

SUPPORT VECTOR MACHINES

Presentation by
Saravanan

Lecture Slides adapted from Campbell

Outline

- Preliminaries
- SVMs for Binary Classification
- SVMs with Soft Margins
- Non Linear SVMs
- Multi Class SVMs
- Sample Usage of SVMs in IR

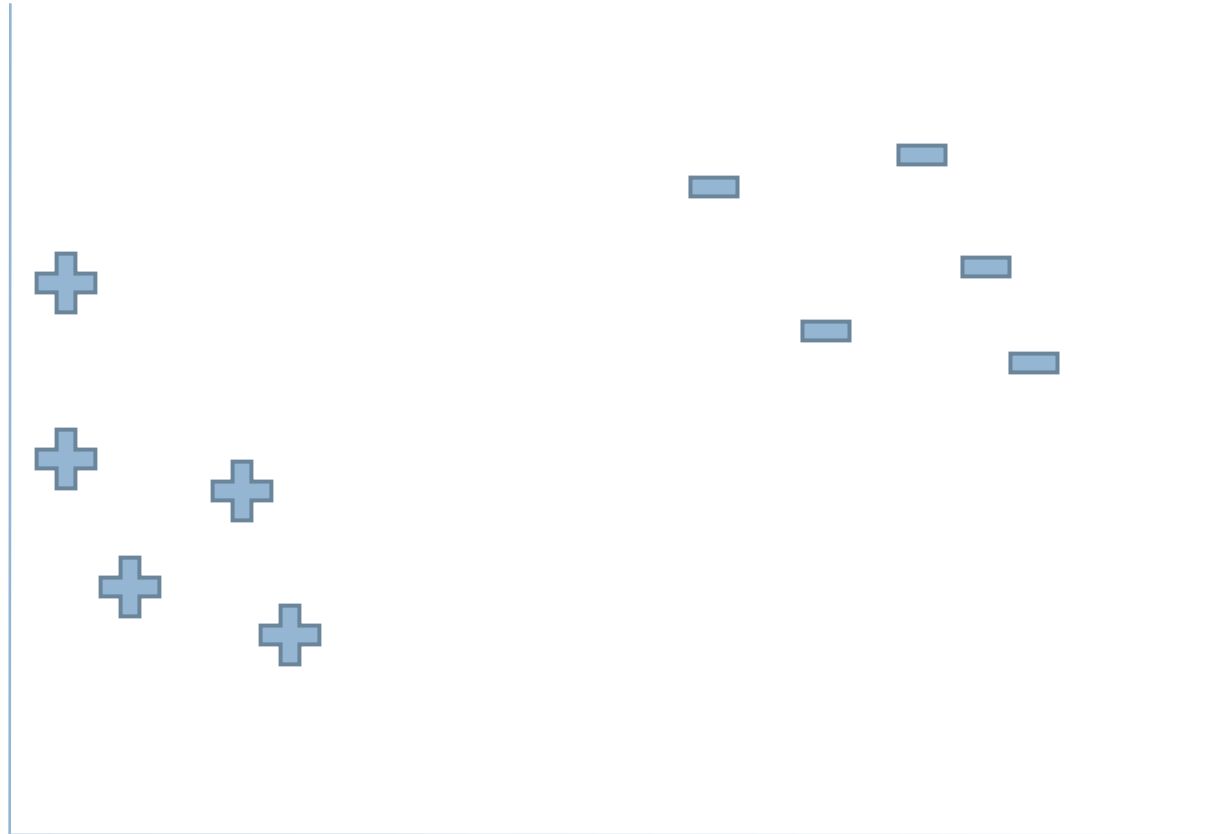
Terminologies

- Supervised Learning
- Two class and Multi class Classifiers
- Linearly separable problems
- Linear Classifiers
- Confidence of a classifier

