EECS 445

Introduction to Machine Learning

Soft Margin SVMs
(Efficient) Feature Maps

Prof. Kutty

Today's Agenda

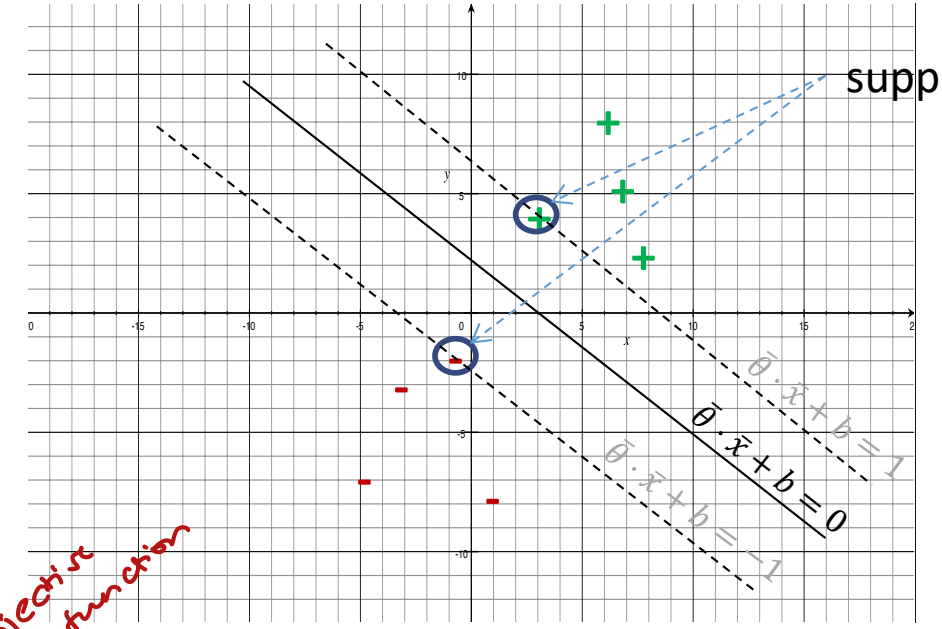
- Recap: Hard Margin SVMs
- Section 1: Soft Margin SVMs
- Section 2: Feature maps

https://forms.gle/ffiBvNbPjHF8ghi77

Recap: Hard Margin SVMs

Hard Margin SVM

Assuming data are linearly separable



support vectors lie on margin boundaries

Want:

- Boundary that classifies the training set correctly
- **That is maximally removed from training examples closest to the decision boundary**

objective function

QP : Quadratic Program

$$\min_{\bar{\theta},b} \frac{\|\bar{\theta}\|^2}{2} \text{ subject to } y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1 \text{ for } i \in \{1, \ldots, n\}$$

constraints

Linear classifier output by this QP: $sign(\bar{\theta} \cdot \bar{x} + b)$

Output of this optimization problem: $\bar{\theta}, b$

Linear Separability

What if data are not linearly separable?

How can we handle such cases?

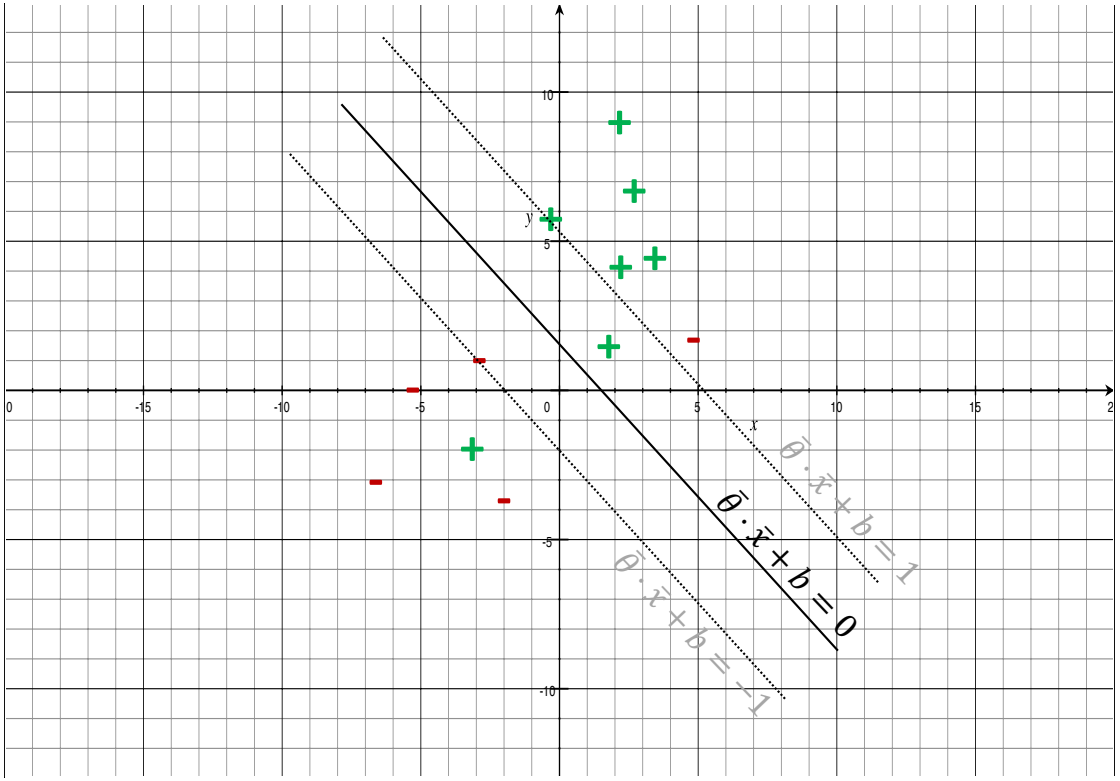
1. Constraints seem too restrictive
   - Fix the constraints: Soft-Margin SVMs

