Slide 1

Support Vector Machines for Data Classification and Regression

Chih-Jen Lin

Department of Computer Science
National Taiwan University

Talk at Ford Research Lab., July 24, 2001

Slide 2

Outline

- Support vector machines (SVM): an introduction
- Optimization algorithms for SVM
- An example on using SVM in practice
- Support Vector Regression
- Discussion and future directions

Slide 3

Data Classification

- Given training data in different classes (labels known)
  Predict test data (labels unknown)
- Examples
  - Handwritten digits recognition
  - Speech recognition
  - Text classification
  - Face recognition
- Methods:
  - Nearest Neighbor
  - Neural Networks

Slide 4

  - Decision Tree
- Support vector machines: a new method
- Early development in Bell Labs from 1990 to 1995
- Becoming more and more popular
- The ambition: replacing other methods on some applications

Slide 5

Support Vector Machines

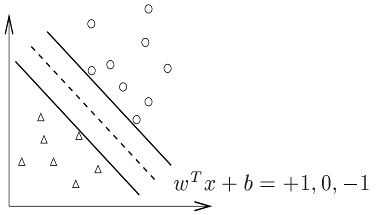
- Training vectors : $x_i, i = 1, \ldots, l$
- Consider a simple case with two classes:

  Define a vector y

$$y_i = \begin{cases} 1 & \text{if } x_i \text{ in class 1} \\ -1 & \text{if } x_i \text{ in class 2,} \end{cases}$$

- A hyperplane which separates all data

Slide 6



- A separating hyperplane: $w^T x + b = 0$

$$\begin{aligned} (w^T x_i) + b > 0 \quad & \text{if } y_i = 1 \\ (w^T x_i) + b < 0 \quad & \text{if } y_i = -1 \end{aligned}$$

- Decision function $f(x) = \text{sign}(w^T x + b)$, x: test data

  Variables: w and b : Need to know coefficients of a plane

  Many possible choices of w and b

Slide 7

- Select w, b with the maximal margin.

  Maximal distance between $w^T x + b = \pm 1$

  Vapnik's statistical learning theory.

$$\begin{aligned} (w^T x_i) + b \geq 1 \quad & \text{if } y_i = 1 \\ (w^T x_i) + b \leq -1 \quad & \text{if } y_i = -1 \end{aligned} \tag{1}$$

- Distance between $w^T x + b = 1$ and -1:

$$2/\|w\| = 2/\sqrt{w^T w}.$$

- $\max 2/\|w\| \equiv \min w^T w/2$

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^T w \\ & y_i((w^T x_i) + b) \geq 1, \qquad \text{from (1)} \\ & i = 1, \ldots, l. \end{aligned}$$

Slide 8

Higher Dimensional Feature Spaces

- Earlier we tried to find a linear separating hyperplane

  Data may not be linear separable

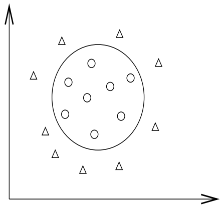
- Non-separable case: allow training errors

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C(\sum_{i=1}^{l} \xi_i) \\ & y_i((w^T x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ i = 1, \ldots, l \end{aligned}$$

- $\xi_i > 1$, x_i not on the correct side of the separating plane
- C: large penalty parameter, most ξ_i are zero

Slide 9

- Nonlinear case: linear separable in other spaces ?



- Higher dimensional (maybe infinite) feature space

$$\phi(x) = (\phi_1(x), \phi_2(x), \ldots).$$

- Example: $x \in R^3, \phi(x) \in R^{10}$

$$\begin{aligned}\phi(x) &= (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2,\\ &\quad x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)\end{aligned}$$

Slide 10

- A standard problem [Cortes and Vapnik, 1995]:

$$\begin{aligned}\min_{w,b,\xi} \quad & \frac{1}{2}w^Tw + C(\sum_{i=1}^{l}\xi_i)\\ & y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i,\\ & \xi_i \geq 0,\ i = 1, \ldots, l\end{aligned}$$

- Other variants (though similar); Example:

$$\begin{aligned}\min_{w,b,\xi} \quad & \frac{1}{2}w^Tw + C(\sum_{i=1}^{l}\xi_i^2)\\ & y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i,\\ & \xi_i \geq 0,\ i = 1, \ldots, l\end{aligned}$$

Slide 11

Finding the Decision Function

- Finding w and b from the standard SVM form

  w: a vector in a high dimensional space $\Rightarrow$ maybe infinite variables

- The dual problem

$$\begin{aligned}\min_{\alpha} \quad & \frac{1}{2}\alpha^TQ\alpha - e^T\alpha\\ & 0 \leq \alpha_i \leq C, i = 1, \ldots, l\\ & y^T\alpha = 0,\end{aligned}$$

where $Q_{ij} = y_iy_j\phi(x_i)^T\phi(x_j)$ and $e = [1, \ldots, 1]^T$

$$w = \sum_{i=1}^{l}\alpha_iy_i\phi(x_i)$$

Slide 12

- Primal and dual : optimization theory. Not trivial.

