- Plain-vanilla computation of SVM params from $\mathbf{a}$:
- $\mathbf{w}$ from opt: $\mathbf{w} = \sum_{i=1}^{N} a_i \, t_i \, \boldsymbol{\phi}(\mathbf{x}_i)$
- SV $(a_i \neq 0)$: $t_i \, y(\mathbf{x}_i) = 1$, $y(\mathbf{x}_i) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b$. Find b
- Prediction for a new point $\mathbf{x}$: $y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b = \sum_{i=1}^{N} a_i \, t_i \, \boldsymbol{\phi}^T(\mathbf{x}_i) \boldsymbol{\phi}(\mathbf{x}) + b = \sum_{i=1}^{N} a_i \, t_i \, k(\mathbf{x}_i, \mathbf{x}) + b$
- $a_i = 0$: region; $[t_i \, y(\mathbf{x}_i) - 1] = 0$: support vectors
- Pred: $y(\mathbf{x}) = \sum_{j \in \mathscr{S}} a_j \, t_j \, k(\mathbf{x}_j, \mathbf{x}) + b$. Sum: Only SVs
- Only SVs to classify, only kernel, not $\mathbf{x}$ space
- SVs: $[t_i \, y(\mathbf{x}_i) - 1] = 0 \implies t_i \, y(\mathbf{x}_i) = 1$. Optimal b:
- $t_i \sum_{j \in \mathscr{S}} a_j \, t_j \, k(\mathbf{x}_i, \mathbf{x}_j) + b = 1$, solve for SV i, better:
- Mult by t_i: $\sum_{j \in \mathscr{S}} a_j \, t_j \, k(\mathbf{x}_i, \mathbf{x}_j) + b = t_i$. Num stable:
- $b = [t_i - \sum_{j \in \mathscr{S}} a_j \, t_j \, k(\mathbf{x}_i, \mathbf{x}_j)]$. Sum $\forall i \in \mathscr{S}$
- $N_{\mathscr{S}} b = \sum_{i \in \mathscr{S}} [t_i - \sum_{j \in \mathscr{S}} a_j \, t_j \, k(\mathbf{x}_i, \mathbf{x}_j)]$. Obtain b

Home Page

Title Page

Contents

Page 24 of 41

Go Back

Full Screen

Close

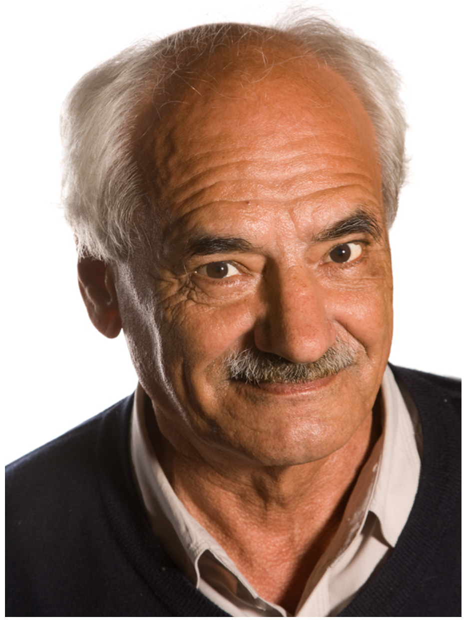
Quit

SVM: The Primal-Dual Question

- If KKT hold, solve the dual instead of the primal
- Primal: $\min \frac{1}{2}||\mathbf{w}||^2$ subject to $t_i \; [\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b] \geq 1$
- $\min \; L(\mathbf{w}, b, \mathbf{a}) \stackrel{\triangle}{=} \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{N} a_i \; (t_i \; [\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b] - 1)$
- Dual: $\max \; \widetilde{L}(\mathbf{a}) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i \, a_j \, t_i \, t_j \, k(\mathbf{x}_i, \mathbf{x}_j)$
- Primal & Dual expressions: identical at extremum
- $L(\mathbf{w}, b, \mathbf{a}) = \widetilde{L}(\mathbf{a})$ at the extremum
- Primal Optimisation: $\min \; L(\mathbf{w}, b, \mathbf{a})$
- Dual optimisation: $\max \; \widetilde{L}(\mathbf{a})$
- Optimisation theory: these are equivalent at KKT
- $\min$ wrt $\mathbf{w}, b$; $\max$ wrt $\mathbf{a}$

The SVM Story

V. Vapnik
[1936-]



A. Y. Chervonenkis
[1938-2014]

http://engineering.columbia.edu/files/engineering/vapnik.jpg

http://www.clrc.rhul.ac.uk/people/photos/AClarge.JPG

1963: SVM (Vapnik, Chervonenkis)
1992: Kernel trick (Boser, Gayoun, Vapnik)
1995: Soft Margin (Cortes, Vapnik)

Important: Why this formulation?