2. Map to a higher dimensional space

Soft Margin SVMs

Soft-Margin SVM

Suppose data are *not* linearly separable



$$\min_{\bar{\theta}, b, \bar{\xi}} \frac{\|\bar{\theta}\|^2}{2} + C \sum_{i=1}^{n} \xi_i$$

slack variables

$\bar{\theta} \in \mathbb{R}^d; b \in \mathbb{R}$

$\bar{\xi} \in \mathbb{R}^n$

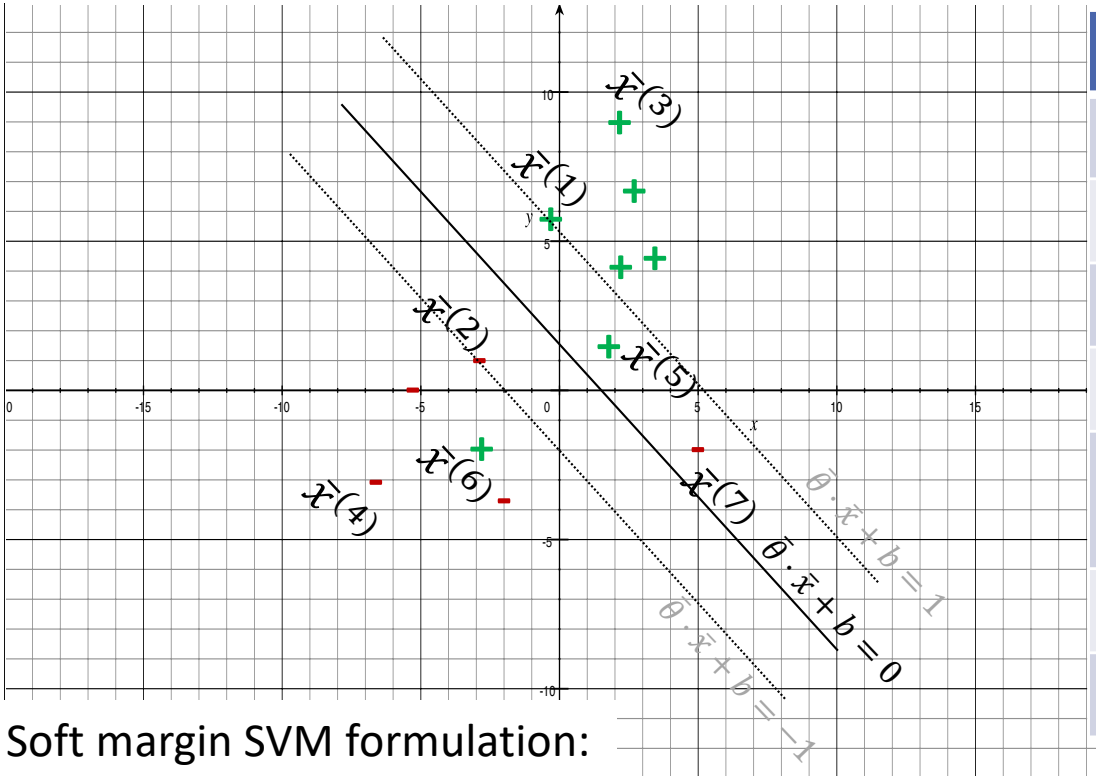
subject to $y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1 - \xi_i$ soft constraints

for $i \in \{1, \ldots, n\}$

penalty

Soft-Margin SVM

Suppose data are *not* linearly separable



CONSTRAINT	PENALTY (ξ_i)
$y^{(1)}\left(\bar{\theta} \cdot \bar{x}^{(1)} + b\right) = 1$	
$y^{(2)}\left(\bar{\theta} \cdot \bar{x}^{(2)} + b\right) = 1$	
$y^{(3)}\left(\bar{\theta} \cdot \bar{x}^{(3)} + b\right) > 1$	
$y^{(4)}\left(\bar{\theta} \cdot \bar{x}^{(4)} + b\right) > 1$	
$y^{(5)}\left(\bar{\theta} \cdot \bar{x}^{(5)} + b\right) < 1$	
$y^{(6)}\left(\bar{\theta} \cdot \bar{x}^{(6)} + b\right) < 1$	
$y^{(7)}\left(\bar{\theta} \cdot \bar{x}^{(7)} + b\right) < 1$	

Soft margin SVM formulation:

https://forms.gle/ffiBvNbPjHF8ghi77

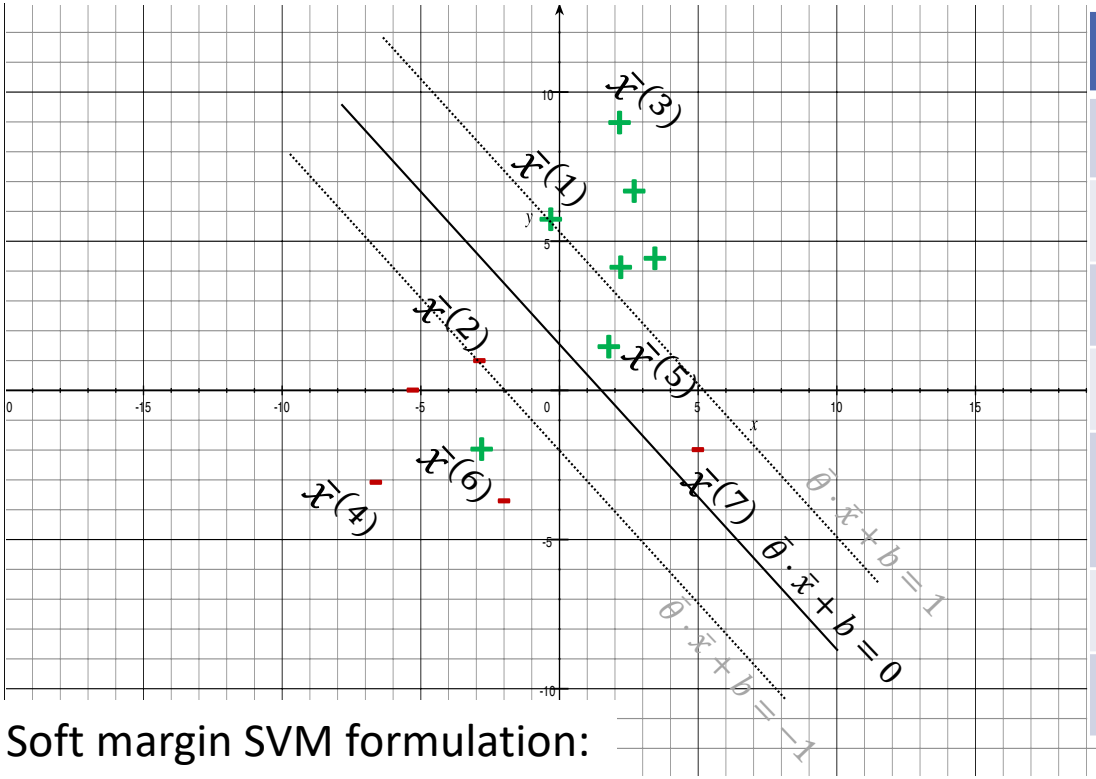
$$\min_{\bar{\theta}, b, \bar{\xi}} \frac{\|\bar{\theta}\|^2}{2} + C \sum_{i=1}^{n} \xi_i$$

