Using

Support Vector Machines

to Classify Multidimensional Data

Václav Bartoš

Brno University of Technology, Faculty of Information Technology
Božetěchova 2, 612 00 Brno, CZ
ibartosv@fit.vutbr.cz

14. prosince 2011

Support Vector Machines

Support Vector Machines

Mathematical concept from area of machine learning.

- Used for:
  - classification
  - (regression analysis)
- Supervised learning method
  - Uses training data set
- Binary classifier
  - Classifies a set of data points into two classes

Classification

Definitions

- Data samples with D attributes are represented by D-dimensional vectors x.
    - Usually $x \in \mathbb{R}^D$
- C is a finite set of classes.
- Function $h : \mathbb{R}^D \to C$ assigns a class to every possible data sample x.
    - h is unknown, we want to find its approximation based on some training data.

Classification problem

Given a training data set $\{(\mathbf{x_1}, c_1), ..., (\mathbf{x}_N, c_N)\}$, where $c_i \in C$ and $c_i = h(x_i)$ for $i \in \{1, ..., N\}$, produce a function $\hat{h}$ which approximates h as close as possible.

Binary classification

In the case of SVM, there are two classes, $C = \{1, -1\}$.

Linear SVM

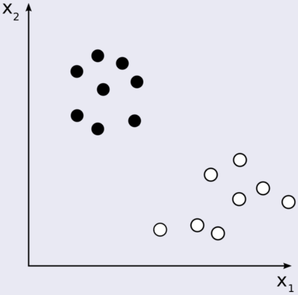
Linearity

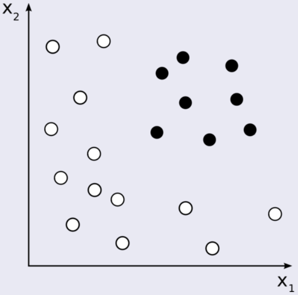
Basic SVM is a linear classifier – data must be linearly separable.

- There is some hyperplane (line for $D = 2$) separating the two classes.

SVM method finds such hyperplane and use it to classify new data points.

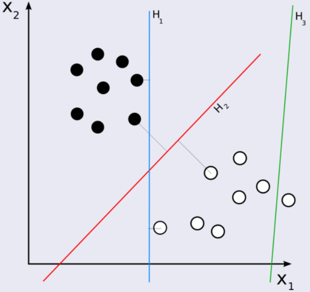
- A class for every new point is assigned according to the side of the hyperplane it lays on.





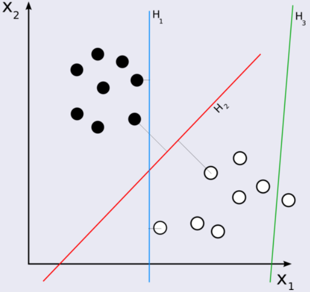
Linear SVM

Which hyperplane (line) is the best for classification?



Linear SVM

Which hyperplane (line) is the best for classification?



Maximum-margin hyperplane

The aim of SVM is to find the maximum-margin hyperplane, i.e. a hyperplane which has the greatest distance to the nearest points from both classes.

Linear SVM

Hyperplane

Set of points $\mathbf{x}$ satisfying:

$$\mathbf{x} \cdot \mathbf{w} - b = 0$$

where:

$\mathbf{w}$ is hyperplane's normal vector

$\frac{b}{\|\mathbf{w}\|}$ is distance of hyperplane from origin.

Margin

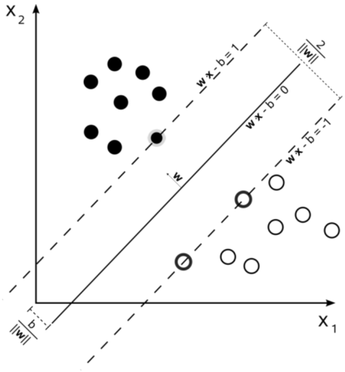
Defined by two parallel hyperplanes:

$$\mathbf{x} \cdot \mathbf{w} - b = 1$$

$$\mathbf{x} \cdot \mathbf{w} - b = -1.$$

Support vectors

Points closest to hyperplane – support vectors



Margin width maximization

Width of margin is $\frac{2}{\|\mathbf{w}\|}$ and we want to maximize it.

$$\text{maximize } \frac{2}{\|\mathbf{w}\|} \equiv \text{minimize } \|\mathbf{w}\| \equiv \text{minimize } \frac{1}{2}\|\mathbf{w}\|^2$$

Constraints to prevent data points from falling into the margin:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} - b &\geq 1 \quad \text{for } x_i \text{ from the first class } (c_i = 1) \\ \mathbf{x}_i \cdot \mathbf{w} - b &\leq -1 \quad \text{for } x_i \text{ from the second class } (c_i = -1). \end{aligned}$$

This can be simplified to:

$$c_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1 \quad \forall i \in \{1, ..., N\}$$

Linear SVM

Optimization problem

Minimize (in $\mathbf{w}$ and b)

$$\frac{1}{2}\|\mathbf{w}\|^2$$

such that

$$c_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1 \quad \forall i \in \{1, ..., N\}$$

Linear SVM

Optimization problem

Minimize (in $\mathbf{w}$ and b)

$$\frac{1}{2}\|\mathbf{w}\|^2$$

such that

$$c_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1 \quad \forall i \in \{1, ..., N\}$$

