 0 \quad (1)$$

where \mathbf{w} is an adjustable weight vector and b is a bias.

- Under this considerations the optimal separating function must be found without knowing the underlying probability distribution $F(\mathbf{x}, y)$.

Optimal hyperplane

- For a given weight vector \mathbf{w} and bias b , the separation between the hyperplane defined in (1) and the closest data point is called the *margin of separation* and denoted by ρ .
- The goal of SVM is to find among all the hyperplanes that minimize the training error (empirical risk), the particular one that maximizes the margin of separation. This hyperplane is called the *optimal hyperplane*.

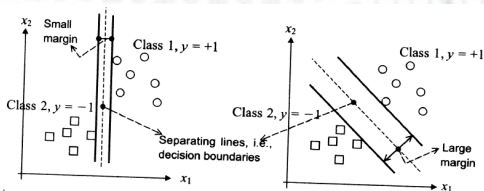


Figure: Two out of separating lines. Right: a good one with a large margin. Left: a less acceptable one with an small margin.

Problem definition

- The issue at hand is to find the parameters \mathbf{w}_o and b_o for the optimal hyperplane given the training set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, $\mathbf{x} \in \mathcal{R}^n$, $y \in \{+1, -1\}$.
- The pair (\mathbf{w}_o, b_o) must satisfy the constraints:

$$\begin{cases} \mathbf{w}_o^T \mathbf{x}_i + b_o \geq 1 & \text{for } y_i = +1 \\ \mathbf{w}_o^T \mathbf{x}_i + b_o \leq -1 & \text{for } y_i = -1 \end{cases} \quad (2)$$

- The particular data points $(\mathbf{x}_i^{(s)}, y_i^{(s)})$ for which one of the constraints is satisfied with the equality sign are called *support vectors*.

Discriminant function, indicator function and decision boundary

- The discriminant function (3) gives an algebraic measure of the distance from \mathbf{x} to the hyperplane defined by (\mathbf{w}, b) .

$$g(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b \quad (3)$$

- The indicator function (4) represents a learning or support vector machine's output.

$$i_F(\mathbf{x}, \mathbf{w}, b) = \text{sign}(g(\mathbf{x}, \mathbf{w}, b)) \quad (4)$$

- Both, the discriminant function and the indicator function, lie in an $(n + 1)$ -dimensional space.
- The decision boundary is an intersection of $g(\mathbf{x}, \mathbf{w}, b)$ and the input space \mathcal{X}^n .

Discriminant function, indicator function and decision boundary

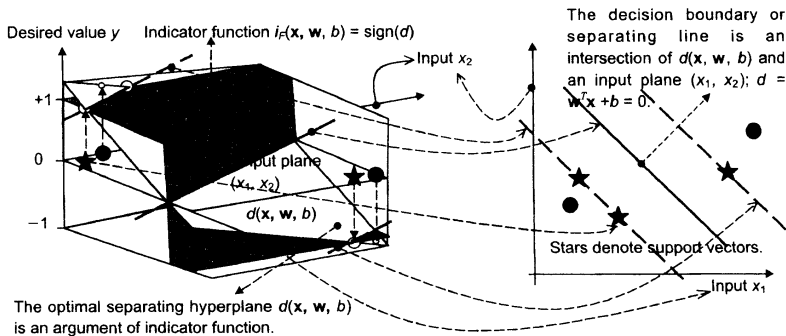


Figure: Discriminant function, indicator function and decision boundary illustration

The algebraic distance

- We have seen that the discriminant function gives an algebraic measure of the distance from \mathbf{x} to the hyperplane defined by (\mathbf{w}, b) .
- \mathbf{x} can be expressed as

$$\mathbf{x} = \mathbf{x}_p + r \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

where \mathbf{x}_p is the normal projection of \mathbf{x} onto the hyperplane, and r is the desired algebraic distance.