- This is one of the many possible formulations
- Mathematically elegant, numerically stable
- One of the simplest SVM formulations
- Historically, the first! [Vapnik, Chervonenkis'63]

Home Page

Title Page

Contents

Page 27 of 41

Go Back

Full Screen

Close

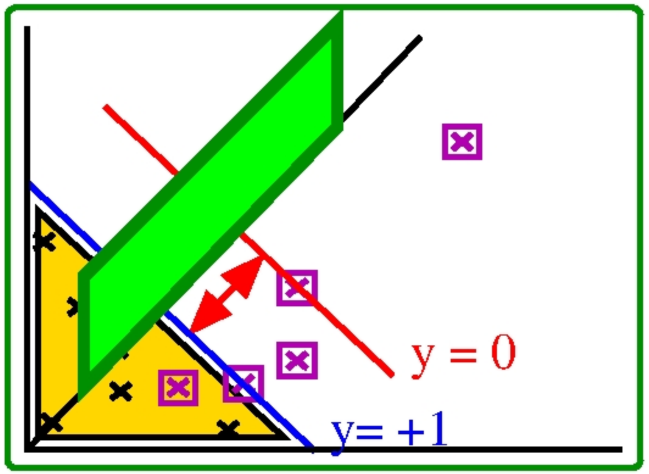
Quit

Application of an SVM

- Use a QP solver on the Dual problem to get $\mathbf{a}$
- Compute $\mathbf{w} = \sum_{i=1}^{N} a_i \, t_i \, \boldsymbol{\phi}(\mathbf{x}_i)$
- Find SVs $\mathscr{S}$: those indices for which $a_i > 0$
- Compute $b = \frac{1}{N_{\mathscr{S}}} \sum_{i \in \mathscr{S}} [t_i - \sum_{j \in \mathscr{S}} a_j \, t_j \, k(\mathbf{x}_i, \mathbf{x}_j)]$
- test point $\mathbf{x}$: $sgn(y(\mathbf{x}))$: $y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b$

Soft-Margin SVMs: Overlapping

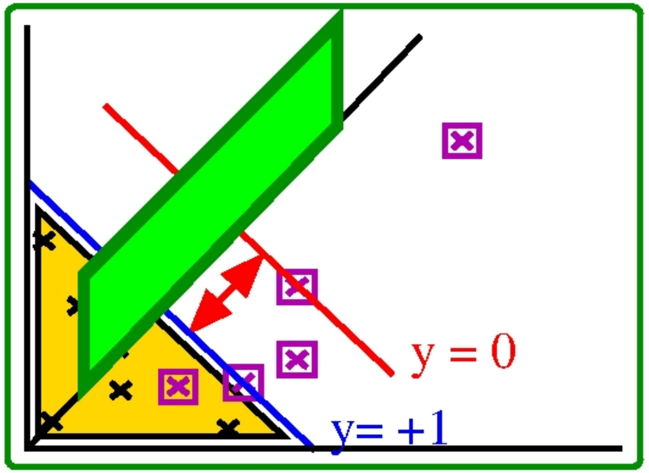
- Practical: when training data not linearly separable in the original $\mathbf{x}-$ or transformed $\boldsymbol{\phi}(\mathbf{x})-$ space
- The Oracle/QP solver earlier gave us direction $\mathbf{w}$ (& b). Adjusted the line to give symmetric margins
- This case: Oracle/QP solver tells us the optimal decision boundary, margins & outliers



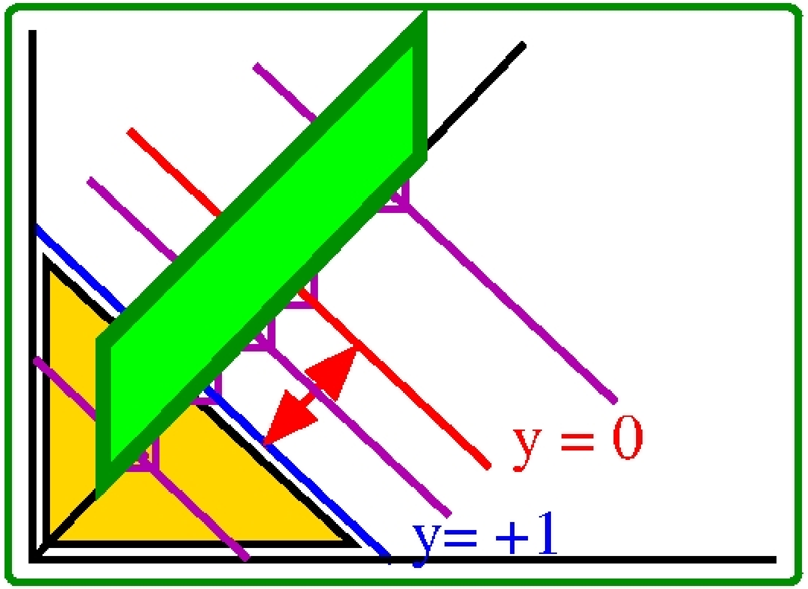
- Points in magenta: new
- Consider wrt one class (+1)
- correct, in the margin zone, on the boundary, outlier
- Points in terms of lines
- $||$ the decision boundary, & the 2 margin lines
- Points can be on the wrong side of decision boundary; penalty $\propto$ distance from boundary

