Support Vector Machines

Wien, June , 2010

Paul Hofmarcher, Stefan Theussl,
WU Wien

Contents

- ✔ Linear Separable
  - ✘ Separating Hyperplanes
- ✔ Non-Linear Separable
  - ✘ Soft-Margin Hyperplanes
  - ✘ Support Vector Machines

Motivation

Support Vector Machines (SVM) are learning method for (binary) classification (Vapnik, 1979,1995).

✔ find hyperplane which separates data perfectly into two classes.

✔ Data is often not linearly separable:

✔ map data into higher dimensional space.

✔ kernel-trick and kernel induced feature space.

Contents

Motivation

Idea

Hyperplanes

Hyperplanes II

Optimal Separating Hyperplanes - LS

Non-linear separable Data

Suppot Vector Machines

Hyperplanes

- ✔ Given n trainingsdata $\{x_i, y_i\}$ for $i = 1, \ldots n$ with $x_i \in \mathbb{R}^k$ and $y_i \in \{-1, 1\}$
- ✔ Define a hyperplane by

$$L = \{x : f(x) = x^T\beta + \beta_0 = 0\}. \tag{1}$$

- ✔ A classification rule induced by $f(x)$ is given by

$$G(x) = sign[x^T\beta + \beta_0]$$

Contents

Motivation

Idea

Hyperplanes

Hyperplanes II

Optimal Separating Hyperplanes - LS

Non-linear separable Data

Suppot Vector Machines

Hyperplanes II

$f(x)$ in Equation 1 gives the signed distance form a point x to hyperplane

- ✔ For two points x_1 and x_2 in L $\beta^T(x_1 - x_2) = 0$ and hence $\beta^* = \beta / \|\beta\|$ and hence is the vector normal to the surface of L.
- ✔ for any point x_0 in L, we have $\beta^T x_0 = -\beta_0$.
- ✔ The signed distance of any point x to L

$$\beta^{*T}(x - x_0) = \frac{1}{\|\beta\|}(\beta^T x + \beta_0) = \frac{1}{\|f'(x)\|} f(x). \qquad (2)$$

Hence $f(x)$ is proportional to the signed distance from x to hyperplane $f(x) = 0$.

linear separable I

Contents

Motivation

Idea

Optimal Separating Hyperplanes - LS

linear separable I

linear separable II

linear separable III

linear separable IV

Non-linear separable Data

Suppot Vector Machines

Optimal separating hyperplane separates two classes and maximizes the distance to the closest point from either class.

$$\max_{\beta_0,\,\beta} C \quad s.t.\ y_i(x_i^T\beta + \beta_0) \geq C\,\|\beta\| \tag{3}$$

This is equivalent to

$$\min_{\beta_0,\,\beta} \frac{1}{2}\,\|\beta\|^2 \quad s.t.\ y_i(x_i^T\beta + \beta_0) \geq 1 \ \ \forall i \tag{4}$$

In light of 2 this defines a margin around linear decision boundary of thickness $1/\,\|\beta\|$ and is a convex optimization problem!

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
linear separable I
linear separable II
linear separable III
linear separable IV
Non-linear separable Data
Suppot Vector Machines

linear separable II

The according Lagrangian (primal) Function is:

$$L_P = \min_{\beta_0, \beta} \frac{1}{2} \|\beta\|^2 + \sum_{i=1}^{N} \alpha_i [y_i (x_i^T \beta + \beta_0) - 1], \tag{5}$$

with the derivatives:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \quad 0 = \sum_{i=1}^{N} \alpha_i y_i \tag{6}$$

and the so-called *Wolfe-dual*

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k \;\; s.t \; \alpha_i \geq 0 \tag{7}$$

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
linear separable I
linear separable II
linear separable III
linear separable IV
Non-linear separable Data
Suppot Vector Machines

linear separable III

Solution is obtained by maximizing L_D.

- ✔ must satisfy Karush-Kuhn-Tucker conditions ((6) and s.t.(7)) and

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - 1] = 0 \quad \forall i. \tag{8}$$

- ✔ if $\alpha_i > 0$ then $y_i(x_i^T\beta + \beta_0) = 1$ w.o.w x_i is on the boundary.
- ✔ if $y_i(x_i^T\beta + \beta_0) > 1$, x_i is not on the boundary and $\alpha_i = 0$

From 6 we see that the solution β is a linear combination of the *support points* x_i, those points defined to be on the boundary of the Hyperplane.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
linear separable I
linear separable II
linear separable III
linear separable IV
Non-linear separable Data
Suppot Vector Machines

linear separable IV

The optimal separating hyperplane produces a function $\hat{f}(x) = x^T\hat{\beta} + \hat{\beta}_0$ for classifying new observations:

$$\hat{G}(x) = sign \hat{f}(x). \tag{9}$$

The intuition is, that a large margin on the training data wil lead to good separation on the test data.