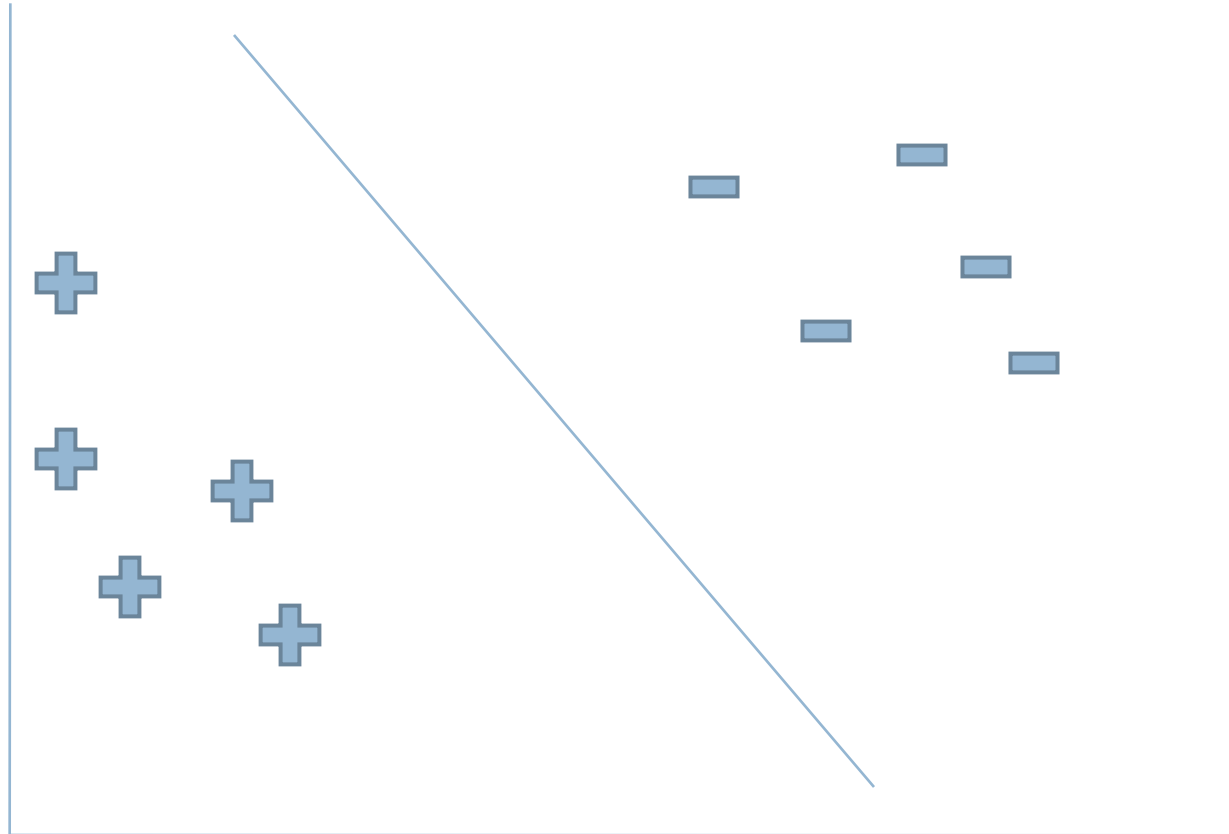
Linear Classifiers - 1

- Decides class membership based on linear combination of features to a threshold.
- Simple eg :
 - $w_1x_1 + w_2x_2 = b$
 - Assign to C if $w_1x_1 + w_2x_2 > b$
 - Assign to C' if $w_1x_1 + w_2x_2 \leq b$
- Generally, $W \cdot X = b$
 - W is the Weight Vector, X is the Input Vector.
 - b is Bias

