# Support vector machines: status and challenges

Chih-Jen Lin
Department of Computer Science
National Taiwan University

Talk at Caltech, November 2006

# Outline

- Basic concepts
- Current Status
- Challenges
- Conclusions

# Outline

- Basic concepts
- Current Status
- Challenges
- Conclusions

# Support Vector Classification

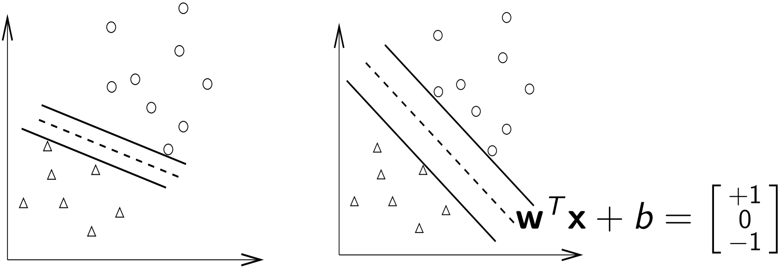
- Training vectors : $\mathbf{x}_i, i = 1, \ldots, l$
- Feature vectors. For example,

  A patient = [height, weight, ...]
- Consider a simple case with two classes:

  Define an indicator vector $\mathbf{y}$

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

- A hyperplane which separates all data




- A separating hyperplane: $\mathbf{w}^T\mathbf{x} + b = 0$

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

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

  Many possible choices of $\mathbf{w}$ and $b$

# Maximal Margin

- Distance between $\mathbf{w}^T\mathbf{x} + b = 1$ and $-1$:

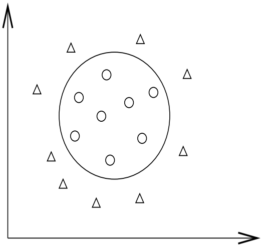
$$2/\|\mathbf{w}\| = 2/\sqrt{\mathbf{w}^T\mathbf{w}}$$

- A quadratic programming problem [Boser et al., 1992]

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \\ & i = 1, \ldots, l. \end{aligned}$$

# Data May Not Be Linearly Separable

- An example:



- Allow training errors
- Higher dimensional ( maybe infinite ) feature space

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

- Standard SVM [Cortes and Vapnik, 1995]

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ i = 1, \ldots, l. \end{aligned}$$

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

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

# Finding the Decision Function

- $\mathbf{w}$: maybe infinite variables
- The dual problem

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \le \alpha_i \le C, i = 1, \ldots, l \\ & \mathbf{y}^T \boldsymbol{\alpha} = 0, \end{aligned}$$

  where $Q_{ij} = y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ and $\mathbf{e} = [1, \ldots, 1]^T$
- At optimum

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

- A finite problem: #variables = #training data

# Kernel Tricks

- $Q_{ij} = y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ needs a closed form
- Example: $\mathbf{x} \in R^3, \phi(\mathbf{x}) \in R^{10}$

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

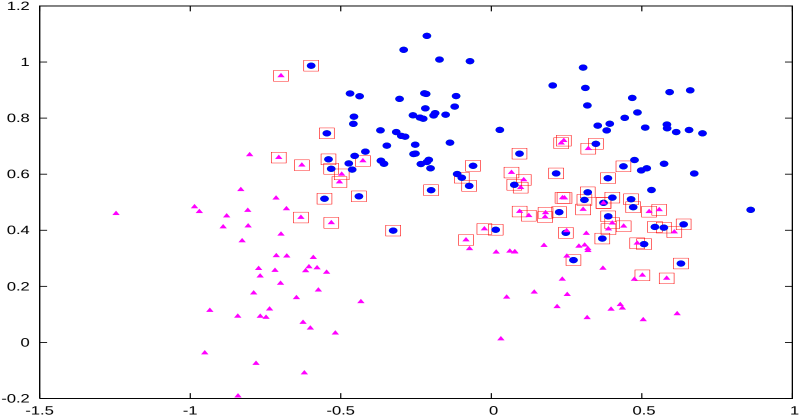
Then $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^2 \quad \Rightarrow \quad K(\mathbf{x}_i, \mathbf{x}_j)$

