Machine Learning and Intelligent Systems

Soft Margin SVM

Maria A. Zuluaga

Nov 17, 2023

EURECOM - Data Science Department

Table of contents

Recap: Hard Margin SVM

Soft Margin SVM

- Slack Variables
- Soft Margin Optimization Problem
- Unconstrained Soft SVM Formulation

Wrap-up

Recap: Hard Margin SVM

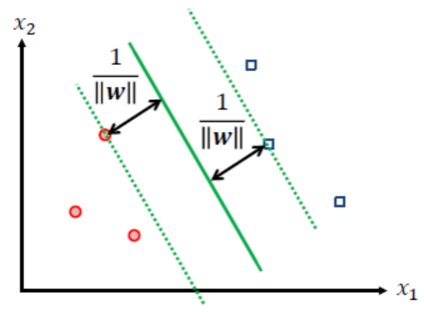
Hard Margin SVM

- **Data Assumptions:** Data is linearly separable
- The decision boundary is a hyperplane:

$$\mathcal{H} = \{\mathbf{x} : \hat{w}^T\mathbf{x} + w_0 = 0\}$$

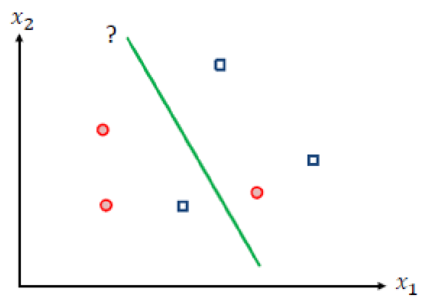
- **Goal:** Find a hyperplane that maximizes the margin width
- Hard margin SVM optimization problem

$$\begin{aligned} &\underset{\mathbf{w}, w_0}{\arg\min} && \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to} && \forall i \; y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \end{aligned}$$



Hard Margin SVM: Limitations

- Real data, most likely, will not meet the of linear separable assumption
- Hard margin loss is too limiting when there is class overlapping
- Hard margin SVM wont be able to deal with it

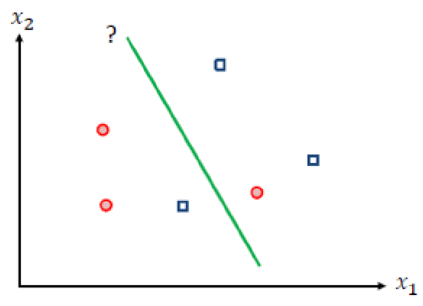


Hard Margin SVM: Limitations

- Real data, most likely, will not meet the of linear separable assumption
- Hard margin loss is too limiting when there is class overlapping
- Hard margin SVM wont be able to deal with it

Possible solutions:

1. Keep hard margin SVM but, transform the data (kernels)
2. **Relax the constraints**
3. Combination of both

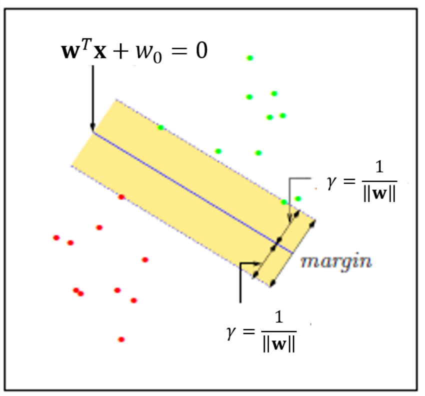


Hard Margin SVM: Loss Function

- Penalization in the hard SVM can be expressed as:

$$l(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } 1 - \mathbf{y}(\mathbf{w}^T\mathbf{x} + w_0) < 0 \\ \infty & \text{otherwise} \end{cases}$$

- Implicitly, hard SVM used a function giving infinite error if a data point was misclassified


$\mathbf{w}^T\mathbf{x} + w_0 = 0$
$\gamma = \frac{1}{\|\mathbf{w}\|}$
margin
$\gamma = \frac{1}{\|\mathbf{w}\|}$

Adapted from Figure 12.1 From ESL.

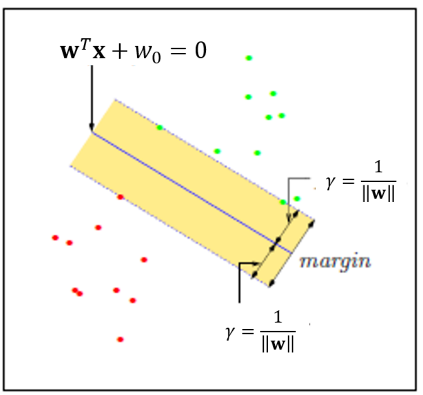
Hard Margin SVM: Loss Function

- Penalization in the hard SVM can be expressed as:

$$l(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } 1 - \mathbf{y}(\mathbf{w}^T\mathbf{x} + w_0) < 0 \\ \infty & \text{otherwise} \end{cases}$$