subject to $y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1 - \xi_i$

for $i \in \{1, \ldots, n\}$ and $\xi_i \geq 0$

Soft-Margin SVM

Suppose data are *not* linearly separable



CONSTRAINT	PENALTY (ξ_i)
$y^{(1)}\left(\bar{\theta} \cdot \bar{x}^{(1)} + b\right) = 1$	$\xi_1 = 0$
$y^{(2)}\left(\bar{\theta} \cdot \bar{x}^{(2)} + b\right) = 1$	$\xi_2 = 0$
$y^{(3)}\left(\bar{\theta} \cdot \bar{x}^{(3)} + b\right) > 1$	$\xi_3 = 0$
$y^{(4)}\left(\bar{\theta} \cdot \bar{x}^{(4)} + b\right) > 1$	$\xi_4 = 0$
$y^{(5)}\left(\bar{\theta} \cdot \bar{x}^{(5)} + b\right) < 1$	$0 < \xi_5 < 1$
$y^{(6)}\left(\bar{\theta} \cdot \bar{x}^{(6)} + b\right) < 1$	$\xi_6 > 1$
$y^{(7)}\left(\bar{\theta} \cdot \bar{x}^{(7)} + b\right) < 1$	$\xi_7 > 1$

Soft margin SVM formulation:

$$\min_{\bar{\theta}, b, \bar{\xi}} \frac{\|\bar{\theta}\|^2}{2} + C \sum_{i=1}^{n} \xi_i$$

subject to $y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1 - \xi_i$

for $i \in \{1, \ldots, n\}$ and $\xi_i \geq 0$

Soft-Margin SVM

Soft-margin SVM advantages

- can handle data that are not linearly separable
- reduce effect of outliers and less chances of overfitting

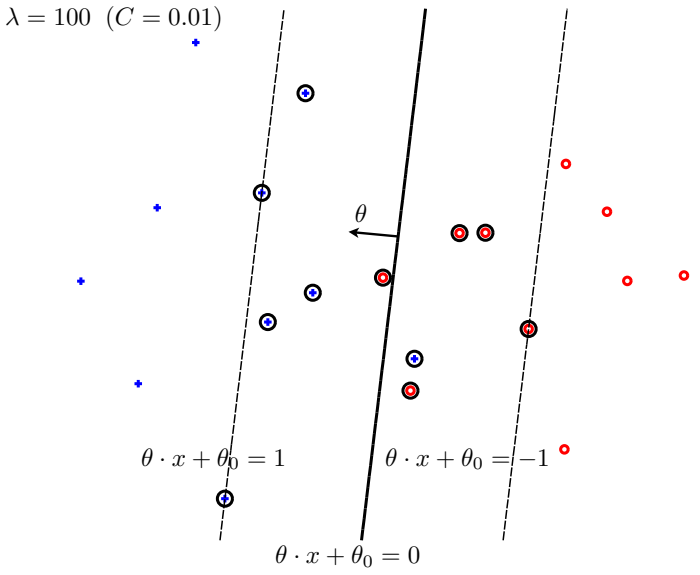


Image credit: Barzilay & Jaakkola

$$\min_{\bar{\theta},b,\bar{\xi}} \frac{\|\bar{\theta}\|^2}{2} + C \sum_{i=1}^{n} \xi_i$$

subject to $y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1 - \xi_i$

for $i \in \{1, \ldots, n\}$ and $\xi_i \geq 0$

hyperparameter

Soft-Margin SVM: hyperparameter

$$\min_{\bar{\theta},b,\bar{\xi}} \quad \frac{\|\bar{\theta}\|^2}{2} + C\sum_{i=1}^{n}\xi_i$$