- Decision function

$$\mathbf{w}^T \phi(\mathbf{x}) + b = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) + b$$

- Only $\phi(\mathbf{x}_i)$ of $\alpha_i > 0$ used $\Rightarrow$ support vectors

# Support Vectors: More Important Data




A 3-D demonstration

www.csie.ntu.edu.tw/˜cjlin/libsvmtools/svmtoy3d

# Outline

# Solving the Dual

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

- $Q_{ij} \neq 0$, $Q$ : an $l$ by $l$ fully dense matrix
- 30,000 training points: 30,000 variables: $(30,000^2 \times 8/2)$ bytes $=$ 3GB RAM to store $Q$:
- Optimization methods cannot be directly applied
- Extensive work has been done
- Now easy to solve median-sized problems

- An example of training 50,000 instances using LIBSVM

```
$ ./svm-train -m 200 -c 16 -g 4 22features
optimization finished, #iter = 24981
Total nSV = 3370
time     5m1.456s
```

- Calculating $Q$ may have taken more than 5 minutes

  $\#\text{SVs} = 3{,}370 \ll 50{,}000$

- SVM properties used in optimization
- A detailed discussion

  www.csie.ntu.edu.tw/~cjlin/talks/rome.pdf

# Parameter/Kernel Selection

- Penalty parameter $C$: balance between generalization and training errors

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i$$

- kernel parameters
- Cross validation

  Data split to training/validation
- Other more efficient techniques

- Difficult if number of parameters is large

  E.g., feature scaling:

$$K(\mathbf{x}, \mathbf{y}) = e^{-\sum_{i=1}^{n} \gamma_i (x_i - y_i)^2}$$

  Some features more important

- A challenging research issue

# Design Kernels

- Still a research issue

  e.g., in bioinformatics and vision, many new kernels
- But, should be careful if the function is a valid one

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$$

- For example, any two strings $s_1, s_2$ we can define edit distance

$$e^{-\gamma \text{edit}(s_1, s_2)}$$

  It's not a valid kernel [Cortes et al., 2003]

# Multi-class Classification

- Combining results of several two-class classifiers
- One-against-the rest
- One-against-one
- And other ways
- A comparison in [Hsu and Lin, 2002]

# Outline

- Basic concepts
- Current Status
- Challenges
- Conclusions

# Challenges

Unbalanced data

- Some classes few data, some classes a lot
- Different evaluation criteria?

Structural data sets

- An instance may not be a vector

  e.g., a tree from a sentence
- Labels in order relationships

  SVM for ranking

# Challenges (Cont'd)

Multi-label classification

- An instance associated with $\geq 2$ labels
- e.g., a video shot includes several concepts

Large-scale Data

- SVM cannot handle large sets if using kernels

  Two possibilities:
- Linear SVMs. In some situations, can solve much larger problems
- Approximation: sub-sampling and beyond

# Challenges (Cont'd)

Semi-supervised learning

- Some available data unlabeled
- How can we guarantee the performance of using only labeled data?

# Outline

- Basic concepts
- Current Status
- Challenges
- Conclusions

# Why is SVM Popular?

No definitive answer; In my opinion

- Reasonably easy to use and often competitive performance
- Rather general: linear/nonlinear
  Gaussian process/RBF networks
- Basic concept relatively easy: maximal margin
- It's lucky

# Conclusions

- We must admit that

  SVM is a rather mature area
- But still quite a few interesting research issues

  Many are extensions of standard classification problems
- Detailed SVM tutorial in Machine Learning Summer School 2006

  www.csie.ntu.edu.tw/~cjlin/talks/MLSS.pdf

# References I

Boser, B. E., Guyon, I., and Vapnik, V. (1992).
A training algorithm for optimal margin classifiers.
In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press.

Cortes, C., Haffner, P., and Mohri, M. (2003).
Positive definite rational kernels.
In *Proceedings of the 16th Annual Conference on Learning Theory*, pages 41–56.

Cortes, C. and Vapnik, V. (1995).
Support-vector network.
*Machine Learning*, 20:273–297.

Hsu, C.-W. and Lin, C.-J. (2002).
A comparison of methods for multi-class support vector machines.
*IEEE Transactions on Neural Networks*, 13(2):415–425.