  Infinite dimensional programming.

- A finite problem:

  #variables = #training data

- $Q_{ij} = y_iy_j\phi(x_i)^T\phi(x_j)$ needs a closed form

  Efficient calculation of high dimensional inner products

- Example: $x_i \in R^3, \phi(x_i) \in R^{10}$

$$\begin{aligned}\phi(x_i) &= (1, \sqrt{2}(x_i)_1, \sqrt{2}(x_i)_2, \sqrt{2}(x_i)_3, (x_i)_1^2,\\ &\quad (x_i)_2^2, (x_i)_3^2, \sqrt{2}(x_i)_1(x_i)_2, \sqrt{2}(x_i)_1(x_i)_3, \sqrt{2}(x_i)_2(x_i)_3),\end{aligned}$$

  Then $\phi(x_i)^T\phi(x_j) = (1 + x_i^Tx_j)^2$.

Slide 13

- Popular methods: $\phi(x_i)^T \phi(x_j) =$

$$e^{-\gamma \|x_i - x_j\|^2}, \text{ (Radial Basis Function)}$$
$$(x_i^T x_j / a + b)^d \text{ (Polynomial kernel)}$$
$$\tanh(a x_i^T x_j + b)$$

- Decision function:

$$\begin{aligned} & w^T \phi(x) + b \\ = & \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)^T \phi(x) + b \end{aligned}$$

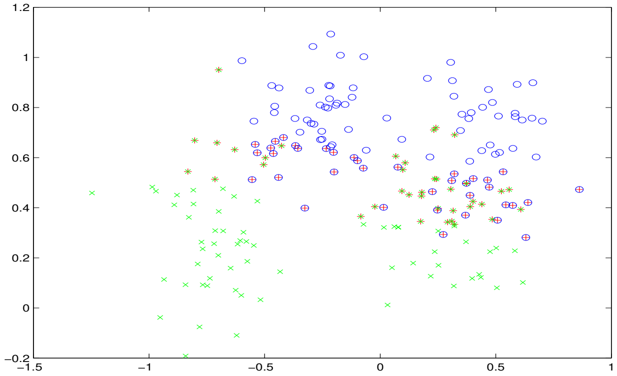
No need to have w

- > 0: 1st class, < 0: 2nd class
- Only $\phi(x_i)$ of $\alpha_i > 0$ used

$\alpha_i > 0 \Rightarrow$ support vectors

Slide 14

Support Vectors: More Important Data



Slide 15

Issues

- Why is this good ? Statistical learning theory
- Solving large quadratic problems: dual variable α
- Multiple-class classifications
  - Several two-class problems or combined together
- Automatic model selection
  - select the best parameters (kernel type, C, etc)
- Comparisons with other methods
- Applications

Slide 16

Large Dense Quadratic Programming

- $\min \frac{1}{2}\alpha^T Q \alpha - e^T \alpha, y^T \alpha = 0, 0 \leq \alpha_i \leq C$
- $Q_{ij} \neq 0$, Q : an l by l fully dense matrix
- 30,000 training points: 30,000 variables:
  $(30,000^2 \times 8/2)$ bytes $=$ 3GB RAM: still difficult
- Traditional methods: Newton, Quasi Newton cannot be directly applied
- Current methods:
  - Decomposition methods (Osuna et al. [1997], Joachims [1998], Platt [1998])
  - Nearest point of two convex hulls (Keerthi et al. [1999a])

Slide 17

Decomposition Methods: Avoid Memory Problem

- Working on a few variable each time
- It is like that for minimizing a function with 10 variables, you sequentially work on one variable in each iteration
- Working set B, $N = \{1, \dots, l\} \backslash B$ fixed

  Size of B usually <= 100
- Submatrices Q_{BB} and Q_{BN} needed

  calculated when needed: avoid the memory problem
- The objective function is decreasing

  Convergence was not fully understood

Slide 18

- Studies on convergence proofs: (Chang et al. [1999], Keerthi and Gilbert [2000], Lin [2000])
- Software: need knowledge of optimization
- Early implementation:

  Working set by heuristics

  Stopping conditions not validated
- Efficient when the percentage of support vectors is small

Slide 19

An Example: Generalization Ability Challenge

- First problem of IJCNN Challenge 2001
- Given time series length $T = 50,000$
- The kth data

$$x_1(k), x_2(k), x_3(k), x_4(k), x_5(k), y(k)$$

- $y(k) = \pm 1$: output, affected only by $x_1(k), \dots, x_4(k)$
- $x_5(k) = 1$ considering $y(k)$ for evaluating accuracy
- 50,000 training data, 100,000 testing data (in two sets)
- Past and future information may affect $y(k)$
- $x_1(k)$: periodically nine 0s, one 1, nine 0s, one 1, and so on.

Slide 20

GAC: Encoding Schemes

- For SVM: each data is a vector
- $x_1(k)$: periodically nine 0s, one 1, nine 0s, one 1, ...

