Margin, Duality and Stability

CS4780/5780 – Machine Learning
Fall 2019

Nika Haghtalab & Thorsten Joachims
Cornell University

Reading: UML 15.3, 15.4

Example: Margin in High-Dimension

Training Sample S_{train}	$\vec{x}$							y
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
$(\vec{x}_1, y_1)$	1	0	0	1	0	0	0	1
$(\vec{x}_2, y_2)$	1	0	0	0	1	0	0	1
$(\vec{x}_3, y_3)$	0	1	0	0	0	1	0	-1
$(\vec{x}_4, y_4)$	0	1	0	0	0	0	1	-1
	$\vec{w}$							**b**
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	
Hyperplane 1	1	1	0	0	0	0	0	0
Hyperplane 2	0	0	0	1	1	-1	-1	0
Hyperplane 3	1	-1	0	0	0	0	0	0
Hyperplane 4	1	-1	1	0	0	0	0	0
Hyperplane 5	0.95	-0.95	0	0.05	0.05	-0.05	-0.05	0
Hyperplane 6	0.67	-0.67	0	0.33	0.33	-0.33	-0.33	0

(Batch) Perceptron Algorithm

Input: $S = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n))$, $\vec{x}_i \in \Re^N$, $y_i \in \{-1, 1\}$,
$I \in [1, 2, ..]$

Algorithm:

- $\vec{w}_0 = \vec{0}$, $k = 0$
- repeat
  - FOR i=1 TO n
    - IF $y_i(\vec{w}_k \cdot \vec{x}_i) \leq 0$ ### makes mistake
      - $\vec{w}_{k+1} = \vec{w}_k + y_i\vec{x}_i$
      - $k = k + 1$
    - ENDIF
  - ENDFOR
- until I iterations reached

Dual (Batch) Perceptron Algorithm

Input: $S = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n))$, $\vec{x}_i \in \Re^N$, $y_i \in \{-1, 1\}$,
$I \in [1, 2, ..]$

Dual Algorithm:

- $\forall i \in [1..n] : \alpha_i = 0$
- repeat
  - FOR i=1 TO n
    - IF $y_i\left(\sum_{j=1}^{n} \alpha_j y_j (\vec{x}_j \cdot \vec{x}_i)\right) \leq 0$
      - $\alpha_i = \alpha_i + 1$
    - ENDIF
  - ENDFOR
- until I iterations reached

Primal Algorithm:

- $\vec{w} = \vec{0}$, $k = 0$
- repeat
  - FOR i=1 TO n
    - IF $y_i(\vec{w} \cdot \vec{x}_i) \leq 0$
      - $\vec{w} = \vec{w} + y_i \vec{x}_i$
    - ENDIF
  - ENDFOR
- until I iterations reached

SVM Solution as Linear Combination

- Primal OP:

$$\begin{aligned} &\text{minimize:} && P(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\vec{w}\cdot\vec{w} + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to:} && \forall_{i=1}^{n} : y_i[\vec{w}\cdot\vec{x}_i + b] \geq 1 - \xi_i \\ &&& \forall_{i=1}^{n} : \xi_i \geq 0 \end{aligned}$$

- Theorem: The solution $\vec{w}^*$ can always be written as a linear combination

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i$$

  of the training vectors with $0 \leq \alpha_i \leq C$.
- Properties:
  - Factor α_i indicates “influence” of training example (x_i, y_i).
  - If $\xi_i > 0$, then $\alpha_i = C$.
  - If $0 \leq \alpha_i < C$, then $\xi_i = 0$.
  - (x_i, y_i) is a Support Vector, if and only if $\alpha_i > 0$.
  - If $0 < \alpha_i < C$, then $y_i(x_i\, w^* + b) = 1$.
  - SVM-light outputs α_i using the “-a” option

Dual SVM Optimization Problem

- Primal Optimization Problem

$$\begin{aligned}
&\text{minimize:} && P(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\,\vec{w}\cdot\vec{w} + C \sum_{i=1}^{n} \xi_i \\
&\text{subject to:} && \forall_{i=1}^{n} : y_i[\vec{w}\cdot\vec{x}_i + b] \geq 1 - \xi_i \\
& && \forall_{i=1}^{n} : \xi_i \geq 0
\end{aligned}$$

- Dual Optimization Problem

$$\begin{aligned}
&\text{maximize:} && D(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\vec{x}_i \cdot \vec{x}_j) \\
&\text{subject to:} && \sum_{i=1}^{n} y_i \alpha_i = 0 \\
& && \forall_{i=1}^{n} : 0 \leq \alpha_i \leq C
\end{aligned}$$

- Theorem: If w* is the solution of the Primal and α* is the solution of the Dual, then

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \vec{x}_i$$

Leave-One-Out (i.e. n-fold CV)

- Training Set: $S = ((x_1, y_1), \dots, (x_n, y_n))$
- Approach: Repeatedly leave one example out for testing.