subject to $y^{(i)}\left(\bar{\theta}\cdot\bar{x}^{(i)}+b\right) \geq 1-\xi_i$

$\xi_i \geq 0$ for $i \in \{1,\ldots,n\}$

Intuitively

as $C\uparrow$, penalty on errors/misclassifications ↑

as $C \to \infty$, in the limit this is the hard margin SVM

Soft-Margin SVM hyperparameter

for higher values of C

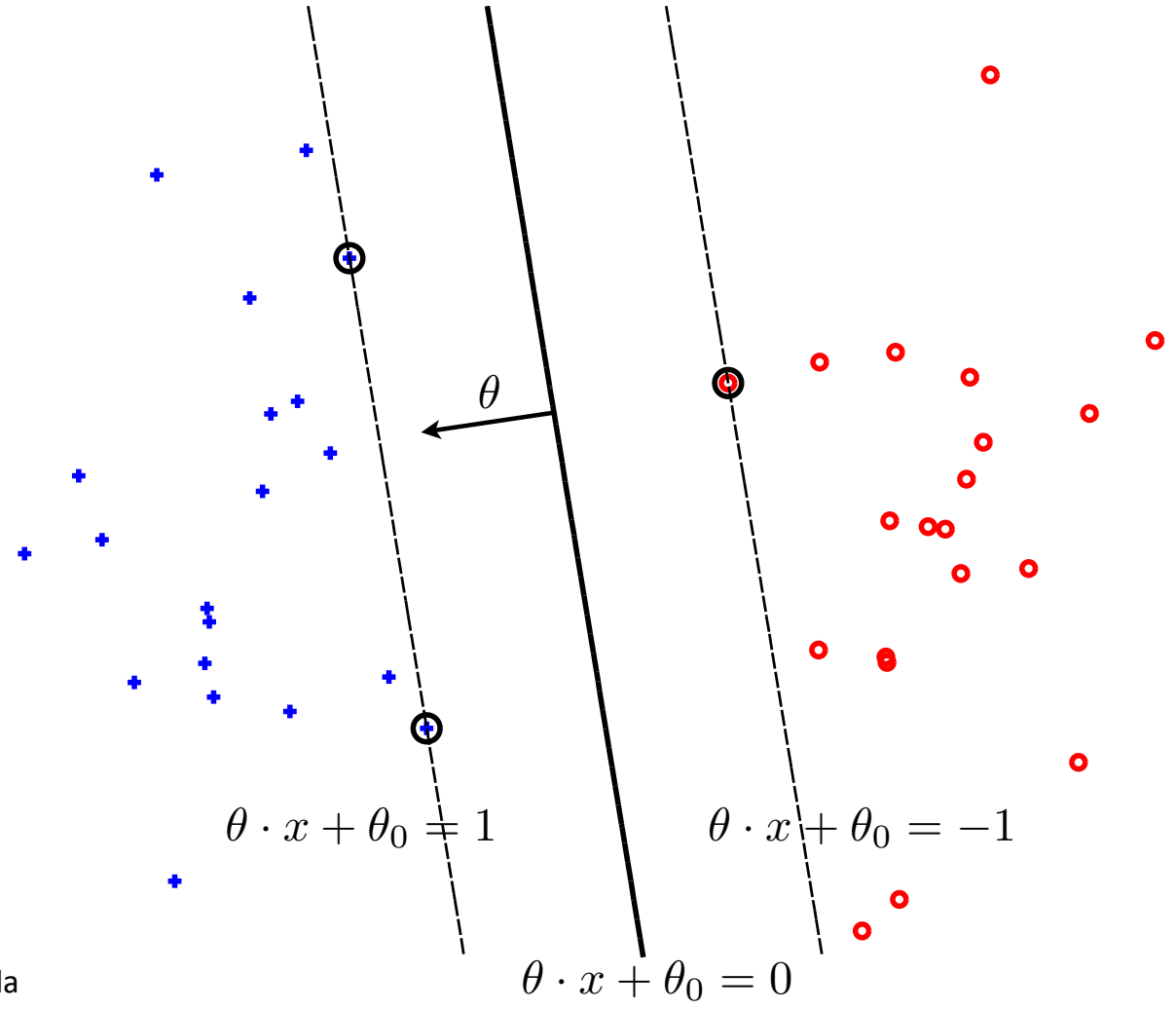


Image credit: Barzilay & Jaakkola

Soft-Margin SVM hyperparameter

for lower values of C

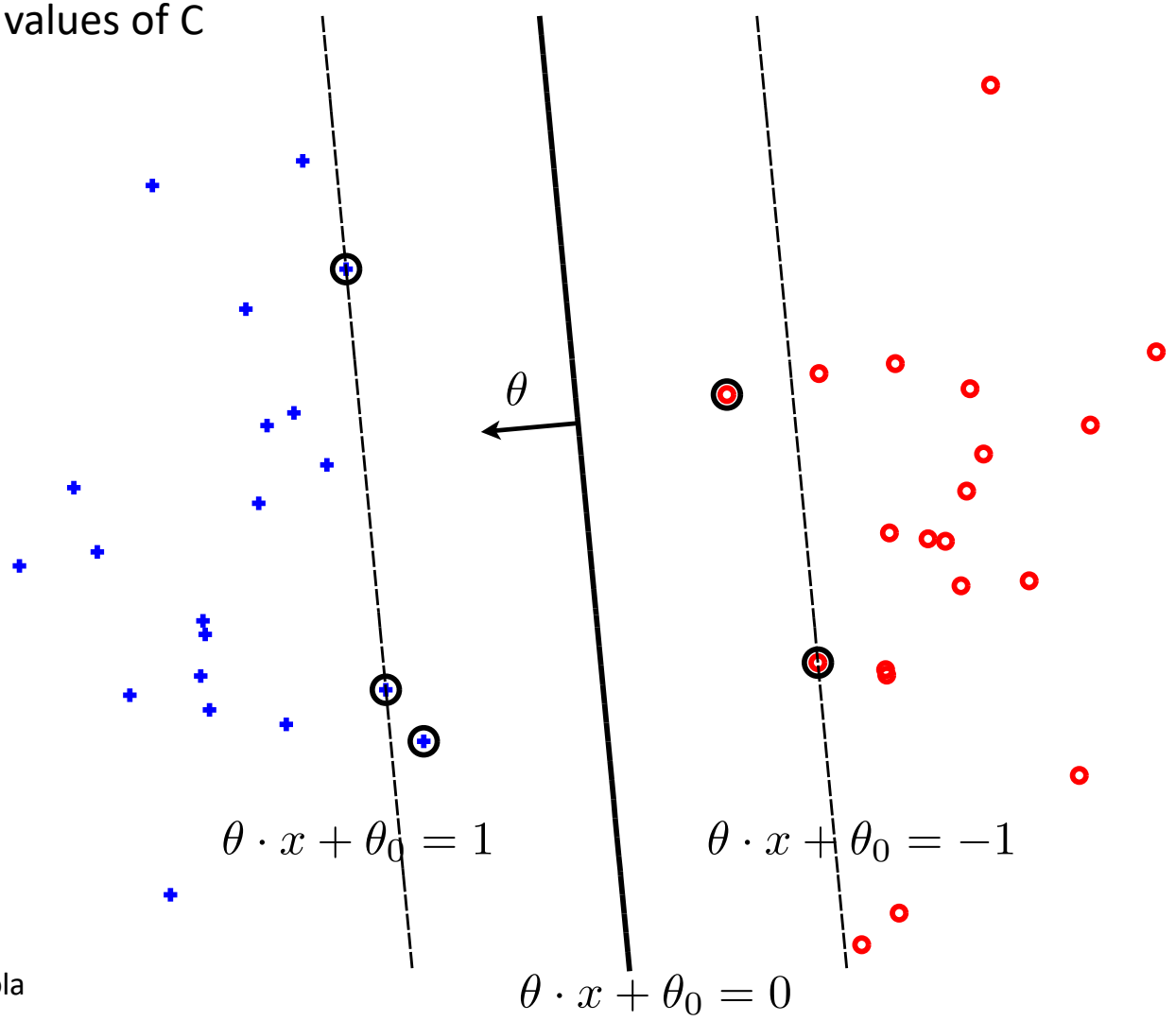


Image credit: Barzilay & Jaakkola

Soft-Margin SVM: exercise

Claim: Soft margin SVM is an optimization problem with hinge loss as objective function and ℓ_2-norm regularizer

$$\min_{\bar{\theta},b,\bar{\xi}} \quad \frac{\|\bar{\theta}\|^2}{2} + C\sum_{i=1}^{n}\xi_i \quad \text{subject to } y^{(i)}\left(\bar{\theta}\cdot\bar{x}^{(i)}+b\right) \geq 1-\xi_i \text{ and } \xi_i \geq 0$$

$$\text{for } i \in \{1,\dots,n\}$$

Hints:

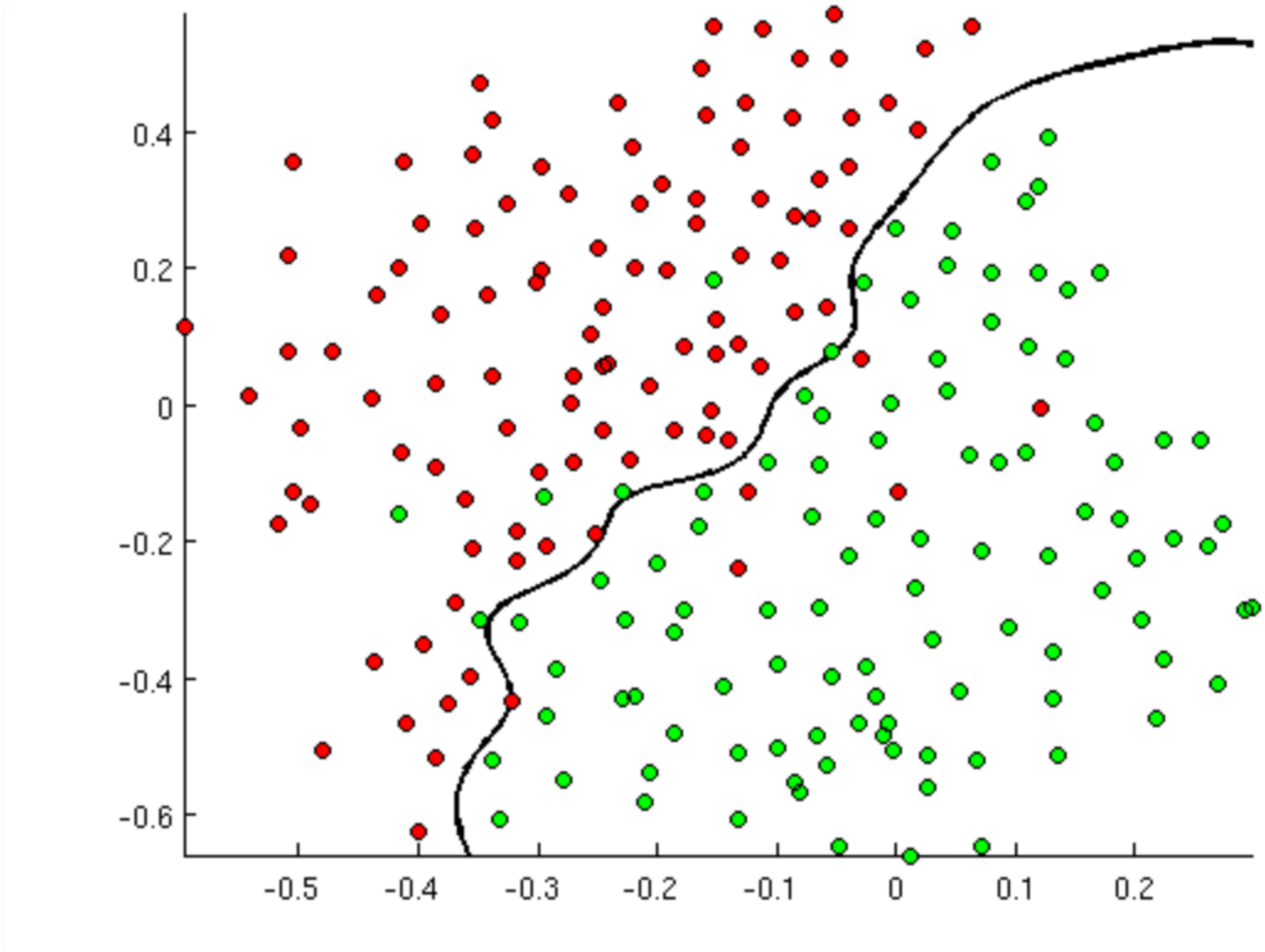
- Write $\xi_i \geq 1 - y^{(i)}\left(\bar{\theta}\cdot\bar{x}^{(i)}+b\right)$ and $\xi_i \geq 0$
- Observe that the objective function includes the terms $\min_{\bar{\xi}} \sum_{i=1}^{n}\xi_i$

$$\min_{\bar{\theta},b} \quad R_n(\bar{\theta}) + \lambda\, Z(\bar{\theta})$$

$R_n(\bar{\theta})$: empirical risk with hinge loss

$Z(\bar{\theta})$: penalty term associated model complexity

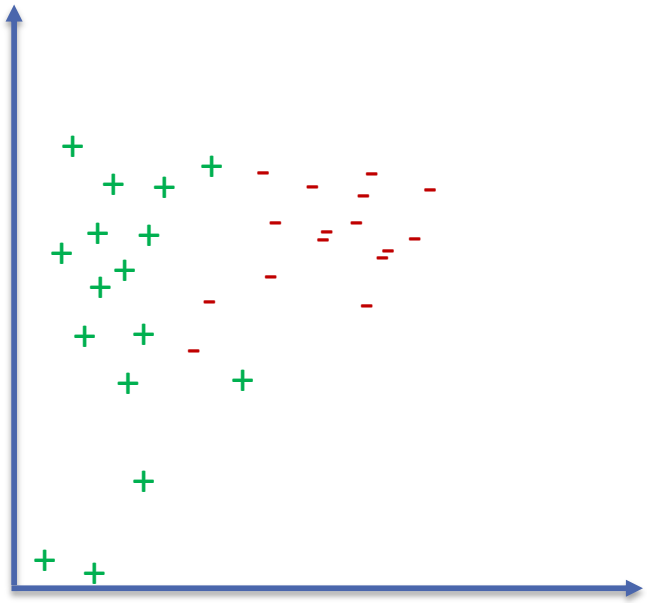
Non-linear decision boundaries with SVMs