  10 binary attributes
  $x_1(k-5), \dots, x_1(k+4)$ for the kth data
- $x_4(k)$ more important

  Including $x_4(k-5), \dots, x_4(k+4)$ for the kth data
- Each training data: 22 attributes

Slide 21

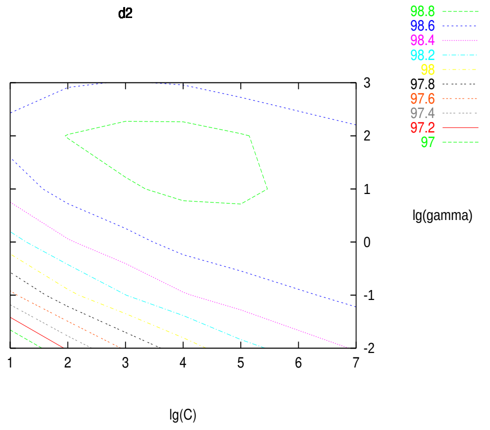
GAC: Training SVM

- Selecting parameters; generate a good model for prediction
- Avoid overfitting
- 5-fold cross validation on 50,000 data

  Data randomly separated to five groups.

  Each time four as training and one as testing
- RBF kernel $K(x_i, x_j) = \phi(x_i)\phi(x_j) = e^{-\gamma\|x_i - x_j\|^2}$
- Two parameters: γ and C
- Use $C = 2^4, \gamma = 2^2$ and train 50,000 data for the final model

Slide 22



Slide 23

- Training SVM: One (C, γ), 50,000 data, 5-fold CV: $<$ 1 hour on PIII-500 with 384 MB RAM
- Test set 1: 656 errors, Test set 2: 637 errors
- About 3000 support vectors of 50,000 training data

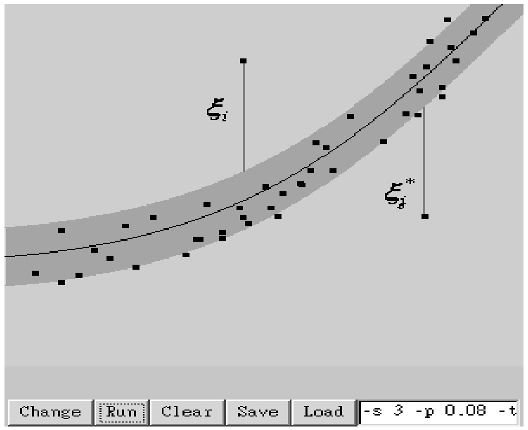
  A good case for SVM
- This is just the outline. There are other details.
- It is essential to do model selection.

  You can see some papers which compare variants of SVM with the same parameters: Not right

  A fair comparison: training and testing sets. Obtain good parameters using only training set by validation

Slide 24

Support Vector Regression



Slide 25

- Primal

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i + C \sum_{i=1}^{l} \xi_i^*$$
$$y_i - w^T \phi(x_i) - b \leq \epsilon + \xi_i,$$
$$w^T \phi(x_i) + b - y_i \leq \epsilon + \xi_i^*,$$
$$\xi_i, \xi_i^* \geq 0, i = 1, \ldots, l.$$

- ϵ-insensitive loss function

$$-\epsilon - \xi_i^* \leq y_i - (w^T \phi(x_i) + b) \leq \epsilon + \xi_i$$

A tube with width ϵ

- There is also a dual
- $\frac{1}{2} w^T w$?

Slide 26

Comparisons with Other Methods

- Result from [Hsu and Lin, 2001]:

Problem	#training	#testing	#class	statlog rate	SVM
dna	2000	1186	3	95.9%	95.447%
satimage	4435	2000	6	90.6%	91.3%
letter	15000	5000	26	93.6%	97.98%
shuttle	43500	14500	7	99.99%	99.924%

- Statlog: best rates by different methods
- SVM is very competitive

Slide 27

Existing Applications

- Most classification problems should be applied without problems
- Many of them have not been tried
- Examples
  - Protein secondary structure prediction
  - Handwritten digits recognition (currently the highest accuracy on MNIST)
  - Text classification
  - Face recognition
  - Microarray data classification

Slide 28

Our Past Work

- Analysis for solving the dual problem
- New and faster algorithms for the dual problem
- Efficient decomposition method for difficult problems
  BSVM: February 2000
- Easy-to-use software for users in other fields
  LIBSVM: April 2000
  - Matlab interface developed at Ohio State University
  - R interface developed at Technische Universität Wien
  - Python interface developed at HP Labs.
  - Used by several data mining/document classification courses

Slide 29

- Model selection software based on **BSVM**
  looms: November 2000
- Multi-class SVM
- Past applications/competitions:
  - Optical Character Recognition (OCR)
    Winner of OCR competition by the UK post office, December 2000
    A regression problem
  - IJCNN Challenge 2001

Slide 30

Conclusions and Future Directions

- On some problems SVM is very competitive
- It is simple. Not a black box
  There are (and will have) too many variants
  May not be too good
- Integration with data mining software
- Applications: key to improve SVM
- All our research results can be found at
  `http://www.csie.ntu.edu.tw/~cjlin`