- ✔ Description in terms of support points suggests that the optimal hyperplane focuses more on points that count.
- ✔ The LDA, on the other hand depends on all of the data, even points that are far away from the decision boundery.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Support Vector Classifier
Support Vector Classifier II
non-separable case
non-separable case
Suppot Vector Machines

Support Vector Classifier

Discuss hyperplanes for cases, where the clases may not be separable by a linear boundery.

- ✔ training data consist of $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$.
- ✔ with a unit vector β, the hyperplane is defined as

$$\{x : f(x) = x^T\beta + \beta_0 = 0\} \tag{10}$$

- ✔ Classification rule is induced by $f(x)$,

$$G(x) = sign[x^T\beta + \beta_0] \tag{11}$$

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Support Vector Classifier
Support Vector Classifier II
non-separable case
non-separable case
Suppot Vector Machines

Support Vector Classifier II

As before, by dropping the norm constraint on β and defining $C := 1/\|\beta\|$ we get

$$min\|\beta\|, \quad s.t.\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i\ \forall i \quad, \xi_i \geq 0, \sum \xi_i \leq const. \tag{12}$$

The value ξ_i (slack variable) is the proportional amount by which the prediction is on the wrong side of its margin. Misclassifications occur when $\xi_i > 1$, so bounding $\sum \xi_i$ at a value K, bounds the total number of training misclassifications.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Support Vector Classifier
Support Vector Classifier II
non-separable case
non-separable case
Suppot Vector Machines

non-separable case

Equation 12 is a quadratic problem with linear constraints, hence a convex optimization problem. Equation 12 can be rewritten to

$$\min_{\beta_0,\beta} 0.5\|\beta\|^2 + \gamma \sum_i \xi_i, \quad s.t.\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \quad , \xi_i \geq 0 \forall i \quad , \tag{13}$$

where γ replaces the constant in 12. As before we want to derive the Lagrangian Wolfe-dual objective function:

$$L_D = \sum_i \alpha_i - 0.5 \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j, \tag{14}$$

which gives a lower bound on the objective function 12 for any feasible point. We maximize L_D subject to $0 \leq \alpha_i \leq \gamma$ and $\sum_i \alpha_i y_i = 0$.

non-separable case

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Support Vector Classifier
Support Vector Classifier II
non-separable case
non-separable case
Suppot Vector Machines

By using the Karush-Kuhn Tucker conditions we get a solution $\hat{\beta}$ of the form

$$\hat{\beta} = \sum_i \hat{\alpha}_i y_i x_i, \tag{15}$$

with nonzero coefficients α_i only for those observations i for which the constraints $y_i(x_i^T \beta + \beta_0) - (1 - \xi_i) = 0$. These observations are the *support vectors*, since $\hat{\beta}$ is represented in terms of them alone.

Given solution $\hat{\beta}, \hat{\beta}_0$, the decision function can be written as

$$\hat{G}(x) = sign[\hat{f}(x)] = sign[x^T \hat{\beta} + \hat{\beta}_0] \tag{16}$$

Idea of Support Vector Machines

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Suppot Vector Machines
Idea of Support Vector Machines
nonlinearity by preprocessing
Support Vector Machines
Support Vector Machines II
Hilbert Space
Kernels
Kernels II
RKHS

Methods described so far, find linear boundaries in the input feature space. SVMs are more flexible by enlarging the feature space using basis expansions such as splines. Linear boundaries in the enlarged space achieve better training-class separation.

- ✔ select basis functions $h_m(x), \quad m = 1, \ldots, M$
- ✔ fit the SV classifier using input features $h(x_i) = (h_1(x_i), \ldots, h_M(x_i)), \quad i = 1, \ldots, N$ and produce the (nonlinear) function $\hat{f}(x) = h(x)^T \hat{\beta} + \hat{\beta}_0$
- ✔ The classifier is $\hat{G}(x) = sign(\hat{f}(x))$
- ✔ SVMs allow the enlarged space to get very, large, even infinite dimension
- ✔ “Problem”: It might seem that computations would become prohibitive.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Suppot Vector Machines
Idea of Support Vector Machines
nonlinearity by preprocessing
Support Vector Machines
Support Vector Machines II
Hilbert Space
Kernels
Kernels II
RKHS

nonlinearity by preprocessing

✔ To make the SV algorithm nonlinear, this could be achieved by simply preprocessing the training patterns x_i by a map $h : \mathcal{X} \rightarrow \mathcal{F}$ into some feature space $\mathcal{F}$.