Train on	Test on
$(x_2,y_2), (x_3,y_3), (x_4,y_4), \dots, (x_n,y_n)$	(x_1,y_1)
$(x_1,y_1), (x_3,y_3), (x_4,y_4), \dots, (x_n,y_n)$	(x_2,y_2)
$(x_1,y_1), (x_2,y_2), (x_4,y_4), \dots, (x_n,y_n)$	(x_3,y_3)
…	…
$(x_1,y_1), (x_2,y_2), (x_3,y_3), \dots, (x_{n-1},y_{n-1})$	(x_n,y_n)

- Estimate: $err_{loo}(A) = \frac{1}{n} \sum_{i=1}^{n} \Delta(h_i(x_i), y_i)$
- Question: Is there a cheaper way to compute this estimate?

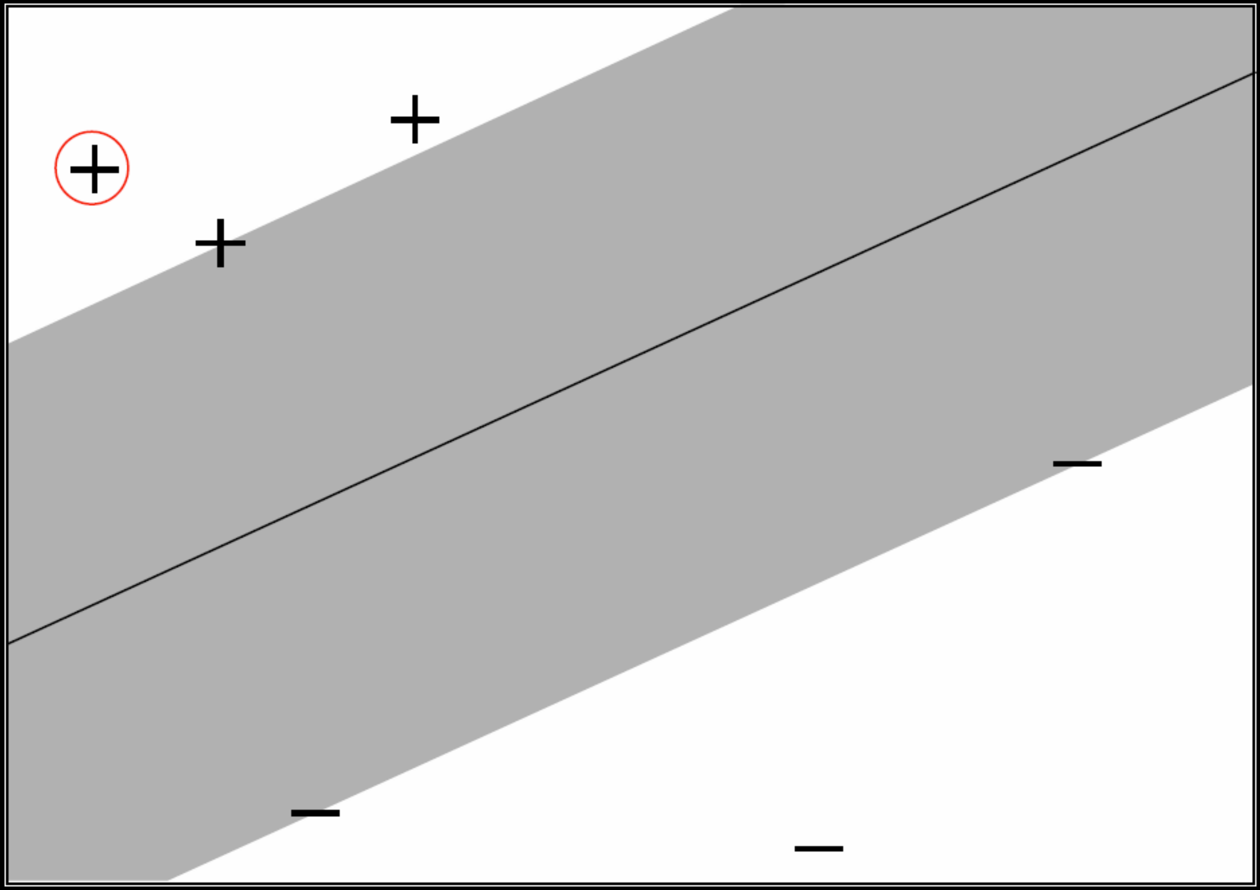
Necessary Condition for Leave-One-Out Error