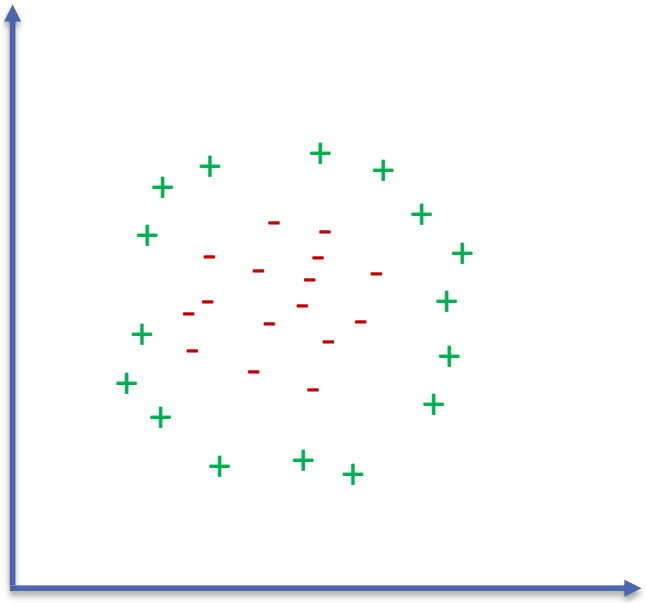


Feature maps

https://forms.gle/ffiBvNbPjHF8ghi77

Data that are not linearly separable





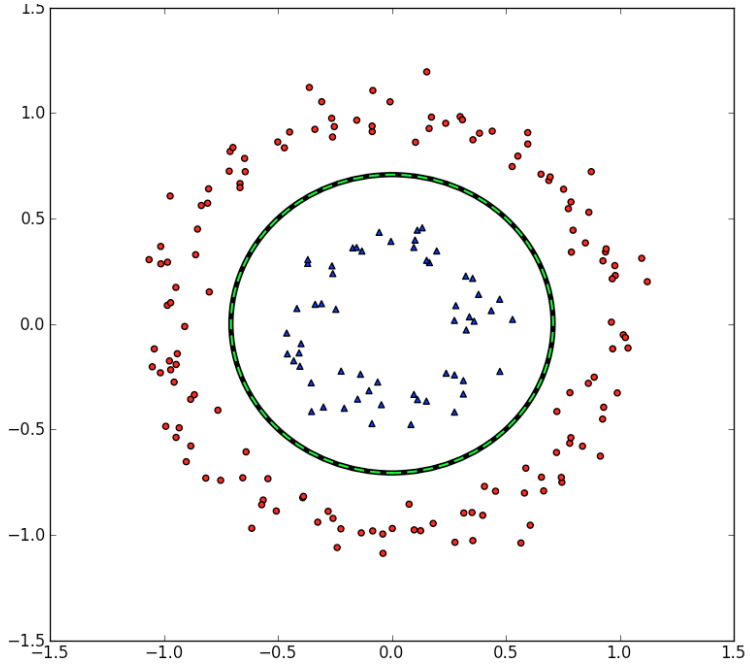
Idea: minimize empirical risk with hinge loss using gradient descent

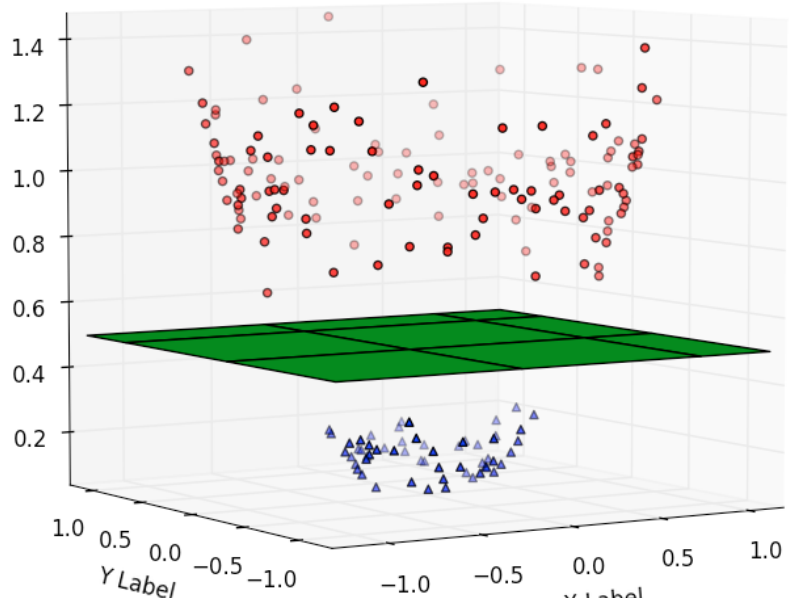
Idea: Soft Margin SVMs

Idea: feature maps

Linear classifiers in higher dimensional spaces: idea

image source: http://www.eric-kim.net





map data to a higher dim. space in which there exists a separating hyperplane (corresponds to a non-linear decision boundary in the original space)