- Implicitly, hard SVM used a function giving infinite error if a data point was misclassified

- **Goal:** Soften this approach



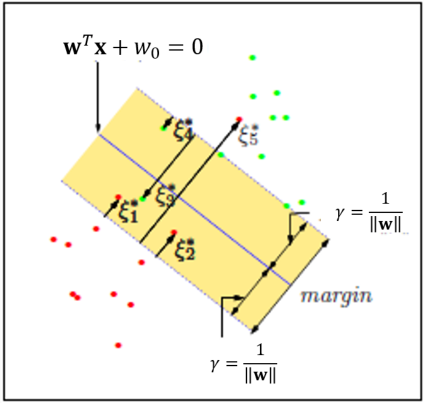
$\mathbf{w}^T\mathbf{x} + w_0 = 0$
$\gamma = \frac{1}{\|\mathbf{w}\|}$
margin
$\gamma = \frac{1}{\|\mathbf{w}\|}$

Adapted from Figure 12.1 From ESL.

Soft Margin SVM

Soft Margin SVM

- Soft margin SVM relaxes the constraint to allow points to be inside the margin or even on the wrong side of the boundary
- The boundaries are penalized by a quantity that reflects the extent of the violation
- We introduce slack variables $\xi_i \geq 0$ for each sample to measure the extent of the violation


$\mathbf{w}^T\mathbf{x} + w_0 = 0$
ξ_4^*
ξ_5^*
ξ_3^*
ξ_1^*
ξ_2^*
$\gamma = \frac{1}{\|\mathbf{w}\|}$
margin
$\gamma = \frac{1}{\|\mathbf{w}\|}$

Adapted from Figure 12.1 From ESL.

Slack Variables

- The original hard margin constraint:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall\ i$$

now becomes

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \quad \forall\ i$$

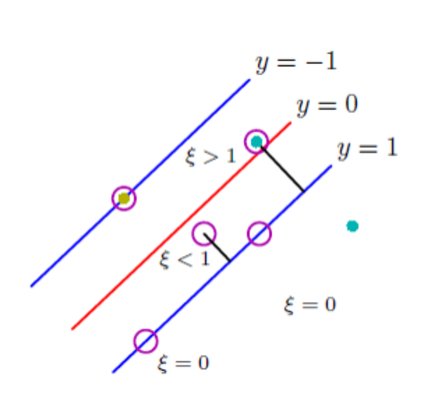


Figure 7.3 from PRML.

Slack Variables

- The original hard margin constraint:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall\ i$$

now becomes

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \quad \forall\ i$$

- For points on or in the correct margin side: $\xi_i = 0$

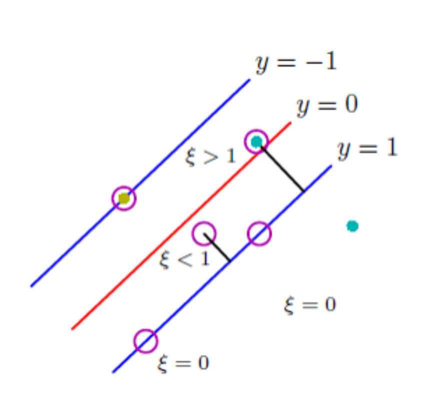


Figure 7.3 from PRML.

Slack Variables

- The original hard margin constraint:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall\ i$$

now becomes

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \quad \forall\ i$$

- For points on or in the correct margin side: $\xi_i = 0$
- For other points: $\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0)$

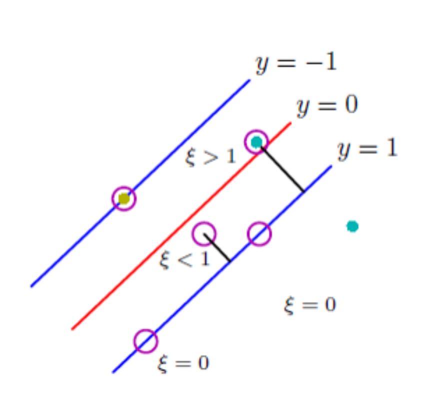


Figure 7.3 from PRML.

Slack Variables

- The original hard margin constraint:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall\ i$$

now becomes

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \quad \forall\ i$$

- For points on or in the correct margin side: $\xi_i = 0$
- For other points: $\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0)$
- If a point is in the decision boundary:

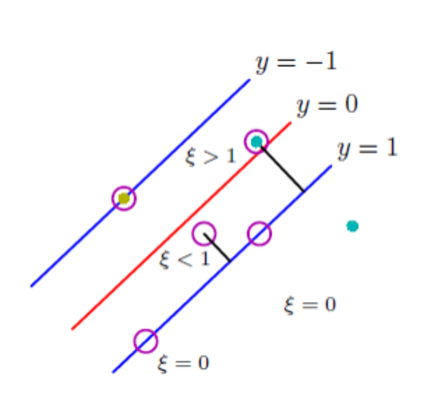


Figure 7.3 from PRML.