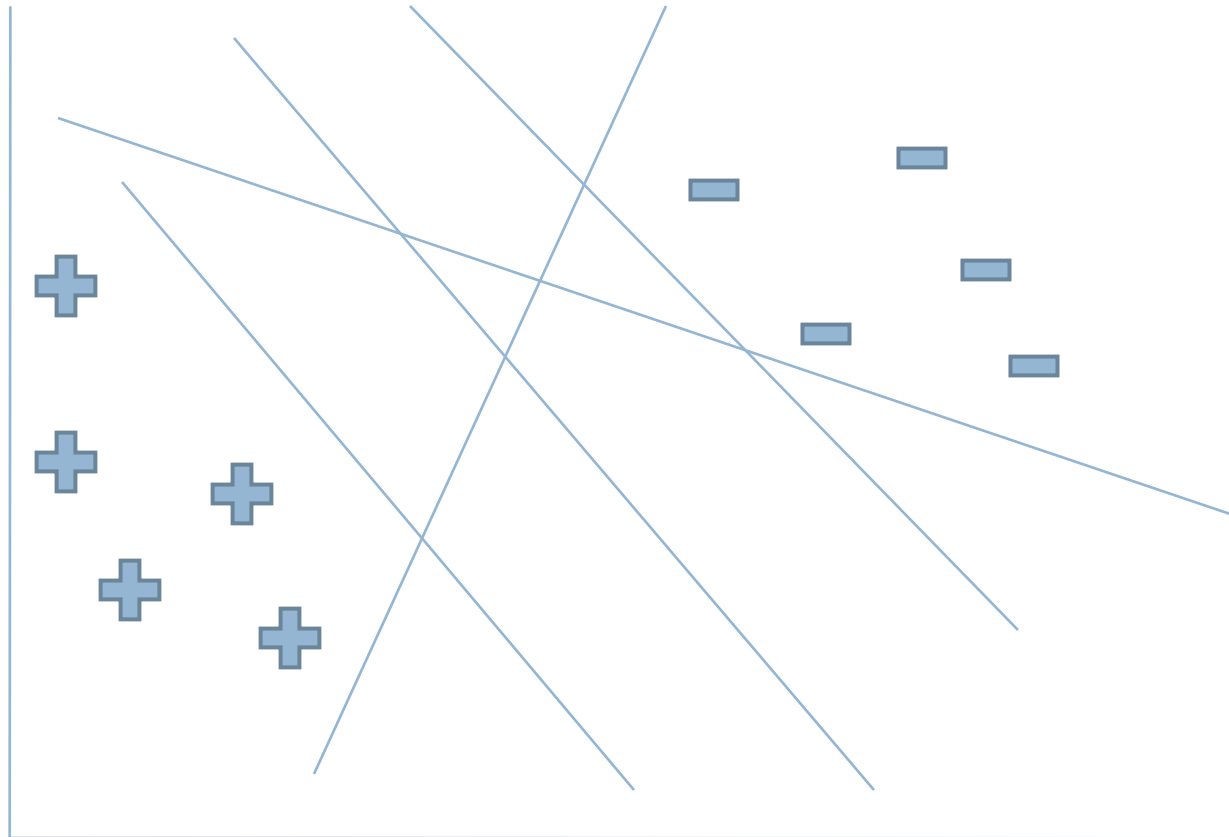
Linear Classifiers - 2



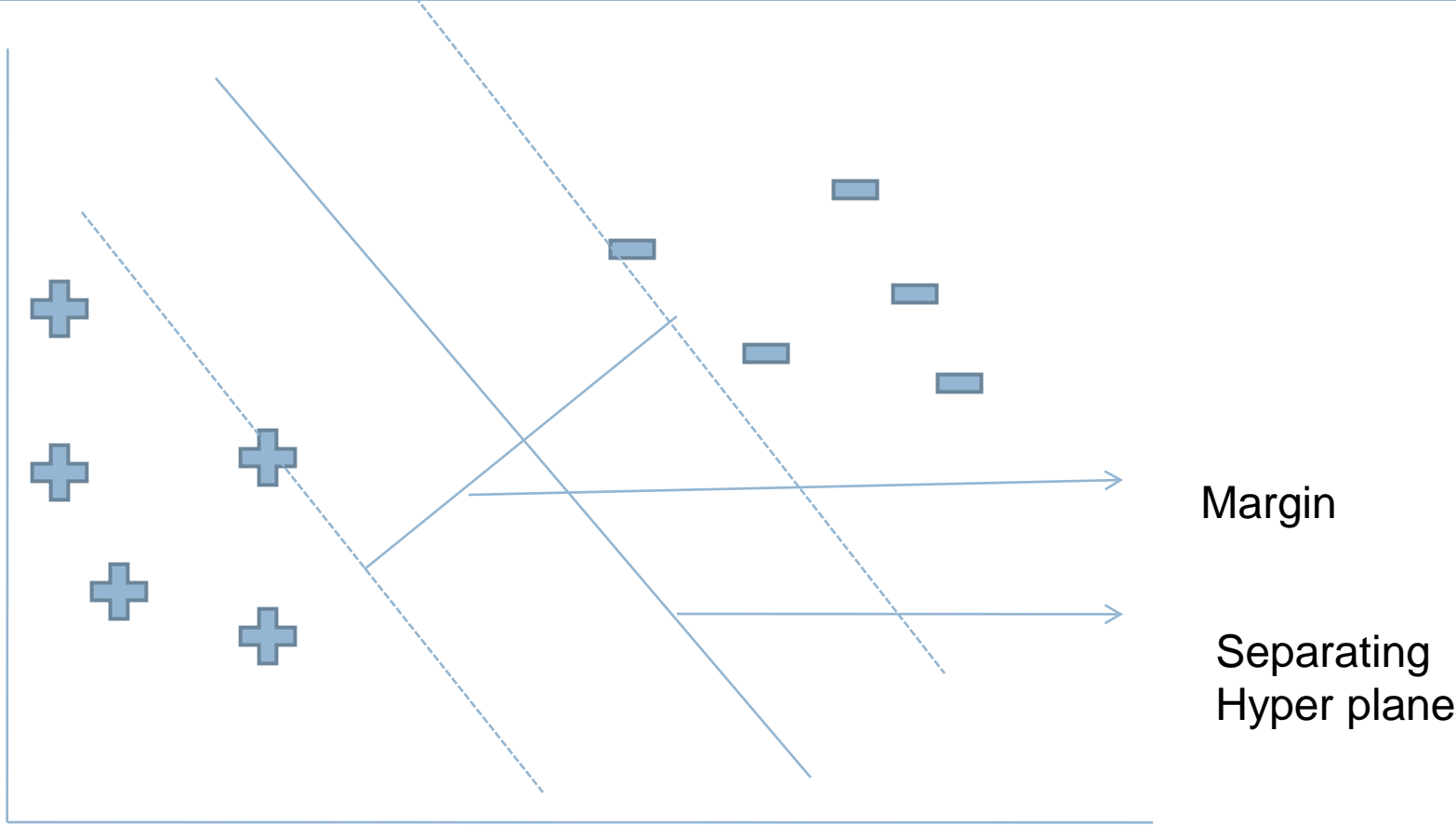
Linear Classifiers - 3



Linear Classifiers - 4



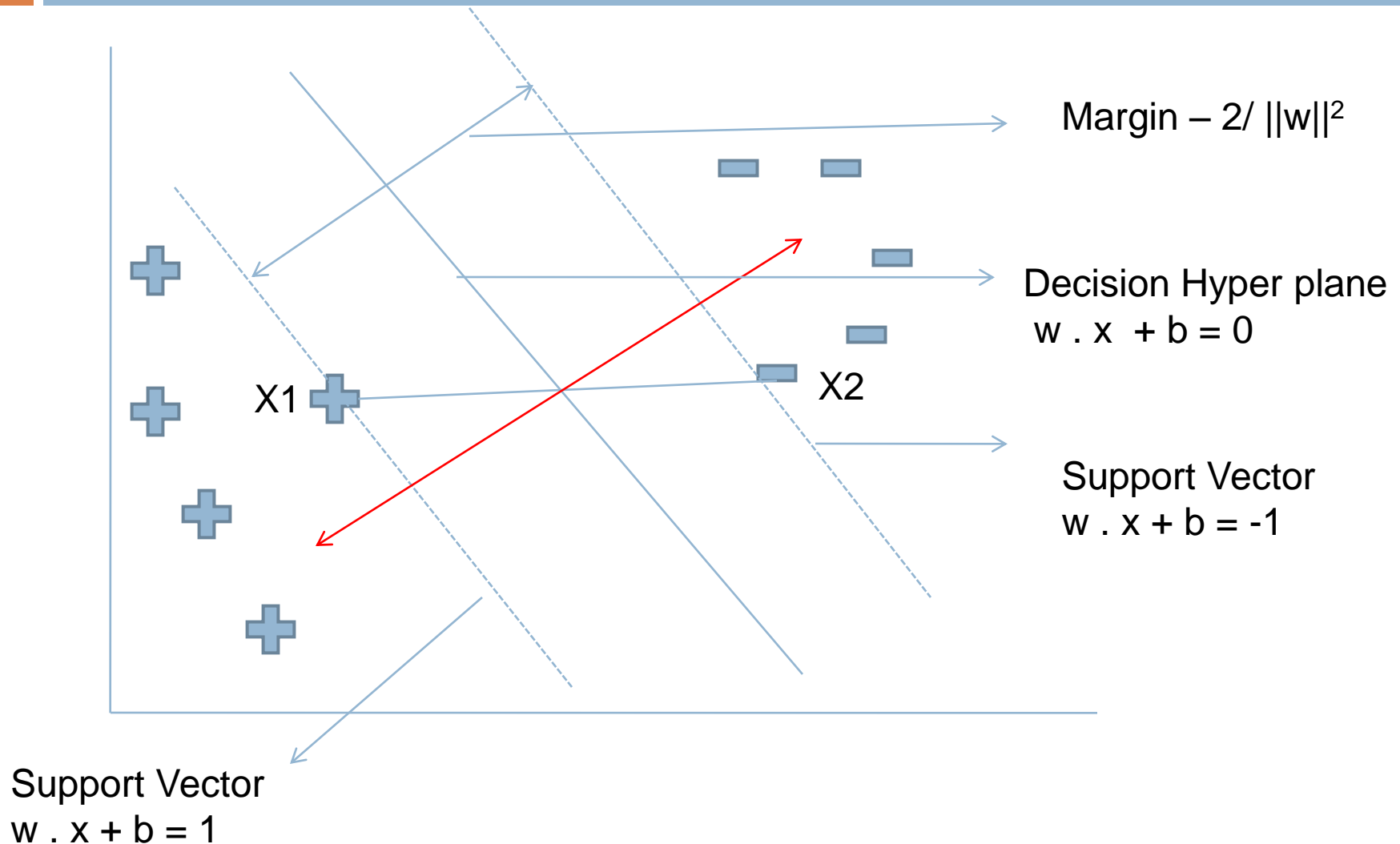
Linear Classifiers - 5



A Detour

- VC Theory
- Consequences
 - ▣ The upper bound on the generalization error does not depend on the dimensionality of the space.
 - ▣ The bound is minimized by maximizing the margin

SVMs for Binary Classification - 1



SVMs for Binary Classification - 2

- Separating Hyper plane is given by,
 - $w \cdot x + b = 0$
- Decision Function
 - $D(x) = \text{sign}(w \cdot x + b)$
- Invariant under scaling of w and b
 - $w = \lambda w$
 - $b = \lambda b$

SVMs for Binary Classification - 3

- Canonical Hyper Planes
 - $w \cdot x + b = 1$ and $w \cdot x + b = -1$
- $D(x)$
 - 1 if $w \cdot x + b \geq 1$
 - -1 if $w \cdot x + b \leq -1$
 - Not sure otherwise
- Margin is $w \cdot (x_1 - x_2) = 2$
- Margin is given by projection of vector $(x_1 - x_2)$ on normal vector to hyper plane ie $w / \|w\|$

SVMs for Binary Classification - 4

- Large Margin Classifier
- Maximize margin = $2/\|w\|^2$
- Or minimize $\|w\|^2 / 2$
- Constraints :
 - $y_i [(w \cdot x_i) + b] \geq 1$
 - Equivalently ,
 - $f(x)$ is
 - 1 if $w \cdot x + b \geq 1$
 - -1 if $w \cdot x + b \leq -1$