- Since, by definition, $g(\mathbf{x}_p) = 0$, it follows that

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = r \|\mathbf{w}\| \quad \text{or} \quad r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

The algebraic distance

Illustration

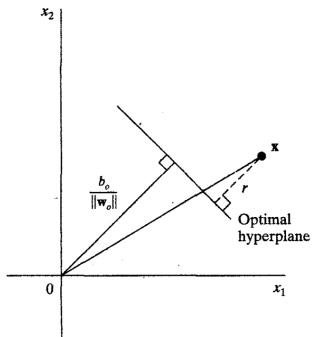


Figure: Geometric interpretation of algebraic distance of points to the optimal hyperplane for a two-dimensional case.



SVM's induction principle for the two separable class problem

- The algebraic distance from the support vector $\mathbf{x}^{(s)}$ to the optimal hyperplane is

$$r = \frac{g(\mathbf{x}^{(s)})}{\|\mathbf{w}_o\|} = \begin{cases} \frac{1}{\|\mathbf{w}_o\|} & \text{if } i_F(\mathbf{x}^{(s)}, \mathbf{w}_o, b_o) = +1 \\ -\frac{1}{\|\mathbf{w}_o\|} & \text{if } i_F(\mathbf{x}^{(s)}, \mathbf{w}_o, b_o) = -1 \end{cases}$$

- Let ρ denote the optimum value of the margin of separation between the two classes that constitute the training set. It follows that

$$\rho = 2r = \frac{2}{\|\mathbf{w}_o\|} \quad (5)$$

- Equation (5) states that maximizing the margin of separation between classes is equivalent to minimizing the euclidean norm of the weight vector.

The primal problem

- Given the training sample $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, $\mathbf{x} \in \mathfrak{R}^n$, $y \in \{+1, -1\}$, find the optimum values of the weight vector \mathbf{w} and bias b such that they satisfy the constraints

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad \text{for } i = 1, \dots, l$$

and the weight vector \mathbf{w} minimizes the cost function

$$\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

- This constrained optimization problem is called the *primal problem* and it's characterized as follows:
 - The cost function $\Phi(\mathbf{w})$ is a convex function of \mathbf{w} .
 - The constraints are linear in \mathbf{w} .

Method of Lagrange multipliers

- The primal problem can be solved using the method of Lagrange multipliers. The Lagrange function is defined as

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^l \alpha_i [y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1] \quad (6)$$

where the auxiliary non-negative variables α_i are called *Lagrange multipliers*.

- The solution to the primal problem is determined by the *saddle point* of the Lagrangian function $J(\mathbf{w}, b, \alpha)$ which has to be minimized with respect to \mathbf{w} and b , and maximized respect to α .

Conditions of optimality

- Differentiating (6) with respect to \mathbf{w} and b and setting the result equal to zero, the following two conditions of optimality are gotten:

$$\text{Condition 1 : } \frac{\partial J(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = 0$$

$$\text{Condition 2 : } \frac{\partial J(\mathbf{w}, b, \alpha)}{\partial b} = 0$$

- Application of condition 1 and condition 2 to the Lagrangian function (6) yields:

$$\mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i \quad \text{and} \quad \sum_{i=1}^l \alpha_i y_i = 0 \quad (7)$$

Considerations about the primal problem

- The solution vector \mathbf{w} is unique by virtue of the convexity of the Lagrangian function but the Lagrange multipliers α_i are not.
- At the saddle point, the product of each Lagrangian multiplier with its corresponding constraints vanishes:

$$\alpha_i [y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1] = 0 \quad \text{for } i = 1, 2, \dots, l \quad (8)$$

- Therefore, only multipliers exactly meeting Eq. (8) can assume non-zero values (Kuhn-Tucker conditions of optimization theory).

