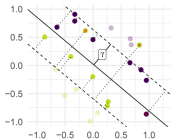


Introduction to Machine Learning

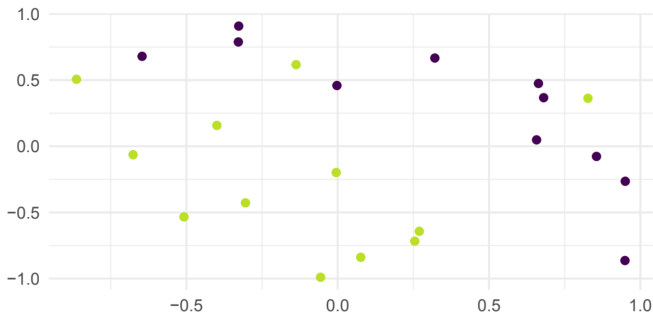
Soft-Margin SVM



Learning goals

- Understand that the hard-margin SVM problem is not 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



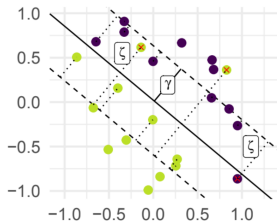
- Assume that dataset \mathcal{D} is not linearly separable.
- Margin maximization becomes meaningless because the hard-margin SVM optimization problem has contradictory constraints and thus an empty **feasible region**.

MARGIN VIOLATIONS

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

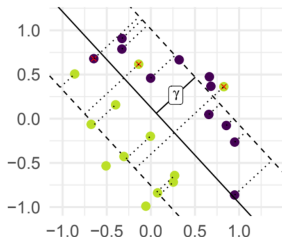
$$y^{(i)} \left(\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \rangle + \theta_0 \right) \geq 1 - \zeta^{(i)}$$

- 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 ...

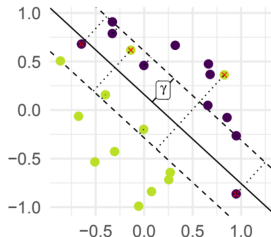


MARGIN VIOLATIONS

- Now we have two distinct and contradictory goals:
 - ➊ Maximize the margin.
 - ➋ Minimize margin violations.
- Let's minimize a weighted sum of them: $\frac{1}{2}\|\theta\|^2 + C \sum_{i=1}^n \zeta^{(i)}$
- Constant $C > 0$ controls the relative importance of the two parts.



$C = 0.5$



$C = 100$

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 $\zeta^{(i)}$.

SOFT-MARGIN SVM DUAL FORM

Can be derived exactly as for the hard margin case.

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

or, in matrix notation:

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

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 a margin > 1 and can be removed from the problem without changing the solution.
- Some SVs are located exactly on the margin and have $yf(\mathbf{x}) = 1$.
- Other SVs are margin violators, with $yf(\mathbf{x}) < 1$, and have an associated positive slack $\zeta^{(i)} > 0$. They are misclassified if $\zeta^{(i)} \geq 1$.