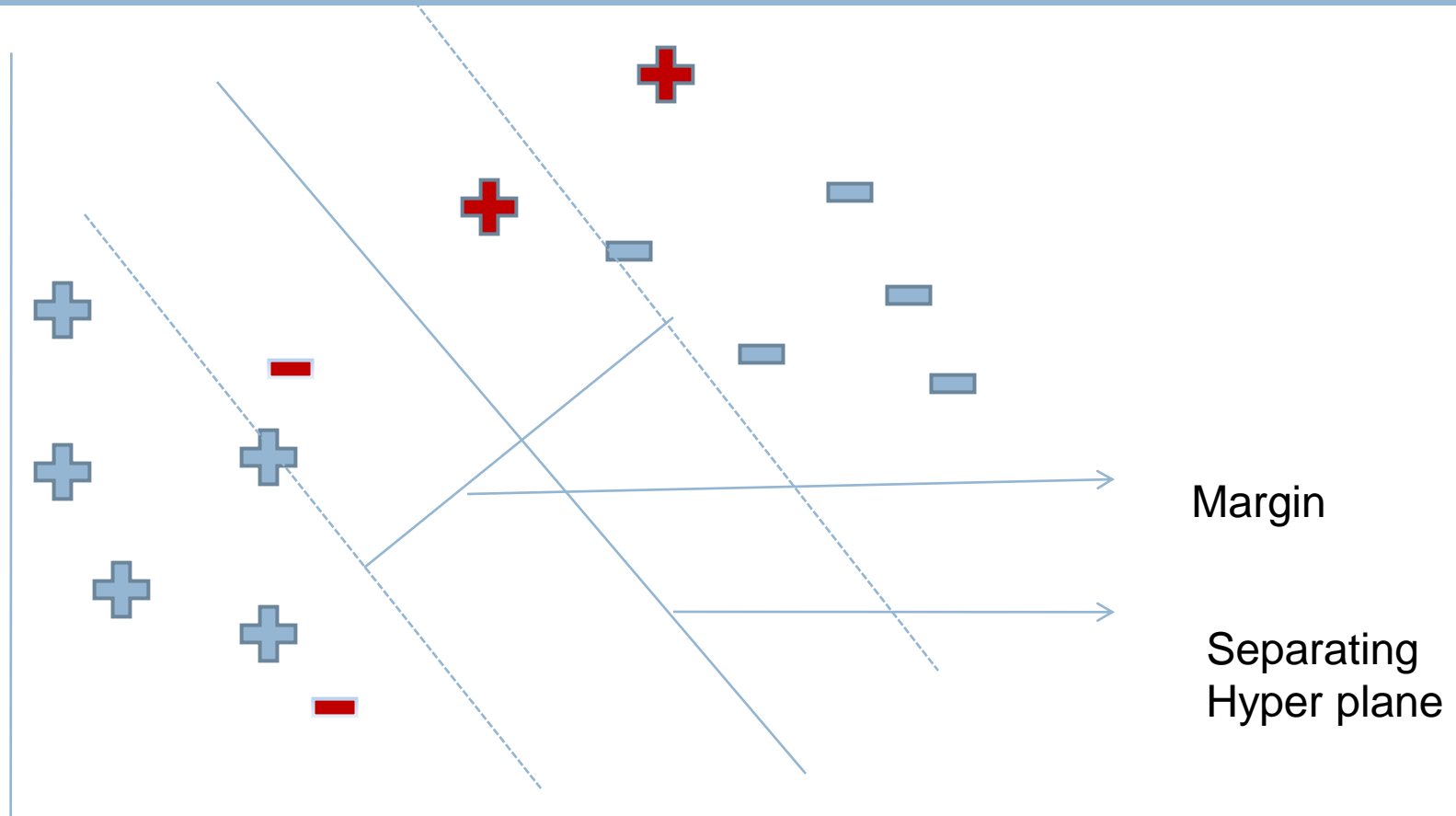
SVMs for Binary Classification - 5

- Constrained Optimization problem
- Lagrangian version :
 - $L(w,b) = \frac{1}{2} (w \cdot w) - \sum \alpha_i [y_i ((w \cdot x_i) + b) - 1]$
- $W(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$
- Constraints :
 - $\alpha_i \geq 0$
 - $\sum \alpha_i y_i = 0$
- $D(z) = \text{sign}(\sum \alpha_i y_i (x_i \cdot z) + b)$

The Story so far

- Linearly separable problem.
- Training set uniquely describes best separating hyper plane.
- Pass data to Quadratic optimization procedure.
- During testing for point z , substitute in $D(z)$ and return class based on sign and don't know otherwise.