The dual problem

- Equivalent to the primal problem, but here the optimal solution is provided by the Lagrange multipliers.
- Duality theorem:
 - If the primal problem has an optimal solution, the dual problem has also an optimal solution, and both optimal values are equal.
 - In order for \mathbf{w}_o to be an optimal primal solution and α_o to be an optimal dual solution, it's necessary and sufficient that \mathbf{w}_o is feasible for the primal problem, and

$$\Phi(\mathbf{w}_o) = J(\mathbf{w}_o, b_o, \alpha_o) = \min_{\mathbf{w}} J(\mathbf{w}, b_o, \alpha_o)$$

Dual problem postulate

- Expanding Eq. (6), term by term, as follows:

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^l \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^l \alpha_i y_i + \sum_{i=1}^l \alpha_i$$

and applying optimality conditions (7), $J(\mathbf{w}, b, \alpha)$ can be reformulated as:

$$Q(\alpha) = J(\mathbf{w}, b, \alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

subject to the constraints:

$$\sum_{i=1}^l \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0 \quad \text{for } i = 1, 2, \dots, l$$

Computing \mathbf{w}_o and b_o

- Having determined the optimum Lagrange multipliers, denoted as $\alpha_{o,i}$, \mathbf{w}_o and b_o are computed by:

$$\mathbf{w}_o = \sum_{i=1}^l \alpha_{o,i} y_i \mathbf{x}_i$$

$$b_o = 1 - \mathbf{w}_o^T \mathbf{x}^{(s)} \quad \text{for } y^{(s)} = +1$$

Statistical properties

- The VC dimension of a learning machine determines the way in which a nested structure of approximating functions should be used.
- The VC dimension of a set of separating hyperplanes in a space of dimensionality m is equal to $h = m + 1$.
- In order to apply the method of structural risk minimization there is a need to construct a set of separating hyperplanes of varying VC dimension such that the empirical risk and the VC dimension are both minimized at the same time.

SRM for the optimal hyperplane problem (I)

Theorem

Let D denote the diameter of the smallest ball containing all the input vectors x_1, \dots, x_l . The set of optimal hyperplanes described by the equation

$$\mathbf{w}_0^T \mathbf{x} + b_0 = 0$$

has a VC dimension h bounded from above as

$$h \leq \min \left\{ \left\lceil \frac{D^2}{\rho^2} \right\rceil, m_0 \right\} + 1$$

where m_0 is the dimensionality of the input space.

- So, there must be exercised control over the VC dimension, independently of the dimensionality m_0 of the input space, by properly choosing the margin of separation ρ .

SRM for the optimal hyperplane problem (II)

- Suppose there is a nested structure described in terms of the separating hyperplanes as follows:

$$S_k = \left\{ w^T x + h : \|w\|^2 \leq c_k \right\}, \quad k = 1, 2, \dots$$

this can be reformulated as:

$$S_k = \left\{ \left[\frac{r^2}{\rho^2} \right] + 1 : \rho^2 \geq a_k \right\}, \quad k = 1, 2, \dots$$

where c_k, a_k are constants.

- Using the optimal hyperplane the SRM requirements can be satisfied.

Outline

- 1 Introduction
 - A brief history of the Learning Problem
 - Vapnik-Chervonenkis (VC) dimension
 - Structural Risk Minimization (SRM) Inductive Principle
- 2 Support Vector Machines (SVM)
 - Optimal hyperplane for linearly separable patterns
 - **Optimal hyperplane for non-separable patterns**
 - SVMs for pattern recognition

Introduction

- Given a training data set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, $\mathbf{x} \in \mathcal{R}^n$, $y \in \{+1, -1\}$.
- It's assumed that the data are NOT linearly separable.
- Given such a set of training data, it is not possible to construct a separating hyperplane without encountering classification errors.
- The margin of separation between classes is said to be *soft* if a data point (\mathbf{x}_i, y_i) violates the following restriction:

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \dots, l \quad (9)$$

Soft margin of separation

- Violation of (9) arises in one of two ways:
 - ① The data point falls inside the region of separation but on the right side of the decision surface (correct classification).
 - ② The data point falls on the wrong side of the decision surface (misclassification).