Linear classifiers in higher dimensional spaces: exercise

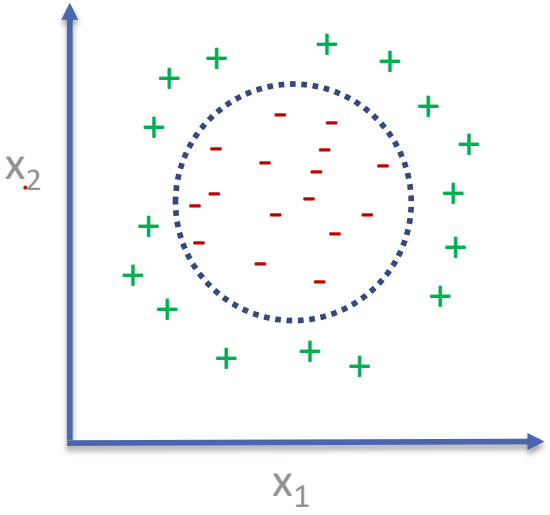
$$(x_1 - 2)^2 + (x_2 - 2)^2 = 1^2$$

$$\Leftrightarrow$$

$$x_1^2 + x_2^2 - 4x_1 - 4x_2 + 7 = 0$$

$$\begin{bmatrix} 1 \\ 1 \\ -4 \\ -4 \end{bmatrix} \cdot \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 \\ x_2 \end{bmatrix} + 7 = 0$$

$$\bar{\theta} \cdot \phi(\bar{x}) + b = 0$$



$\bar{x} \in \mathbb{R}^2$

'ideal' decision boundary

center at (2,2)

radius = 1

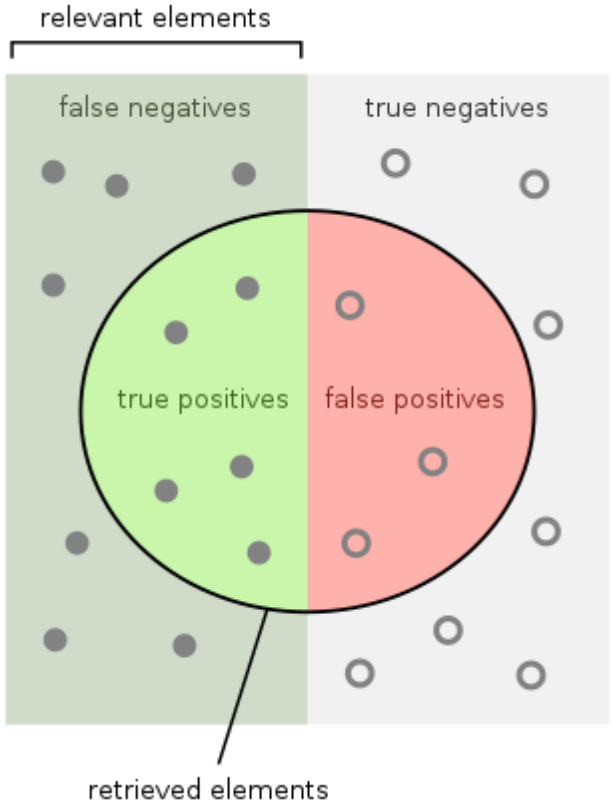
$$\phi(\bar{x}) = [x_1^2 \quad x_2^2 \quad x_1 \quad x_2 \quad 1]$$

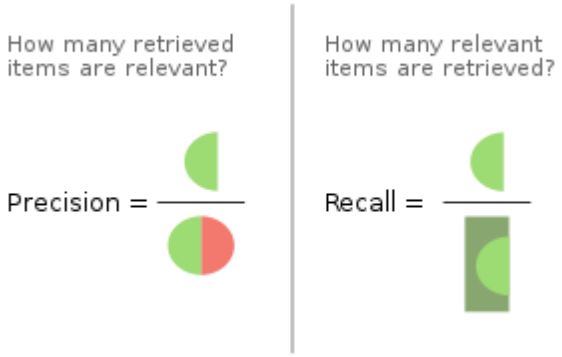
$$\bar{\theta} = [1 \quad 1 \quad -4 \quad -4 \quad 7]$$

Linear Classifier $\text{sign}(\bar{\theta} \cdot \phi(\bar{x}))$ separates positive from negative examples

Evaluating Classifiers

image source https://commons.wikimedia.org/w/index.php?title=File:Precisionrecall.svg&oldid=697135202





Common Performance Metrics

Classification Error:	FP+FN
Accuracy:	(TP+TN)/N
False Positive Rate:	FP/(TN+FP)
True Positive Rate:	TP/(TP+FN)
Sensitivity (Recall):	TPR
Precision:	TP/(TP+FP)
Specificity:	TN/(TN+FP)

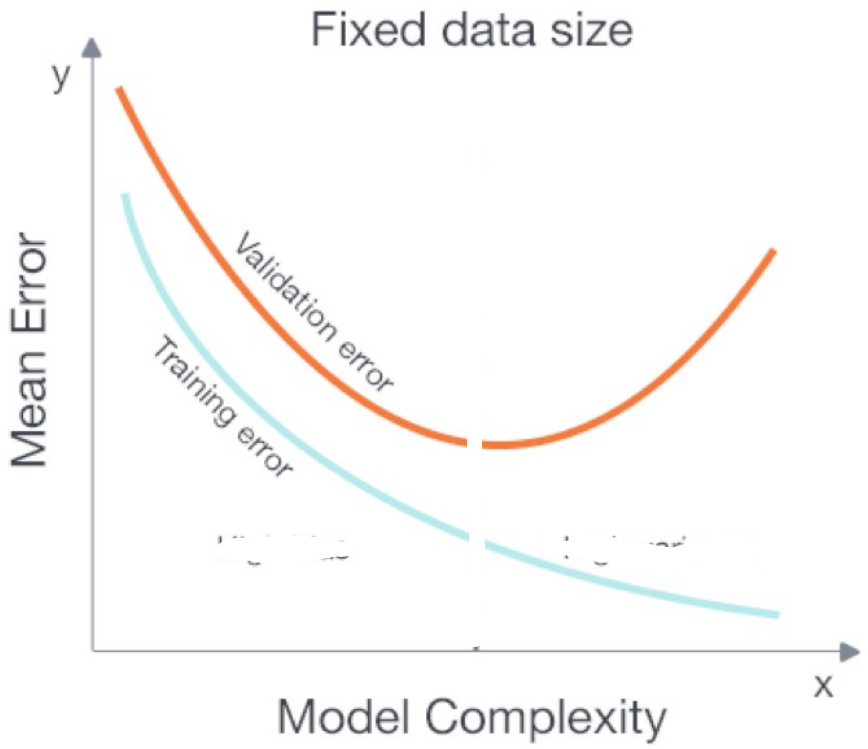
F1 score:
harmonic mean of
precision and recall

