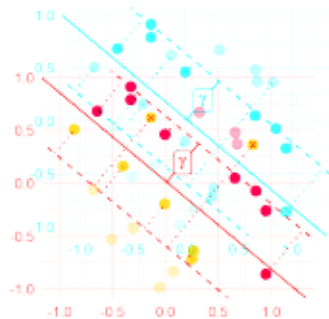


# Introduction to Machine Learning

## Linear Support Vector Machines

### Soft-Margin SVM

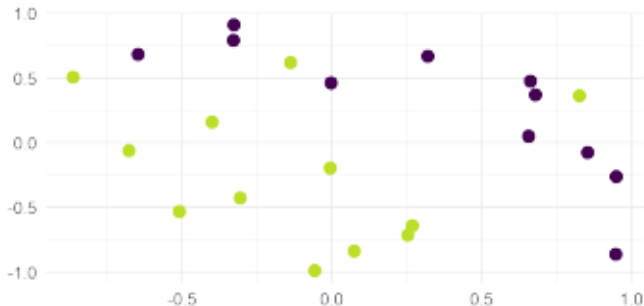


### Learning goals

#### Learning goals

- Understand that the hard-margin SVM problem is only solvable for linearly separable data
- Know that the soft-margin SVM problem therefore allows margin violations
- The degree to which margin violations are tolerated is controlled by a hyperparameter

# NON-SEPARABLE DATA



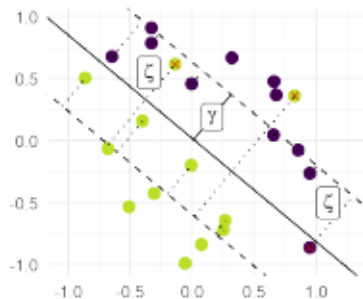
- Assume that dataset  $\mathcal{D}$  is not linearly separable.
- Margin maximization becomes meaningless because the hard-margin SVM optimization problem has contradictory constraints and thus an empty **feasible region**.

# MARGIN VIOLATIONS

- We still want a large margin for most of the examples.
- We allow violations of the margin constraints via slack vars  $\zeta^{(i)} \geq 0$

$$y^{(i)} \left( \langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \rangle + \theta_0 \right) \geq 1 - \zeta^{(i)}$$

- Even for separable data, a decision boundary with a few violations and a large average margin may be preferable to one without any violations and a small average margin.



We assume  $\gamma = 1$  to not further complicate presentation.



## SOFT-MARGIN SVM

The linear **soft-margin** SVM is the convex quadratic program:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \theta_0, \zeta^{(i)}} \quad & \frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \zeta^{(i)} \\ \text{s.t.} \quad & y^{(i)} \left( \langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \rangle + \theta_0 \right) \geq 1 - \zeta^{(i)} \quad \forall i \in \{1, \dots, n\}, \\ \text{and} \quad & \zeta^{(i)} \geq 0 \quad \forall i \in \{1, \dots, n\}. \end{aligned}$$



This is called “soft-margin” SVM because the “hard” margin constraint is replaced with a “softened” constraint that can be violated by an amount  $\zeta^{(i)}$ .

# LAGRANGE FUNCTION AND KKT

The Lagrange function of the soft-margin SVM is given by:

$$\mathcal{L}(\boldsymbol{\theta}, \theta_0, \zeta, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2} \|\boldsymbol{\theta}\|_2^2 + C \sum_{i=1}^n \zeta^{(i)} - \sum_{i=1}^n \alpha_i \left( y^{(i)} \left( \langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \rangle + \theta_0 \right) - 1 + \zeta^{(i)} \right) - \sum_{i=1}^n \mu_i \zeta^{(i)} \quad \text{with Lagrange multipliers } \boldsymbol{\alpha} \text{ and } \boldsymbol{\mu}.$$



The KKT conditions for  $i = 1, \dots, n$  are:

$$\begin{aligned} \alpha_i &\geq 0, & \mu_i &\geq 0, \\ y^{(i)} \left( \langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \rangle + \theta_0 \right) - 1 + \zeta^{(i)} &\geq 0, & \zeta^{(i)} &\geq 0, \\ \alpha_i \left( y^{(i)} \left( \langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \rangle + \theta_0 \right) - 1 + \zeta^{(i)} \right) &= 0, & \zeta^{(i)} \mu_i &= 0. \end{aligned}$$

With these, we derive (see our optimization course) that

$$\boldsymbol{\theta} = \sum_{i=1}^n \alpha_i y^{(i)} \mathbf{x}^{(i)}, \quad 0 = \sum_{i=1}^n \alpha_i y^{(i)}, \quad \alpha_i = C - \mu_i \quad \forall i = 1, \dots, n.$$

## SOFT-MARGIN SVM DUAL FORM

Can be derived exactly as for the hard margin case.

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} \langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \\ & \sum_{i=1}^n \alpha_i y^{(i)} = 0, \end{aligned}$$



or, in matrix notation:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \mathbf{1}^T \alpha - \frac{1}{2} \alpha^T \text{diag}(\mathbf{y}) \mathbf{K} \text{diag}(\mathbf{y}) \alpha \\ \text{s.t.} \quad & \alpha^T \mathbf{y} = 0, \\ & 0 \leq \alpha \leq C, \end{aligned}$$

with  $\mathbf{K} := \mathbf{X}\mathbf{X}^T$ .

## COST PARAMETER C

- The parameter  $C$  controls the trade-off between the two conflicting objectives of maximizing the size of the margin and minimizing the frequency and size of margin violations.
- It is known under different names, such as “trade-off parameter”, “regularization parameter”, and “complexity control parameter”.
- For sufficiently large  $C$  margin violations become extremely costly, and the optimal solution does not violate any margins if the data is separable. The hard-margin SVM is obtained as a special case.





# SUPPORT VECTORS

There are three types of training examples:

- Non-SVs have  $\alpha_i = 0$  ( $\Rightarrow \mu_i = C \Rightarrow \zeta^{(i)} = 0$ ) and can be removed from the problem without changing the solution. Their margin  $yf(\mathbf{x}) \geq 1$ . They are always classified correctly and are never inside of the margin.
- SVs with  $0 < \alpha_i < C$  ( $\Rightarrow \mu_i > 0 \Rightarrow \zeta^{(i)} = 0$ ) are located exactly on the margin and have  $yf(\mathbf{x}) = 1$ .
- SVs with  $\alpha_i = C$  have an associated slack  $\zeta^{(i)} \geq 0$ . They can be on the margin or can be margin violators with  $yf(\mathbf{x}) < 1$  (they can even be misclassified if  $\zeta^{(i)} \geq 1$ ).



As for hard-margin case: on the margin we can have SVs and non-SVs.

