# Support Vector Machines

## Miguel A. Veganzones

Grupo Inteligencia Computacional
Universidad del País Vasco

# Outline

# Outline

# 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 can generalize.
- The Perceptron was constructed for solve 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 and started 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 is 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.

- Example: overfitting is a problem of "false structure" (ill-posed problems) solved in statistical analysis by regularization techniques.

# 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 posseses 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 an 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

# Introduction

- The VC dimension is a property of a set of approximatting 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.
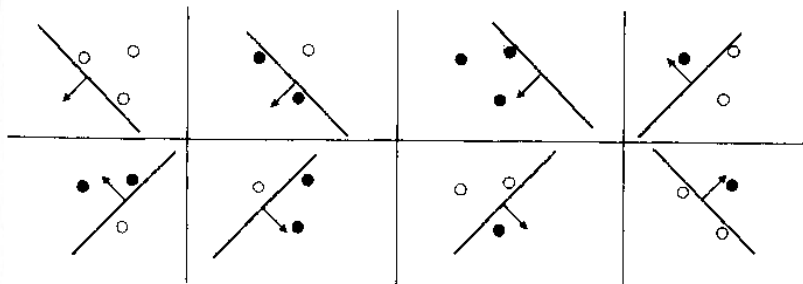
# Two-class pattern recognition case
## Indicator functions

- 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.
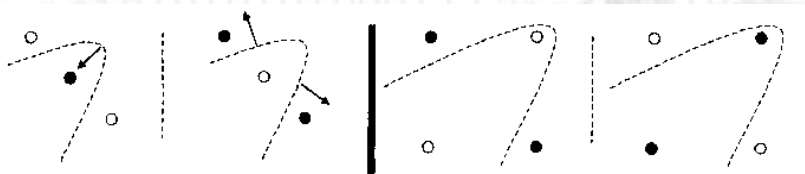
# Two-class pattern recognition case
## Possible ways in $\Re^2$



Figure: Three points in all possible $2^3 = 8$ ways by an indicator function $i_F(\mathbf{x}, \mathbf{w}) = sign(u) = sign(w_1 x_1 + w_2 x_2 + w_0)$ represented by the oriented straight line $u = 0$.

# Two-class pattern recognition case
## Labelings that cannot be shattered in $\Re^2$



Figure: Left: two labelings of a three co-linear points that cannot be shattered by $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$. Right: $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$ 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, $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$, 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 $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$.
  - In a two-dimensional set of inputs no set of four points can be shattered by $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$.

# 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, $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$, is equal to the number of unknown parameters that are elements of the weight vector $w = [w_0 w_1 \ldots w_n]$.

- It's a coincidence and the VC dimension does not necessarily increases with the number of weights vector parameters.

  - Example: the indicator function $i_F(\mathbf{x}, \mathbf{w}) = sign(\sin(wx))$, $w, x \in \Re$, 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 Linear Functions

- The VC dimension of a set of linear functions as given by

$$f_a(\mathbf{x}, \alpha) = \sum_{i=1}^{N} \alpha_i x_i + \alpha_0$$

is equal to $h = N + 1$, where $N$ is the dimensionality of the sample space.

# 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 dimnesion of the sum can be infinite.

# Outline

# 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 $l/h < 20$.
- To construct small sample size

# Outline

# Binary classification problem definition

- Given a training data set $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_l}, y_l)$, $\mathbf{x} \in \mathfrak{R}^n$, $y \in \{+1, -1\}$.

- 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

$$\mathbf{w^T x} + b = 0 \tag{1}$$

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 $\rho$.
- 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*.
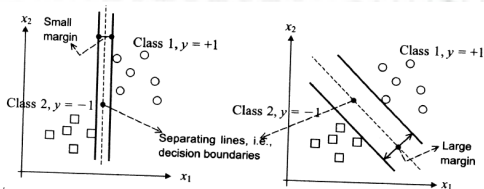


Figure: Two out of separating lines. Right: a good one with a large margin. Left: a less acedable one with a small margin.