Slack Variables

- The original hard margin constraint:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall\ i$$

now becomes

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \quad \forall\ i$$

- For points on or in the correct margin side: $\xi_i = 0$
- For other points: $\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0)$
- If a point is in the decision boundary: $\xi_i = 1$

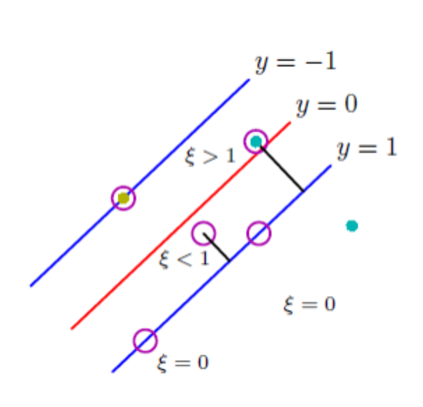


Figure 7.3 from PRML.

Slack Variables

- The original hard margin constraint:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall\ i$$

now becomes

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \quad \forall\ i$$

- For points on or in the correct margin side: $\xi_i = 0$
- For other points: $\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0)$
- If a point is in the decision boundary: $\xi_i = 1$

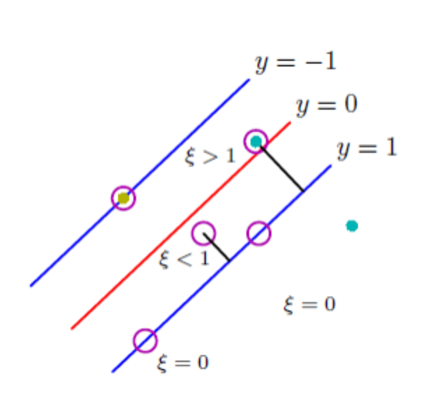


Figure 7.3 from PRML.

Slack Variables: Interpretation

Zone 1: $\xi_i > 1$

- Zone of points that are in the wrong side
- $\xi_i = 1 - (\text{negative value}) > 1$

Zone 2: $\xi_i < 1$

- Zone of points in the good side of the boundary decision but inside the margin
- $\xi_i = (1 - (\text{positive value} < 1)) < 1$

The slack variables ξ_i compensate to make sure the constraint is satisfied

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \quad \forall\, i$$

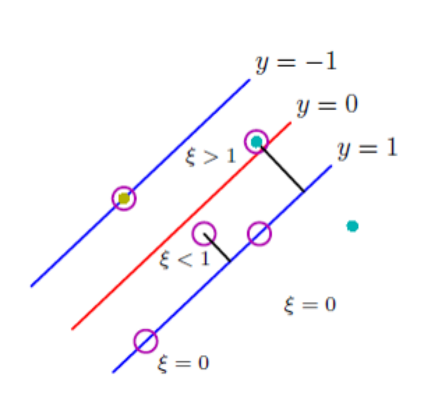
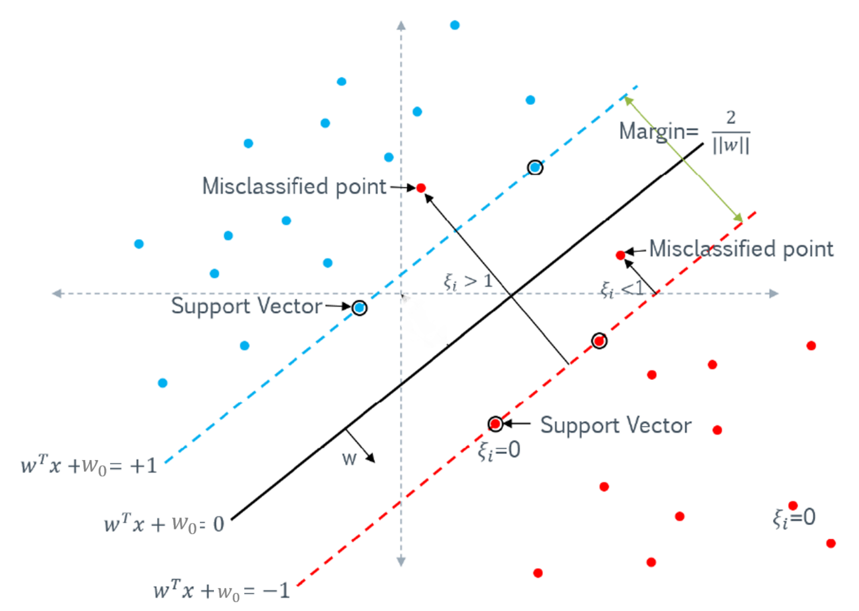


Figure 7.3 from PRML.