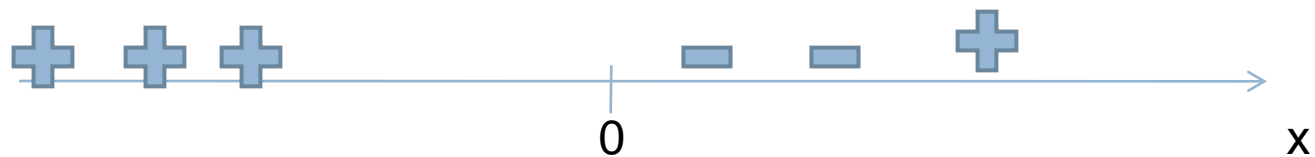
Soft Margin Classification - 1



Soft Margin Classification - 2

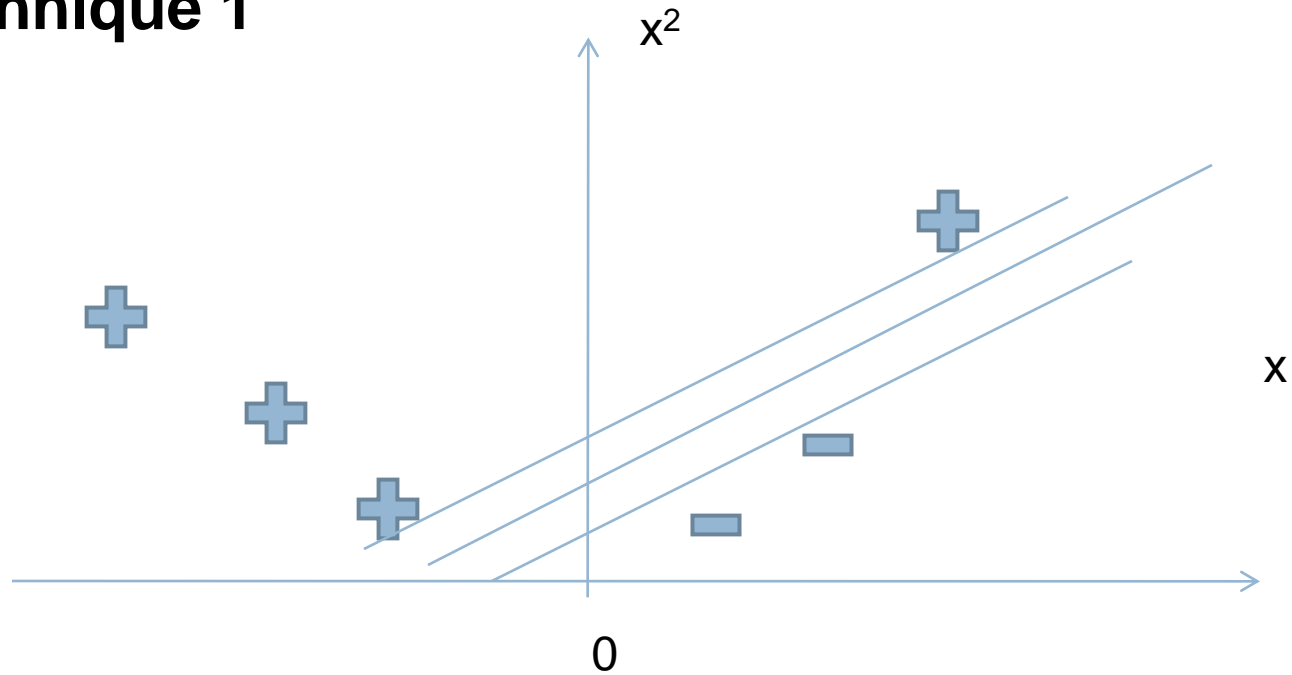
- Slack variables.
- L1 Error Norm : $0 \leq \alpha_i \leq C$
 - Impact of C
- Minimize : $\|w\|^2 / 2 + C (\sum \xi_i)$
- Decision function : $y_i (w \cdot x_i + b) \geq 1 - \xi_i$
- Equivalently,
 - 1 if $w \cdot x + b \geq 1 - \xi_i$
 - -1 if $w \cdot x + b \leq -1 + \xi_i$

Non Linear SVMs - 1



Non Linear SVMs - 2

Technique 1



We are using first consequence of VC theory.