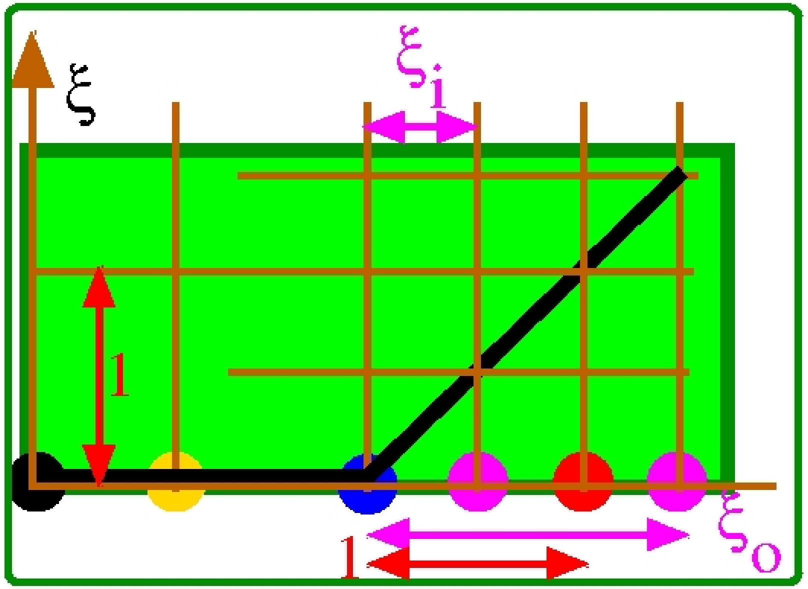
The ξ Story...

- 'Correct' part: $\xi_i \triangleq 0$. Else, $\xi_i \triangleq |t_i - y(\mathbf{x}_i)|$
- Penalty: linear fn of distance from boundary
- Why this? Simple. Elegant. Historically, the first!
- [Bennet'92], [Cortes & Vapnik '95]

