Slide 1

Implementation of Support Vector Machines Software: Theory and Practice

Chih-Jen Lin

Department of Computer Science and
Information Engineering
National Taiwan University

Talk at Ohio State University, August 29, 2000

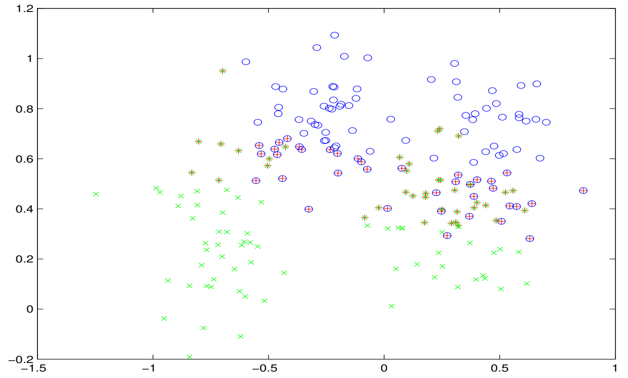
Slide 2

Outline

- Support vector machines (SVM): an introduction
  - SVM: a new machine learning technique
- Optimization algorithms for SVM
- Implementation for large data sets.
- Challenges and future directions

Slide 3

Support Vector Machines for Data Classification



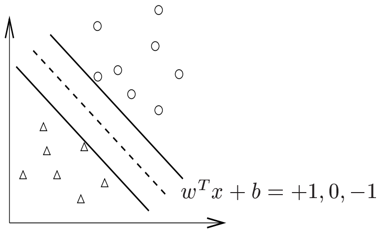
Slide 4

Support Vector Machines

- Training and testing
- Training vectors : $x_i, i = 1, \ldots, l$
- Consider examples with two classes:

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

Slide 5



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

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

- Variables: w and b : Need to know coefficients of a plane
  Decision function $w^Tx + b$, x: test data
  Many possible choices
- Select w, b with the maximal margin.

Slide 6

Maximal distance between $(w^Tx_i) + b = \pm 1$
Vapnik's statistical learning theory.

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

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

$$2/\|w\|^2 = 2/w^Tw.$$

- $\max 2/w^Tw \equiv \min w^Tw/2$

$$\begin{aligned} \min \quad & \frac{1}{2}w^Tw \\ & y_i((w^Tx_i) + b) \geq 1, \\ & i = 1, \ldots, l. \end{aligned}$$

Slide 7

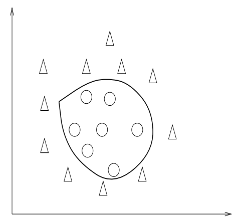
Higher Dimensional Feature Spaces

- Data may not be linear separable
- Non-separable case: allow training errors

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

- $\xi_i > 1$, x_i on the other side of the separating plane
- C: large penalty parameter, most ξ_i are zero
- Nonlinear case: separable in other spaces ?

Slide 8



- Higher dimensional (maybe infinite) feature space

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

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

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

Slide 9

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

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

- Othere variants are similar

Slide 10

Find the Decision Function

- w: a vector in an infinite dimensional space
  infinite variables
- The dual problem

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

  where $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$.

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

- Primal and dual: based on optimization theory. Not trivial.
  Infinite dimensional programming.

Slide 11

- $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$ needs a closed form: efficient calculation of high dimensional inner products
- Example: $x_i \in R^3, \phi(x_i) \in R^{10}$

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

  Then $\phi(x_i)^T \phi(x_j) = (1 + x_i^T x_j)^2$.
- Popular methods: $\phi(x_i)^T \phi(x_j) =$

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

- A finite problem
- #variables = #training data

Slide 12

- Decision function:

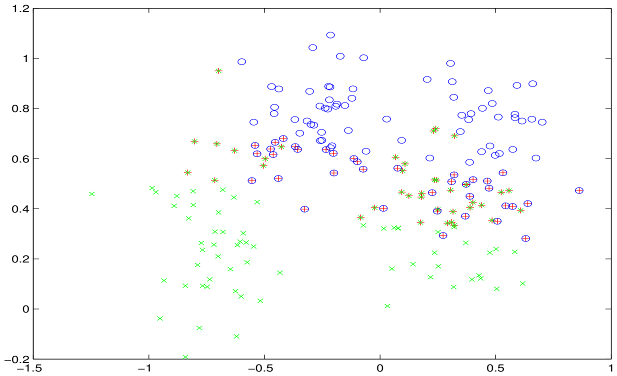
$$
\begin{aligned}
& w^T \phi(x) + b \\
= \ & \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)^T \phi(x) + b
\end{aligned}
$$