Primal form

Using Lagrange multipliers α, we get (minimize L_P):

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i c_i(\mathbf{x_i} \cdot \mathbf{w_i} - b) + \sum_{i=1}^{N} \alpha_i$$

Linear SVM

Dual form

We can also derive a dual form (maximize L_D):

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j c_i c_j \mathbf{x_i} \cdot \mathbf{x_j}$$
$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \alpha^T \mathbf{H} \alpha$$

where $H = \{h_{ij}\}_{i,j=1}^{N}$ and $h_{ij} = c_i c_j \mathbf{x_i} \cdot \mathbf{x_j}$
subject to:

$$\alpha_i \geq 0 \;\; \forall_i, \;\; \sum_{i=1}^{N} \alpha_i c_i = 0$$

This can be solved by standard quadratic programming techniques and programs.

Classification of new data point

Classification of new data point

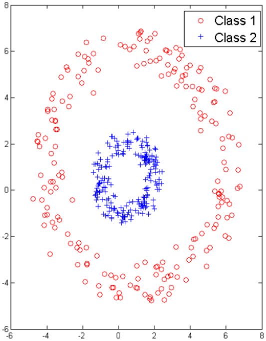
Once we have found the maximum-margin hyperplane (given by $\mathbf{w}$ and b), we can define classifier as:

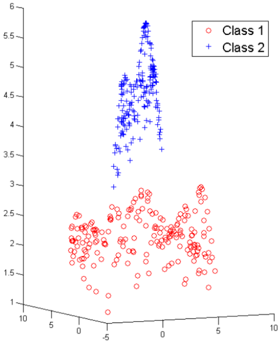
$$\hat{h}(\mathbf{x}) = sqn(\mathbf{w} \cdot \mathbf{x} - b)$$

(We simply look on which side of the hyperplane the new point is.)

Nonlinear classification

If data are not linearly separable, it's possible to find some mapping $\varphi : \mathbb{R}^D \rightarrow \mathbb{R}^{D'}$, $D' \geq D$, which transforms data into a higher dimensionality space in which they are linearly separable.





Nonlinear classification

In fact, such transformation is not needed to be computed explicitly.

Dual form of maximization problem in transformed space:

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \varphi(\mathbf{x_i}) \cdot \varphi(\mathbf{x_j})$$

Label the dot product of transformed vectors as a function k

$$k(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i}) \cdot \varphi(\mathbf{x_j})$$

So we get

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x_i}, \mathbf{x_j})$$

Nonlinear classification

Kernel trick

We can use any (nonlinear) kernel function as k to get a nonlinear classifier.
We don't need to know transforation φ explicitly, it's implicitly represented by k.

Kernel function examples

Some commonly used kernel functions:

- Gaussian radial basis kernel:
  $k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2)$ for $\gamma > 0$
- Polynomial kernel:
  $k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + \alpha)^\beta$
- Sigmoidal kernel:
  $k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\alpha\mathbf{x_i} \cdot \mathbf{x_j} - \beta)$

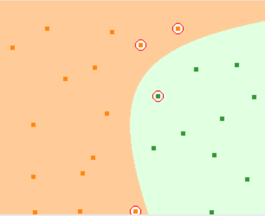
It is also possible to define kernel functions working with more complex data structures (sets, strings, DNA sequences, ...)

Nonlinear classification

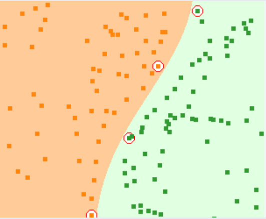
Examples of nonlinear SVM

(hyperplane transformed back to the original space)

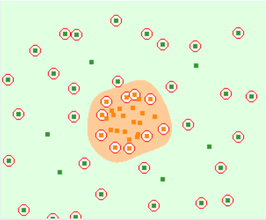
Polynomial (degree 2)



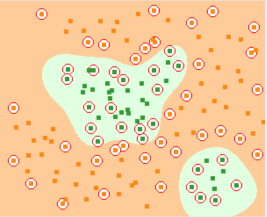
Polynomial (degree 3)



Radial basis function



Radial basis function



Multi-class classification

SVM is binary classifier, but it can be used to multi-class classification as well.

- Multi-class classification can be decomposed into several binary classification tasks.
  - One-versus-All, One-versus-One, ...
- There are also some extensions to basic SVM, which allows multi-class classification.
  - Optimization problem becomes much more complex.

Conclusion

Conclusion

- SVM – binary linear classifier
  - Finds maximum-margin hyperplane
  - Quadratic programming optimization
- Using kernel trick it can be changed to nonlinear
  - It is usually used in this way.
- Effective especially for large number of dimensions
- Very robust and often used in practice

Thank you for your attention.

Questions?

Sources

Main sources

- Tristan Fletcher: Support Vector Machines Explained
  `http://www.tristanfletcher.co.uk/SVM%20Explained.pdf`
- Wikipedia: Support Vector Machines
  `http://en.wikipedia.org/wiki/Support_vector_machine`
- Hakan Serce: SVM Applet
  `http://www.eee.metu.edu.tr/~alatan/Courses/Demo/AppletSVM.html`