Non Linear SVMs - 3

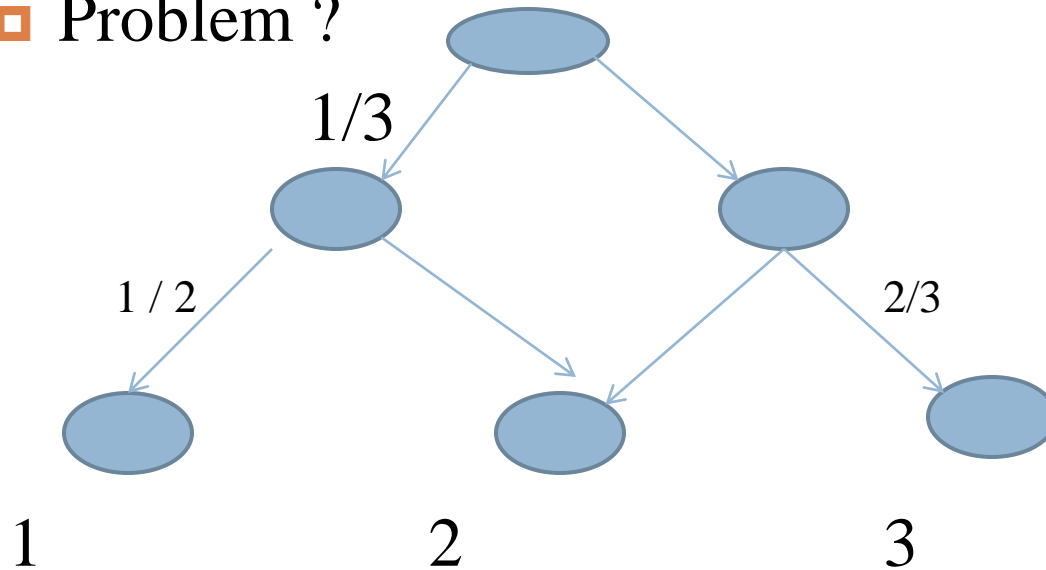
- Kernel trick
- $W(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$
- Data points x_i and x_j come as inner product.
- We can use a mapping function s.t.
 - $x_i \cdot x_j \rightarrow \phi(x_i) \cdot \phi(x_j)$
- $\phi(x_i) \cdot \phi(x_j)$ is the kernel and represented as $K(x_i \cdot x_j)$
- Feature space
- Constraint : Must be Hilbert space

Non Linear SVMs - 4

- Popular kernels :
 - Gaussian
 - Polynomial
- Mercer theorem.

Multi Class SVMs - 1

- Technique 1 :
 - ▣ Create a DAG by reducing learning to binary classification.
 - ▣ Problem ?



Multi Class SVMs - 2

- Technique 2 : One Versus All (OVA)
- Technique 3 :
 - Structural SVMs
 - Two class classifier over $\phi(x, y)$
 - Choose class, $y = \operatorname{argmax}_{y'} w \phi(x, y')$

Hierarchical Classification of Web Content using SVMs – 1

- Hierarchical Classification.
 - Examples : Yahoo directory , DMOZ , Usenet
 - Why not clustering ?
- Previous approaches and disadvantages
 - Flattening
 - Good features may not be useful discriminators.
 - Treating it as m-ary problem
 - Accuracy over large data and large features.

Hierarchical Classification of Web Content using SVMs - 2

- Data selection from LookSmart.
- Pre-processing
 - Extract plain text from web page.
 - Get meta data like title, description, keywords and ALT text.
 - Generate summaries.
 - Binary vector for each term for each page summary.
 - Top 1000 terms for each category using mutual information between term and category.
 - Term frequency and document length ignored.

Hierarchical Classification of Web Content using SVMs - 3

- Classification using SVMs :
 - Training phase.
 - SMO algorithm to solve optimization.
 - Testing
 - After learning weights, calculate $w \cdot x$.
 - Sum of weights of features as we store binary vector.
 - Algorithm also produces posterior probabilities – used to compare across categories.

Hierarchical Classification of Web Content using SVMs - 4

- Feature Selection :
 - Ignore rare words.
 - $MI(F,C) = \sum \sum P(F,C) \log P(F,C) / P(F) P(C)$
 - F over f and f' and C over c and c'
 - Used 1000 features.
 - SVM parameters :
 - $C = 0.01$
 - Results
 - Accuracy and Efficiency

Other Issues

- Important points :
 - Impact of outliers on hyper plane
 - Non necessity of non support vectors.
 - Learning rate
 - Classification efficiency.
- Applications

Conclusion

- VC theorem
- SVMs for Binary Classification
- SVMs with Soft Margins
- Non Linear SVMs
- Multi Class SVMs
- Advantages