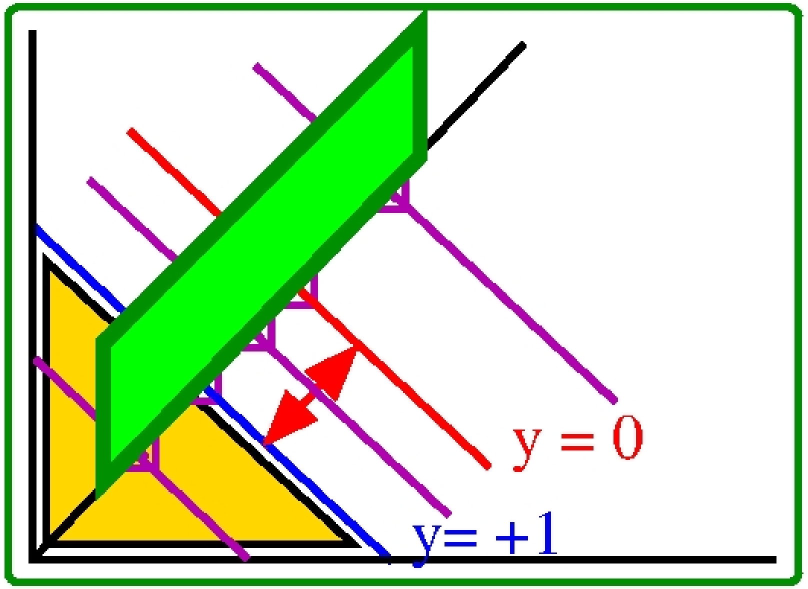


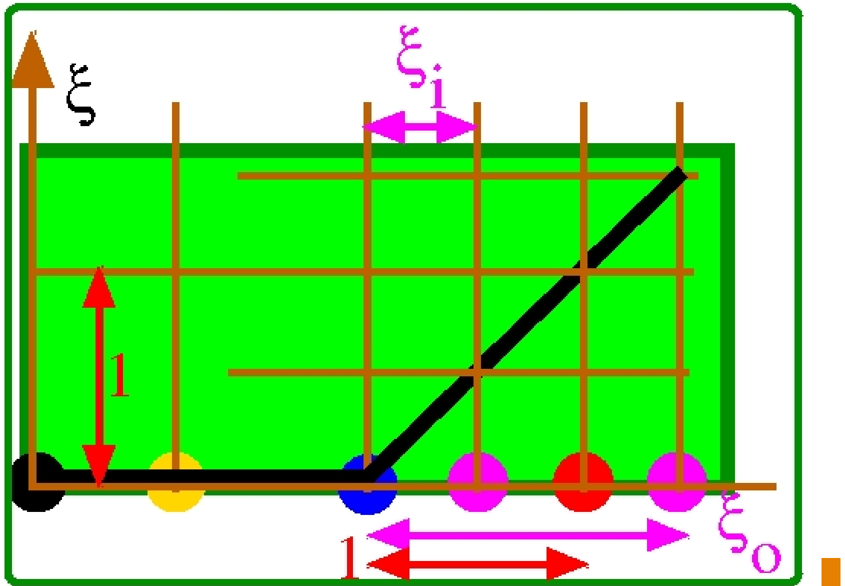
- Earlier: hard margin. Utopian: training data was linearly separable in original $\mathbf{x}-$ or kernel $\phi(\mathbf{x})-$ space
- Now relax this: even training data not linearly separable
- A green vertical board atop the 'distance' line
- 'Green-board space': Vertical axis: ξ_i. Horizontal axis: distance from the 'near' margin line
- 'Green-board space': (0,0): 'near' margin, $t_i = +1$





- 'near' margin: $\xi_i = |t_i - y(\mathbf{x}_i)|$=$|+1-(+1)|$=0: [0,0]
- boundary: $\xi_i = |t_i - y(\mathbf{x}_i)|$=$|+1-0|$=1: [1,1]
- An inliner excursion point in the 'zone': Consider a point at a distance ξ_i from the 'near' margin.
  – $\frac{\xi_i}{||\mathbf{w}||}$ farther from origin than the 'near' margin
  – eqn: $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+(b-1+\xi_i)=0$: $y_i = 1-\xi_i$
  – $\xi_i = |t_i - y(\mathbf{x}_i)|$=$|+1-(1-\xi_i)|$=$\xi_i$: $[\xi_i, \xi_i]$





- An outlier excursion point beyond the 'boundary': Consider a point at $\xi_o \stackrel{\triangle}{=} \xi_i$ from 'near' margin.
  - $\frac{\xi_o}{||\mathbf{w}||}$ farther from origin than the 'near' margin
  - eqn: $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + (b - 1 + \xi_o) = 0$: $y_i = 1 - \xi_o$
  - $\xi_i = |t_i - y(\mathbf{x}_i)| = |+1-(1-\xi_o)| = \xi_i$: $[\xi_o, \xi_o]$
  - Aliter: $\frac{\xi_o - 1}{||\mathbf{w}||}$ farther from the decision boundary
  - eqn: $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b + (\xi_o - 1) = 0$: $y_i = 1 - \xi_o$

- Similar formulation for the other side: $y = -1$ class
- $|\cdot|$ imposes symmetry
- For the +1 class, $\xi_i = |t_i - y(\mathbf{x}_i)|$, $\xi_i = 1 - y(\mathbf{x}_i)$
- Required hard region: $y(\mathbf{x}_i) \geq 1$, 'golden' region
- Required (relaxed) region: $y(\mathbf{x}_i) \geq 1 - \xi_i$
- Generalising from the -1 class also, $t_i\, y(\mathbf{x}_i) = 1 - \xi_i$
- Complete (relaxed) region: $t_i\, y(\mathbf{x}_i) \geq 1 - \xi_i$
- Hard to Soft: allow some data be misclassified
- Slack: allow for overlapping class distributions

Basic Soft-Margin Formulation

- Framework still sensitive to outliers as misclassification penalty increases linearly with ξ_i
- 1 possible formulation, QP-suitable & num stable
- Primal Formul'n: $\sim$hard-margin primal(simple!)
- To minimise $\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$, 1st term: previous
- 2nd term: penalises far points/excursions
- C: trade-off b/w margin & slack penalty, empirical
- $\sum_{i=1}^{N}\xi_i$: upper bounds # misclassifications Why?
- Correct classification: $\xi_i = 0$, or ≤ 1
- ($\forall$ Miscl: $\xi_i > 1$) $\sum_{i=1}^{K}\xi_i > K \Longrightarrow$ #miscl $< \sum_{i=1}^{K}\xi_i$
- $L(\mathbf{w}, b, \mathbf{a}, \boldsymbol{\xi}, \boldsymbol{\mu}) \triangleq \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{N} a_i\{t_i\left[\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b\right] - 1 + \xi_i\} + C\sum_{1=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i$