- Lemma: For SVM, $[h_i(\vec{x}_i) \neq y_i] \Rightarrow [2\alpha_i R^2 + \xi_i \geq 1]$
- Input:
  - α_i dual variable of example i
  - ξ_i slack variable of example i
  - $||\vec{x}_i|| \leq R$ bound on length
- Example:

Value of $2\ \alpha_i\ R^2 + \xi_i$	Leave-one-out Error?
0.0	Must be Correct
0.7	Must be Correct
3.5	Error
0.1	Must be Correct
1.3	Correct
…	…

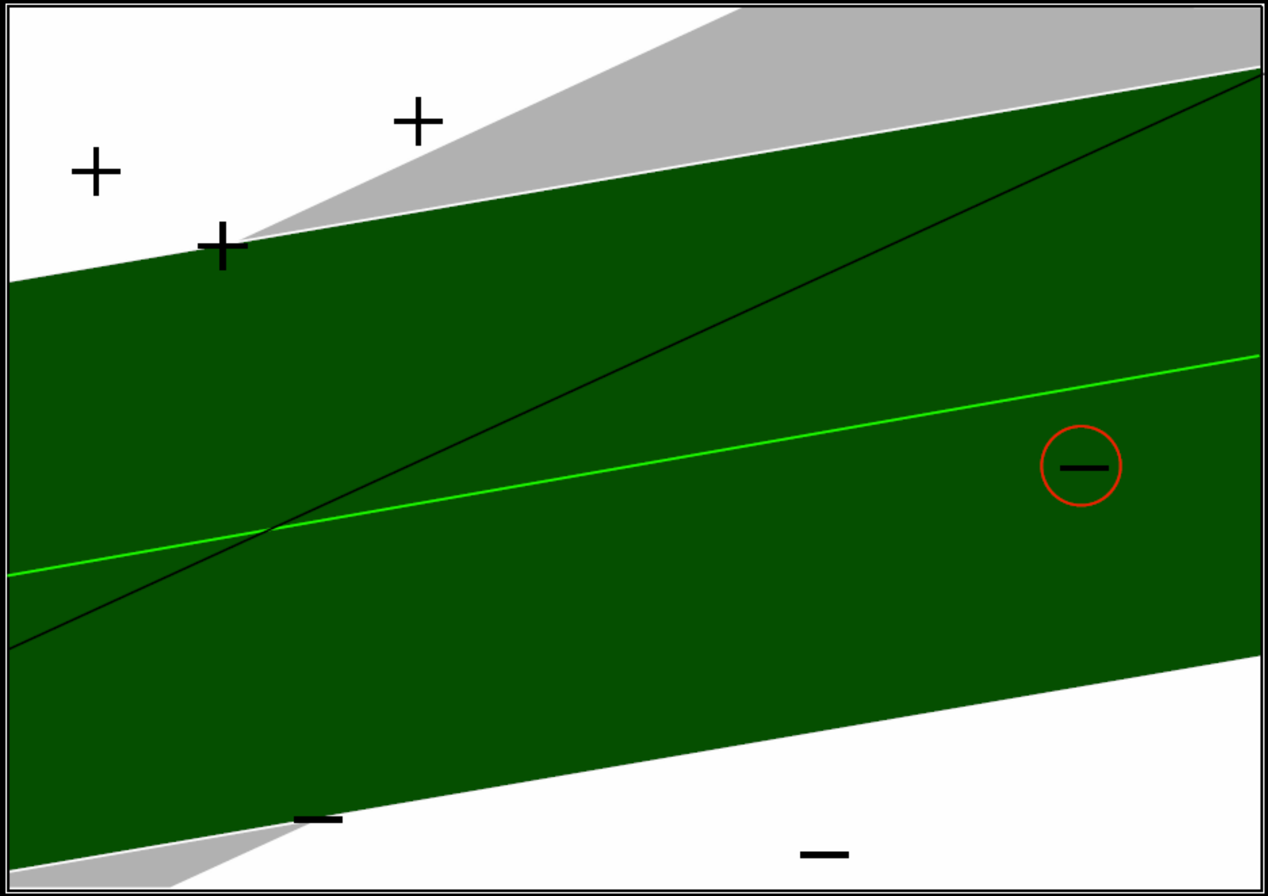
Case 1: Example is not SV

Criterion: (α_i = 0) and (ξ_i=0), so (2 α_i R^2 + ξ_i < 1): Correct



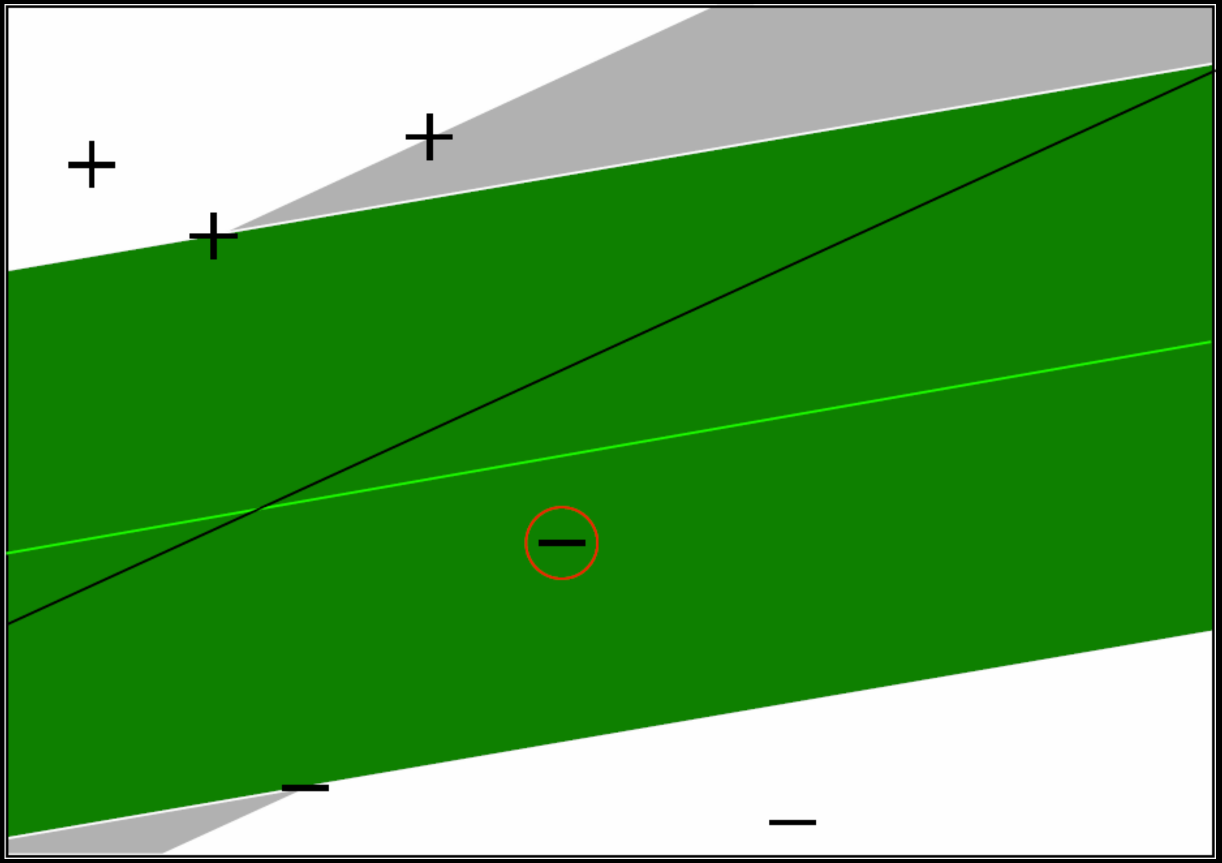
Case 2: Example is SV with Low Influence

Criterion: (α_i<0.5/R^2 < C) and (ξ_i=0), so ($2\alpha_i R^2+\xi_i$ < 1): Correct



Case 3: Example has Small Training Error

Criterion: ($\alpha_l = C$) and ($\xi_i < 1\text{-}2CR^2$), so ($2\alpha_i R^2 + \xi_i < 1$): Correct



Experiment: Reuters Text Classification

Experiment Setup

- 6451 Training Examples
- 6451 Test Examples to estimate true Prediction Error
- Comparison between Leave-One-Out upper bound and error on Test Set (average over 10 train/test splits)