Margin= $\frac{2}{||w||}$
Misclassified point
Misclassified point
$\xi_i > 1$
$\xi_i < 1$
Support Vector
Support Vector
ξ_i=0
ξ_i=0
w
$w^T x + w_0 = +1$
$w^T x + w_0 = 0$
$w^T x + w_0 = -1$

Revisiting the Optimization Problem

- What we have done so far in words: If the constraint cannot be satisfied, we subtract some ξ_i on the right until it is satisfied

Revisiting the Optimization Problem

- What we have done so far in words: If the constraint cannot be satisfied, we subtract some ξ_i on the right until it is satisfied
- However, we want to subtract as little as possible
- Therefore, we want to minimize that quantity too

Revisiting the Optimization Problem

- What we have done so far in words: If the constraint cannot be satisfied, we subtract some ξ_i on the right until it is satisfied
- However, we want to subtract as little as possible
- Therefore, we want to minimize that quantity too
- The optimization problem now looks like

$$\begin{aligned} \underset{\mathbf{w}, w_0}{\arg\min} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & \forall i \; y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \\ & \forall i \; \xi_i \geq 0 \end{aligned}$$

- Blue: New terms
- $C > 0$ Hyper-parameter controlling the trade-off between slack variable penalty and the margin. $C \to \infty$: Hard margin SVM

Soft Margin Optimization Constraints

Soft Margin SVM:

$$\begin{aligned} \underset{\mathbf{w}, w_0}{\arg\min} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & \forall i \; y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \\ & \forall i \; \xi_i \geq 0 \end{aligned}$$

The Soft SVM remains a quadratic optimization problem that can be solved with standard solvers.

Soft Margin Optimization Constraints

Soft Margin SVM:

$$\begin{aligned} \underset{\mathbf{w}, w_0}{\arg\min} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & \forall i \; y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \\ & \forall i \; \xi_i \geq 0 \end{aligned}$$

The Soft SVM remains a quadratic optimization problem that can be solved with standard solvers.

Question: When would you use a Soft SVM? can you think of some examples?

Unconstrained Soft SVM Formulation

- From the analysis we did about ξ_i (slides 6-7), we can reformulate it as:

$$\xi_i = \begin{cases} 0 & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \\ 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0) & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) < 1 \end{cases}$$

Unconstrained Soft SVM Formulation

- From the analysis we did about ξ_i (slides 6-7), we can reformulate it as:

$$\xi_i = \begin{cases} 0 & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \\ 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0) & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) < 1 \end{cases}$$

- This is equivalent to:

$$\xi_i = \max(1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0), 0)$$

Unconstrained Soft SVM Formulation

- From the analysis we did about ξ_i (slides 6-7), we can reformulate it as:

$$\xi_i = \begin{cases} 0 & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \\ 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0) & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) < 1 \end{cases}$$

- This is equivalent to:

$$\xi_i = \max(1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0), 0)$$

- We can use this closed-form expression to reformulate the soft SVM constrained optimization

Unconstrained Soft SVM Formulation

Constrained Soft SVM Optimization Problem:

$$\begin{aligned} \underset{\mathbf{w}, w_0}{\arg\min} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & \forall i \; y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \\ & \forall i \; \xi_i \geq 0 \end{aligned}$$

Unconstrained Soft SVM Optimization Problem:

$$\underset{\mathbf{w}, w_0}{\arg\min} \; \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \max(1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0), 0)$$

About the Unconstrained Soft SVM Formulation

$$\underset{\mathbf{w}, w_0}{\arg\min} \quad \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{l2-\text{regularizer}} + C \sum_{i=1}^{N} \underbrace{\max(1 - y_i(\mathbf{w}^T \mathbf{x}_i + w_0), 0)}_{\text{hinge loss}}$$

This formulation allows to find the SVM parameters using gradient descent, as it has been done for logistic regression or the perceptron.

The key difference is the change of loss function. Here, we use the **hinge loss**.

The role of the regularizer will be covered later.

Wrap-up

Wrap-up

- We discussed the limitation of hard margin SVMs
- We introduced soft margin SVM as an alternative to relax the hard margin constraints
- We derived the objective of soft margin SVM both as a constrained and unconstrained optimization problem
- We presented the hinge loss function
- We introduced the concept of regularization

Key Concepts

- Soft Constraints
- Slack Variables
- Hinge Loss
- Soft Margin SVM

References

Further Reading and Useful Material

Source	Notes
Support Vector Networks - Cortes and Vapnik	original publication (link)
Pattern Recognition and Machine Learning	Ch 7
The Elements of Statistical Learning	Ch 12