Confusion Matrix

Actual Class		*Predicted Class* Pos.	Neg.
	Pos.	#TP	#FN
	Neg.	#FP	#TN

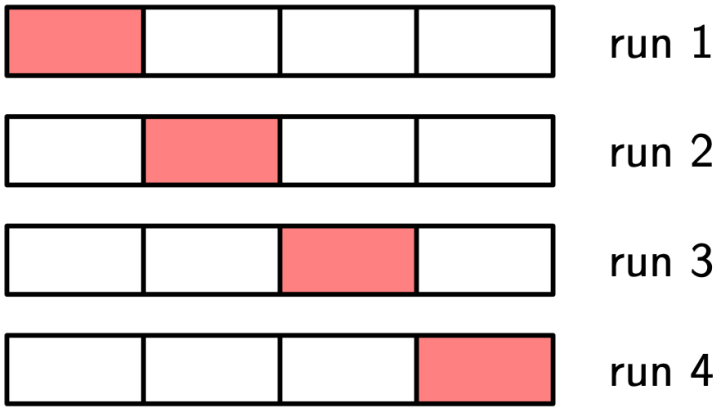
How to measure generalization error?

- *training dataset*
- *validation dataset*
- *test dataset*



k-fold Cross Validation

The technique of S-fold cross-validation, illustrated here for the case of $S = 4$, involves taking the available data and partitioning it into S groups (in the simplest case these are of equal size). Then $S - 1$ of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is then repeated for all S possible choices for the held-out group, indicated here by the red blocks, and the performance scores from the S runs are then averaged.



Note: we say k-fold Cross Validation

image source: Bishop 2006

Implications for SVM

$$\min_{\bar{\theta}, b, \bar{\xi}} \quad \frac{\|\bar{\theta}\|^2}{2} + C \sum_{i=1}^{n} \xi_i$$

subject to $y^{(i)}\left(\bar{\theta} \cdot \phi(\bar{x}^{(i)}) + b\right) \geq 1 - \xi_i$

and $\xi_i \geq 0$ for $i \in \{1, \ldots, n\}$

$\bar{\theta} \in \mathbb{R}^p$

$S_n = \{(\bar{x}^{(i)}, y^{(i)})\}_{i=1}^n$

$S_n' = \{(\phi(\bar{x}^{(i)}), y^{(i)})\}_{i=1}^n$

$\phi: \mathbb{R}^d \to \mathbb{R}^p$

Issue: potentially *very* inefficient

Linear Separability

What if data are not linearly separable?

How can we handle such cases?

1. Soft-Margin SVMs

2. Map to a higher dimensional space

3. SVM dual and the kernel trick

Support Vector Machines

QP formulation

$$\min_{\bar{\theta}} \frac{\|\bar{\theta}\|^2}{2} \text{ subject to } y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)}\right) \geq 1 \text{ for } i \in \{1, \ldots, n\}$$

Goal: rewrite in dual form

$$\max_{\bar{\alpha}, \alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)} \cdot \bar{x}^{(j)}$$

Dual Formulation in General

https://forms.gle/ffiBvNbPjHF8ghi77

Relating the Lagrangian to $f(\bar{w})$

primal variables

original problem: $\min_{\bar{w}} f(\bar{w}) \quad \text{s.t.} \quad h_i(\bar{w}) \leq 0 \quad \text{for } i = 1, \ldots, n$

Lagrangian: $L(\bar{w}, \bar{\alpha}) = f(\bar{w}) + \sum_{i=1}^{n} \alpha_i h_i(\bar{w}) \qquad \alpha_i \geq 0$

↳ dual

Define: $g_p(\bar{w}) = \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

Claim: $g_p(\bar{w}) = \begin{cases} f(\bar{w}), & \text{if constraints are satisfied} \\ \infty, & \text{otherwise} \end{cases}$

Case 1: constraints are satisfied

$$g_p(\bar{w}) = \max_{\bar{\alpha}} \left(f(\bar{w}) + \sum_{i=1}^{n} \alpha_i h_i(\bar{w}) \right)$$

by definition

$\alpha_i \cdot h_i(\bar{w}) = 0$

$$= \max_{\bar{\alpha}} \left(f(\bar{w}) + \alpha_1 h_1(\bar{w}) + \cdots + \alpha_j h_j(\bar{w}) + \cdots + \alpha_n h_n(\bar{w}) \right)$$

$= f(\bar{w})$

Case 2: constraints are *not* satisfied

$$g_p(\bar{w}) = \max_{\bar{\alpha}} \left(f(\bar{w}) + \alpha_1 h_1(\bar{w}) + \cdots + \alpha_j h_j(\bar{w}) + \cdots + \alpha_n h_n(\bar{w}) \right)$$

> 0

$= \infty$

Primal formulation

original problem: $\min_{\bar{w}} f(\bar{w})$ s. t. $h_i(\bar{w}) \leq 0$ for $i = 1, \dots, n$

Lagrangian: $L(\bar{w}, \bar{\alpha}) = f(\bar{w}) + \sum_{i=1}^{n} \alpha_i h_i(\bar{w})$ $\alpha_i \geq 0$

Define: $g_p(\bar{w}) = \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

$$g_p(\bar{w}) = \begin{cases} f(\bar{w}), & \text{if constraints are satisfied} \\ \infty, & \text{otherwise} \end{cases}$$

Note that $\min_{\bar{w}} g_p(\bar{w}) = \min_{\bar{w}} \max_{\bar{\alpha}} L(\bar{w}, \bar{\alpha}) = \min_{\bar{w}} f(\bar{w})$

if constraints are satisfiable

Primal formulation $\min_{\bar{w}} \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

Primal vs Dual formulation

original problem: $\min_{\bar{w}} f(\bar{w})$ s. t. $h_i(\bar{w}) \leq 0$ for $i = 1, \dots, n$

Lagrangian: $L(\bar{w}, \bar{\alpha}) = f(\bar{w}) + \sum_{i=1}^{n} \alpha_i h_i(\bar{w})$ $\alpha_i \geq 0$

Define: $g_p(\bar{w}) = \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

$$g_p(\bar{w}) = \begin{cases} f(\bar{w}), & \text{if constraints are satisfied} \\ \infty, & \text{otherwise} \end{cases}$$

Primal formulation $\min_{\bar{w}} \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

Dual formulation $\max_{\bar{\alpha}, \alpha_i \geq 0} \min_{\bar{w}} L(\bar{w}, \bar{\alpha})$