## Problem definition

- The issue at hand is to find the parameters $\mathbf{w_0}$ and $b_0$ for the optimal hyperplane given the training set $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_l}, y_l)$, $\mathbf{x} \in \Re^n$, $y \in \{+1, -1\}$.

- The pair $(\mathbf{w_0}, b_0)$ must satisfy the constraints:

$$\begin{cases} \mathbf{w_0^T x_i} + b_0 \geq 1 & for \quad y_i = +1 \\ \mathbf{w_0^T x_i} + b_0 \leq 1 & for \quad y_i = -1 \end{cases} \tag{2}$$

- The particular data points $\left(\mathbf{x_i^{(s)}}, y_i^{(s)}\right)$ 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 $(\mathbf{w}, b)$.

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

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

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

- 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 $\Re^n$.

# Discriminant function, indicator function and decision boundary
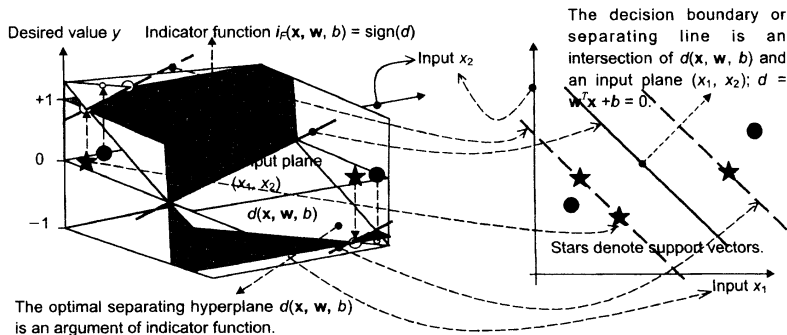


Figure: Discriminant function, indicator function and decision boundary illustration

# 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

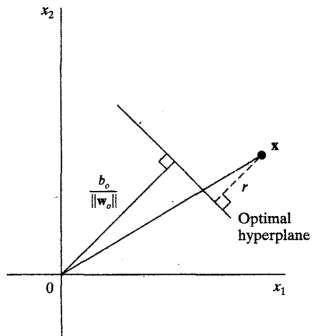$$\mathbf{x} = \mathbf{x_p} + r \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

  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

$$g(\mathbf{x}) = \mathbf{w^T}\mathbf{x} + b = r \|\mathbf{w_0}\| \qquad or \qquad r = \frac{g(\mathbf{x})}{\|\mathbf{w_0}\|}$$

# The algebraic distance
## Illustraton



Figure: Geometric iterpretation 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 $\mathbf{x}^{(\mathbf{s})}$ to the optimal hyperplane is

$$r = \frac{g\left(x^{(s)}\right)}{\|w_0\|} = \begin{cases} \frac{1}{\|w_0\|} & if \quad i_F\left(\mathbf{x}^{(\mathbf{s})}, \mathbf{w_0}, b_0\right) = +1 \\ -\frac{1}{\|w_0\|} & if \quad i_F\left(\mathbf{x}^{(\mathbf{s})}, \mathbf{w_0}, b_0\right) = -1 \end{cases}$$

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

$$\rho = 2r = \frac{2}{\|\mathbf{w_0}\|} \tag{5}$$

- 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 $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_l}, y_l)$, $\mathbf{x} \in \Re^n$, $y \in \{+1, -1\}$, find the optimum values of the weight vector $w$ and bias $b$ such that they satisfy the constraints

$$y_i \left( \mathbf{w^T x_i} + b \right) \geq 1 \qquad for \quad i = 1, \ldots, l$$

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

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

- 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

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

  where the auxiliary non-negative variables $\alpha_i$ are called *Lagrange multipliers*.

- The solution to the primal problem is determined by he *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 $\alpha$.

# Conditions of optimality

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

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

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

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

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x_i} \qquad and \qquad \sum_{i=1}^{l} \alpha_i y_i = 0$$

# 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
- SVMs for regression

# Outline

# Outline

# 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.

📑 Neural Networks: A Comprenhesive Foundation, 2$^{nd}$ Edition. Simon Haykin. ISBN: 81-7808-300-0. 1999.

📑 Learning and Soft Computing: Support Vector Machines, Neural Netowrks 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