✔ **Example:** Consider a map $h : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ with $h(x_{i1}, x_{i2}) = (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2)$.

✔ Training a linear SV machine on the preprocessed features would yield a quadratic function.

✔ This approach can easily become computationally infeasible for both polynomial features of higher order or higher dimension.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Suppot Vector Machines
Idea of Support Vector Machines nonlinearity by preprocessing
Support Vector Machines
Support Vector Machines II
Hilbert Space
Kernels
Kernels II
RKHS

Support Vector Machines

The Lagrangian dual function of our new optimization problem gets the form:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle h(x_i) h(x_j) \rangle . \tag{17}$$

Analogous to the previous cases (separable, non-separable), the solution function $f(x)$ can be written as

$$f(x) = h(x)^T \beta + \beta_0 = \sum_i \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0 \tag{18}$$

As before, α_i, β_0 can be determined by solving $f(x_i) = 0$ for any x_i which $0 < \alpha_i < \gamma$

Support Vector Machines II

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Suppot Vector Machines
Idea of Support Vector Machines
nonlinearity by preprocessing
Support Vector Machines
Support Vector Machines II
Hilbert Space
Kernels
Kernels II
RKHS

So both, 17 and 18 involve $h(x)$ only through inner products. And in fact we do not need to specify the transformation $h(x)$ at all, but only require knowledge of the kernel function

$$k(x, x') = \langle h(x), h(x') \rangle, \tag{19}$$

that computes inner products in the transformed space.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Suppot Vector Machines
Idea of Support Vector Machines
nonlinearity by preprocessing
Support Vector Machines
Support Vector Machines II
Hilbert Space
Kernels
Kernels II
RKHS

Hilbert Space

- ✔ A Hilbert Space is essentially an infinite-dimensional Euclidean space. It is a vector space, i.e., closed under addition and scalar multiplication...
- ✔ It is endowed with an inner product $\langle \cdot, \cdot \rangle$, a bilinear form. From this inner product we get a norm $\| \cdot \|$ via $\|x\| = \sqrt{\langle x, x \rangle}$, which allows to define notions of convergence.
- ✔ Adding all limit points of Cauchy sequences to our space yields a Hilbert space.
- ✔ We will use kernel functions k to construct Hilbert Spaces.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Suppot Vector Machines
Idea of Support Vector Machines
nonlinearity by preprocessing
Support Vector Machines
Support Vector Machines II
Hilbert Space
Kernels
Kernels II
RKHS

Kernels

- ✔ Let $\mathcal{X}$ be a (non-empty) set. A mapping

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \ \ (x, x') \to k(x, x'), \tag{20}$$

is called a kernel if k is symmetric, i.e., $k(x, x') = k(x', x)$
- ✔ A kernel k is *positive definite*, if its Gram Matrix $K_{i,j} := k(x_i, x_j)$ is positive definite $\forall x$.
- ✔ The Cauchy-Schwarz inequality holds for p.d. kernels.
- ✔ Define a *reproducing kernel map:*

$$\Phi : x \to k(\cdot, x), \tag{21}$$

i.e., to each point x in the original space we associate a function $k(\cdot, x)$.

Kernels II

Contents

Motivation

Idea

Optimal Separating Hyperplanes - LS

Non-linear separable Data

Suppot Vector Machines

Idea of Support Vector Machines

nonlinearity by preprocessing

Support Vector Machines

Support Vector Machines II

Hilbert Space

Kernels

Kernels II

RKHS

✔ **Example:** Gaussian kernel. Each point x maps to a Gaussian distribution centered at that point.

$$k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \tag{22}$$

✔ Construct a vector space containing all linear combinations of functions $k(\cdot, x)$:

$$f(\cdot) = \sum_i \alpha_i k(\cdot, x_i). \tag{23}$$

This will be our RKHS.

Contents
Motivation
Idea
Optimal Separating Hyperplanes - LS
Non-linear separable Data
Suppot Vector Machines
Idea of Support Vector Machines
nonlinearity by preprocessing
Support Vector Machines
Support Vector Machines II
Hilbert Space
Kernels
Kernels II
RKHS

RKHS

- ✔ We now have to define an inner product in RKHS. Let $g(\cdot) = \sum_j \beta_j k(\cdot, x_j)$ and define:

$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j k(x_i, x'_j) \tag{24}$$

- ✔ One needs to verify that this is in fact an inner product (Symmetry, Linearity, $\langle f, f \rangle = 0 \rightarrow f = 0$).