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

$\alpha_i > 0 \Rightarrow$ support vectors

Slide 13

Support Vectors: More Important Data



Slide 14

Issues

- Solving large quadratic problems
- Multiple-class classifications
  - Several two-class problems
- Automatic model selection
  - Try different parameters
- Comparisons with other methods
- Applications
- Other areas: Regression analysis, function estimation

Slide 15

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: 3GB RAM
  Still a difficult problem
- 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 16

Decomposition Methods: Work on Few Variables

Each time

1. Given $q \ll l$, α^1: initial solution. $k \leftarrow 1$.
2. If α^k an optimum, stop. Find a working set $B \subset \{1, \ldots, l\}$, $|B| = q$. Define $N \equiv \{1, \ldots, l\} \backslash B$, α_B^k and α_N^k
3. Solve a sub-problem:
$$\begin{aligned} \min \quad & \frac{1}{2}\alpha_B^T Q_{BB}\alpha_B - (e_B - Q_{BN}\alpha_N^k)^T\alpha_B \\ & 0 \leq (\alpha_B)_i \leq C, i = 1, \ldots, q, \\ & y_B^T\alpha_B = -y_N^T\alpha_N^k, \end{aligned}$$
4. Set α_B^{k+1} and α_N^{k+1}, $k \leftarrow k+1$ and goto Step 2.

Slide 17

- Q_{BB} is calculated when required: avoid the memory problem
- The objective function is decreasing
  Convergence not fully understood yet
- Studies on convergence proofs: (Chang et al. [1999], Keerthi [2000], Lin [2000])
- Convergence may not be that important
  However, software should follow basic principles of optimization algorithms.

Slide 18

Working Set Selection

- Size of the working set
  - Reducing the size to two
    Sequential Minimization Optimization (SMO) by Platt
    Analytic solution of the sub-problem
  - in general ≤ 100
- Selection
  - Random selection ?
  - First systematic way: Joachims [1998] (SVM^{light})
  - Without systematic selection, difficult to prove convergence (e.g. SMO may not converge, Keerthi et al. [1999b])

Slide 19

Computational Time

- Cache for Q
- #iterations ⇒ rate of convergence
- Easy problems ⇒ 50,000 data solved in minutes
- Difficult problems ⇒ very slow convergence
  50,000 data can take hours
- Local convergence of optimization algorithms
  - Newton: quadratic
  - Quasi Newton: superlinear
  - Decomposition method: even does not consider the whole vector

Slide 20

Better Working Set Selection

- We have a software BSVM
  ⇒ faster convergence on difficult cases
- Still not satisfactory
  More investigation and new techniques need to be done

Slide 21

Challenges and Future Directions

- Model selection:
  - Mapping data into infinite dimensional spaces: training error $\downarrow 0$.
  - Overfitting
  - Training error not important
    Test error counts : generalization property
  - $\min \frac{1}{2}\alpha^T Q\alpha - e^T\alpha, y^T\alpha = 0, 0 \leq \alpha \leq C_1$
  - $\min \frac{1}{2}\alpha^T Q\alpha - e^T\alpha, y^T\alpha = 0, 0 \leq \alpha \leq C_2$
  - Parameters: $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j) = e^{-\gamma\|x_i - x_j\|^2}$
  - Cross validation
    * Data separated to k groups, $k-1$ for traing and 1 for testing

Slide 22

  - Leave one out
    * Each group only 1 element: l quadratic programs
  - Initial solution (warm start)
  - Reuse of Q
- User interface
  - Use of optimization solvers
  - Combining SVM classification and regression
  - Additional parameters: C^+, C^-

$$\frac{1}{2}w^T w + C^+(\sum_{i \in \text{class } 1} \xi_i) + C^-(\sum_{i \in \text{class } 2} \xi_i)$$

  - Input format: sparse vs. dense