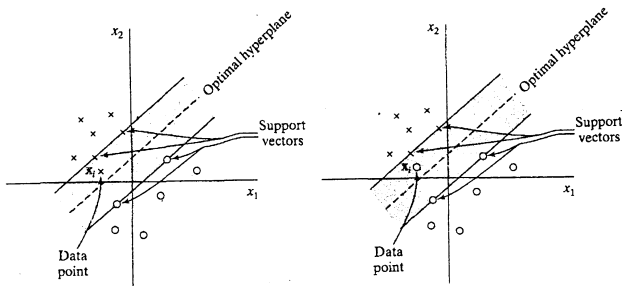


Figure: Left: violation case 1. Right: violation case 2.



Slack variables

- To formally treat non-separable data, a new set of non-negative scalar variables, $\{\xi_i\}_{i=1}^l$, called *slack variables*, are introduced into the definition of the separation hyperplane:

$$y_i (w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, l \quad (10)$$

- The slack variables measure the deviation of a data point from the ideal condition of pattern separability:
 - For $0 \leq \xi_i \leq 1$, the data point falls inside the region of separation but on the right side of the decision surface.
 - For $\xi_i > 1$, it falls on the wrong side of the separation hyperplane.
 - Support vectors are those data points that satisfy (10) even if $\xi_i > 0$.

Problem definition

- The goal is to find a separating hyperplane for which the missclassification error, averaged on the training set, is minimized.
- This is done by minimizing the functional

$$\Phi(\xi) = \sum_{i=1}^l I(\xi_i - 1)$$

with respect to the weight vector \mathbf{w} , subject to the constraints on $\|\mathbf{w}\|^2$ and (10). $I(\xi)$ is an indicator function defined by:

$$I(\xi) = \begin{cases} 0 & \text{if } \xi \leq 0 \\ 1 & \text{if } \xi > 0 \end{cases}$$

- Non-convex optimization problem: NP-complete.



Approximation

- Mathematically tractable approximation:

$$\Phi(\xi) = \sum_{i=1}^l \xi_i$$

- Moreover, the functional is simplified by formulating it to be minimized with respect to the weight vector \mathbf{w} :

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i \quad (11)$$

where C is a regularization parameter.

- Minimizing the first term of (11) is related to minimizing the VC dimension on the SVM. The second term on (11) is an upper bound on the number of test errors.

Optimization problem

- The optimization problem for non-separable patterns includes the optimization problem for linearly separable patterns as a special case where:

$$\xi_i = 0, \forall i$$



Primal problem

- Given the training sample $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, find the optimum values of the weight vector \mathbf{w} and bias b such that they satisfy the constraint

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \text{for } i = 1, 2, \dots, l$$

$$\xi_i \geq 0 \quad \forall i$$

and such that the weight vector \mathbf{w} and the slack variables ξ_i minimize the cost functional

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i$$

where C is an user-specific positive parameter.

Dual problem

- Given the training sample $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, find the Lagrange multipliers $\{\alpha_i\}_{i=1}^l$ that maximize the objective function

$$Q(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

subject to the constraints:

$$\sum_{i=1}^l \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \quad \text{for } i = 1, 2, \dots, l$$

where C is an user-specific positive parameter.



Computing \mathbf{w}_0

- The optimum solution for the weight vector is given by

$$\mathbf{w}_0 = \sum_{i=1}^{N_s} \alpha_{o,i} y_i \mathbf{X}_i$$

where N_s is the number of support vectors.

Computing b_0

- The Kunn-Tucker conditions are now defined by:

$$\alpha_i [y_i (w^T x_i + b) - 1 + \xi_i] = 0 \quad i = 1, 2, \dots, l \quad (12)$$

$$\mu_i \xi_i = 0 \quad i = 1, 2, \dots, l \quad (13)$$

where the μ_i are Lagrange multipliers that have been introduced to enforce the non-negativity of the slack variable ξ_i .

