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)

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)

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

1. Introduction
   - A brief history of the Learning Problem
   - Vapnik-Chervonenkis (VC) dimension
   - Structural Risk Minimization (SRM) Inductive Principle
2. Support Vector Machines (SVM)

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 $\Re^2$

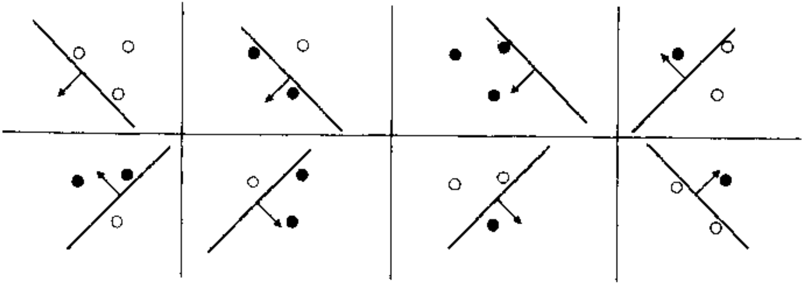

Figure: Three points in all possible $2^3 = 8$ ways by an indicator function $i_F(\mathbf{x}, \mathbf{w}) = sign(u) = sign(w_1 x_1 + w_2 x_2 + w_0)$ represented by the oriented straight line $u = 0$.

Two-class pattern recognition case

Labelings that cannot be shattered in $\Re^2$

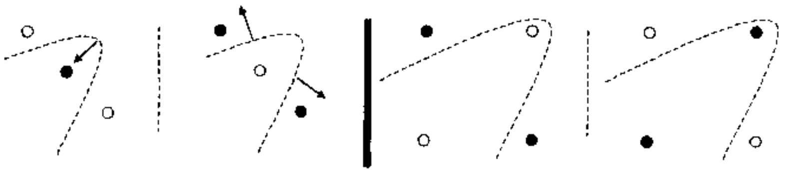

Figure: Left: two labelings of a three co-linear points that cannot be shattered by $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$. Right: $i_F(\mathbf{x}, \mathbf{w}) = sign(u)$ cannot shatter the depicted two out of sixteen labelings of four points. A quadratic indicator function (dashed line) can easily 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 $w = [w_0 w_1 \dots w_n]$.
- 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

1. Introduction
   - A brief history of the Learning Problem
   - Vapnik-Chervonenkis (VC) dimension
   - Structural Risk Minimization (SRM) Inductive Principle
2. Support Vector Machines (SVM)

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 Comprehensive Foundation, 2nd 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