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: \( x_i, i = 1, \ldots, l \)
- Feature vectors. For example, 
  A patient = [height, weight, ...]
- Consider a simple case with two classes:
  Define an indicator vector \( y \)

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

- A hyperplane which separates all data
A separating hyperplane: \( \mathbf{w}^T \mathbf{x} + b = 0 \)

\[
(w^T x_i) + b \geq 1 \quad \text{if } y_i = 1 \\
(w^T x_i) + b \leq -1 \quad \text{if } y_i = -1
\]

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 $-1$:
  
  $$\frac{2}{\|\mathbf{w}\|} = \frac{2}{\sqrt{\mathbf{w}^T \mathbf{w}}}$$

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

  $$\min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

  subject to

  $$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1,$$

  $$i = 1, \ldots, l.$$
Data May Not Be Linearly Separable

- An example:

- Allow training errors

- Higher dimensional (maybe infinite) feature space

\[ \phi(x) = (\phi_1(x), \phi_2(x), \ldots). \]
Standard SVM [Cortes and Vapnik, 1995]

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

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

\[
\phi(\mathbf{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, \\\nx_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)
\]
Finding the Decision Function

- **w**: maybe infinite variables
- The dual problem

\[
\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\
\text{subject to} & \quad 0 \leq \alpha_i \leq C, \ i = 1, \ldots, l \\
& \quad y^T \alpha = 0,
\end{align*}
\]

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

- At optimum

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

- A finite problem: \#variables = \#training data
Kernel Tricks

- \( Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j) \) needs a closed form
- Example: \( x \in \mathbb{R}^3, \phi(x) \in \mathbb{R}^{10} \)

\[
\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3) \]

Then \( \phi(x_i)^T \phi(x_j) = (1 + x_i^T x_j)^2 \Rightarrow K(x_i, x_j) \)

- Decision function

\[
w^T \phi(x) + b = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)^T \phi(x) + b
\]

- Only \( \phi(x_i) \) of \( \alpha_i > 0 \) used \( \Rightarrow \) support vectors
Support Vectors: More Important Data

A 3-D demonstration
www.csie.ntu.edu.tw/~cjlin/libsvmtools/svmttoy3d
Outline

- Basic concepts
- Current Status
- Challenges
- Conclusions
Solving the Dual

\[
\min_{\alpha} \quad \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\
\text{subject to} \quad 0 \leq \alpha_i \leq C, \ i = 1, \ldots, l \\
y^T \alpha = 0
\]

- \( Q_{ij} \neq 0, \ Q : \text{an} \ l \ \text{by} \ l \ \text{fully dense matrix} \)
- 30,000 training points: 30,000 variables: 
  \((30,000^2 \times 8/2) \ \text{bytes} = 3\text{GB} \ \text{RAM to store} \ Q:\)
- Optimization methods \textbf{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 22 features
optimization finished, #iter = 24981
Total nSV = 3370
time 5m1.456s

• Calculating Q may have taken more than 5 minutes
#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

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

- Kernel parameters
- Cross validation
  - Data split to training/validation
- Other more efficient techniques
Difficult if number of parameters is large

E.g., feature scaling:

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

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

\[ K(x, y) = \phi(x)^T \phi(y) \]

- 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
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](http://www.csie.ntu.edu.tw/~cjlin/talks/MLSS.pdf)