- At the saddle point the derivative of the Lagrangian function for the primal problem with respect to the slack variable ξ_i is zero, which yields:

$$\alpha_i + \mu_i = C \quad (14)$$



Computing b_0

- Combining (13) and (14) yields:

$$\xi_i = 0 \quad \text{if} \quad \alpha_i < C$$

- The optimum bias b_0 can be determined by taking any data point (\mathbf{x}_i, y_i) in the training set for which $0 < \alpha_{o,i} < C$ and therefore $\xi_i = 0$, and using the data point in (12).
- From a numerical perspective is better to take the mean value of b_0 resulting from all such data points in the training sample.

Outline

- 1 Introduction
 - A brief history of the Learning Problem
 - Vapnik-Chervonenkis (VC) dimension
 - Structural Risk Minimization (SRM) Inductive Principle
- 2 Support Vector Machines (SVM)
 - Optimal hyperplane for linearly separable patterns
 - Optimal hyperplane for non-separable patterns
 - SVMs for pattern recognition



SVM's idea

- 1 Non-linear mapping of an input vector into a high-dimensional feature space that is hidden from both, the input and the output.
 - 2 Construction of an optimal hyperplane for separating the features discovered.
- Operation 1 is performed in accordance with Cover's theorem on the separability of patterns: a multidimensional space may be transformed into a new feature space where the patterns are linearly separable with high probability, providing that the transformation is non-linear and that the dimensionality of the feature space is high enough.
 - However, it's only that by using an optimal separating hyperplane, the VC dimension is minimized and generalization is achieved.

Non-linear transformation

- Let \mathbf{x} denote a vector drawn from the input space, assumed to be of dimension m_0 .
- Let $\{\varphi_j(\mathbf{x})\}_{j=1}^{m_1}$ denote a set of non-linear transformations from the input space to the feature space with dimension m_1 .
- Given such a set of non-linear transformations, an hyperplane acting as a decision surface can be defined as:

$$\sum_{j=1}^{m_1} w_j \varphi_j(\mathbf{x}) + b = 0 \quad (15)$$

Decision surface

- Define the vector

$$\boldsymbol{\varphi}(\mathbf{x}) = [\varphi_0(\mathbf{x}), \varphi_1(\mathbf{x}), \dots, \varphi_{m_1}(\mathbf{x})]^T$$

where, by definition, $\varphi_0(\mathbf{x}) = 1$, for all \mathbf{x} .

- $\boldsymbol{\varphi}(\mathbf{x})$ represents the “image” induced in the feature space due to the input vector \mathbf{x} .
- In terms of this image the decision surface (15) can be defined in a more compact form:

$$\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = 0$$

Lagrange conditions

- Adapting Lagrange conditions (7) to the present situation involving a feature space where “linear” separability of patterns can be seen, it follows:

$$\mathbf{w} = \sum_{i=1}^l \alpha_i y_i \varphi(\mathbf{x}_i) \quad (16)$$

- Substituting (16) in (15) yields:

$$\sum_{i=1}^l \alpha_i y_i \varphi^T(\mathbf{x}_i) \varphi(\mathbf{x}) = 0$$



The inner-product kernel

- The term $\phi^T(\mathbf{x}_i) \phi(\mathbf{x})$ represents the inner product of two vectors induced in the feature space by the input vector \mathbf{x} and the input pattern \mathbf{x}_i pertaining to the i th sample.
- Let's introduce the *inner-product kernel* denoted by $K(\mathbf{x}, \mathbf{x}_i)$ and defined by:

$$K(\mathbf{x}, \mathbf{x}_i) = K(\mathbf{x}_i, \mathbf{x}) = \phi^T(\mathbf{x}) \phi(\mathbf{x}_i), \quad \text{for } i = 1, 2, \dots, l$$

- The inner-product kernel can be used to construct the optimal hyperplane in the feature space without having to consider the feature space itself in explicit form (Kernel trick):

$$\sum_{i=1}^l \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) = 0$$

Mercer's condition

Theorem

For a mapping ϕ and an symmetric expansion, defined in the closed interval $\mathbf{a} \leq \mathbf{x}, \mathbf{x}' \leq \mathbf{b}$ as

$$K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}'), \quad \lambda_i < 0$$

to be valid and for it to converge absolutely and uniformly, it is necessary and sufficient that the condition

$$\int_b^a \int_b^a K(\mathbf{x}, \mathbf{x}') \psi(\mathbf{x}) \psi(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

holds for any $\psi(\cdot)$ such that

$$\int_b^a \psi^2(\mathbf{x}) d\mathbf{x} < \infty$$

Mercer's theorem observations

- The functions $\varphi_i(\mathbf{x})$ are called eigenfunctions of the expansion and the scalars λ_i are called eigenvalues.
- the fact that all of the eigenvalues are positive means that the kernel $K(\mathbf{x}, \mathbf{x}')$ is positive definite.
- In theory, the dimensionality of the feature space can be infinitely large.
- Mercer's theorem only tells whether or not a candidate kernel is actually an inner-product kernel in some space and therefore admissible for its use in SVM.
- But it says nothing about how to construct the functions $\varphi_i(\mathbf{x})$.



Inner-product kernels examples

- Polynomial learning machine:

$$K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T \mathbf{x}_i + 1)^p$$

where p is given a priori.

- Radial-basis function network:

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}_i\|^2\right)$$

where σ is given a priori.

- Two-layer perceptrons:

$$K(\mathbf{x}, \mathbf{x}_i) = \tanh(\beta_0 \mathbf{x}^T \mathbf{x}_i + \beta_1)$$

where Mercer's conditions are satisfied only for some values of β_0 and β_1 .

Optimum design of a SVM

- Giving the training sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^l$ find the Lagrange multipliers $\{\alpha_i\}_{i=1}^l$ that maximize the objective function:

$$Q(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to the constraints:

$$\sum_{i=1}^l \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \quad \text{for } i = 1, 2, \dots, l$$








Computing \mathbf{w}_o and b_o

- Having found the optimum values of Lagrange multipliers, denoted by $\alpha_{o,i}$, the corresponding optimum value of the linear weight vector is given by:

$$\mathbf{w}_o = \sum_{i=1}^l \alpha_{o,i} y_i \boldsymbol{\varphi}(\mathbf{x}_i)$$

where $\boldsymbol{\varphi}(\mathbf{x}_i)$ is the image induced in the feature space due to \mathbf{x}_i and the first component of \mathbf{w}_o represent the optimum bias b_o .

For Further Reading

-  The Nature of Statistical Learning Theory. Vladimir N. Vapnik. ISBN: 0-387-98780-0. 1995.
-  Statistical Learning Theory. Vladimir N. Vapnik. ISBN: 0-471-03003-1. 1998.
-  A tutorial on Support Vector Machines for Pattern Recognition. Christopher J. C. Burges. Data Mining and Knowledge Discovery, Vol.2, pp: 121-167. 1998.
-  Neural Networks: A Comprehensive Foundation, 2nd Edition. Simon Haykin. ISBN: 81-7808-300-0. 1999.
-  Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models. Vojislav Kecman. ISBN: 0-262-11255-8. 2001.

Questions?

Thank you very much for your attention.

- Contact:
 - Miguel Angel Veganzones
 - Grupo Inteligencia Computacional
 - Universidad del País Vasco - UPV/EHU (Spain)
 - E-mail: miguelangel.veganzones@ehu.es
 - Web page: <http://www.ehu.es/computationalintelligence>