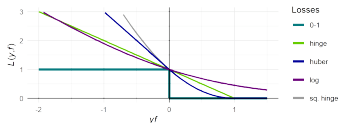


# Introduction to Machine Learning

## Linear Support Vector Machines

## SVMs and Empirical Risk Minimization

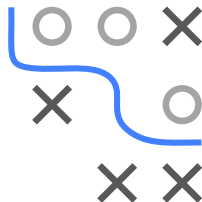


### Learning goals

- Know why the SVM problem can be understood as (regularized) empirical risk minimization problem
- Know that the corresponding loss is the hinge loss

# REGULARIZED EMPIRICAL RISK MINIMIZATION

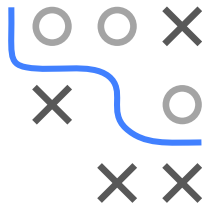
- We motivated SVMs from a geometrical point of view: The margin is a distance to be maximized.
- This is not really true anymore under margin violations: The slack variables are not really distances. Instead,  $\gamma \cdot \zeta^{(i)}$  is the distance by which an observation violates the margin.
- This already indicates that transferring the geometric intuition from hard-margin SVMs to the soft-margin case has its limits.
- There is an alternative approach to understanding soft-margin SVMs: They are **regularized empirical risk minimizers**.



# SOFT-MARGIN SVM WITH ERM AND HINGE LOSS

We derived this QP for the soft-margin SVM:

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



In the optimum, the inequalities will hold with equality (as we minimize the slacks), so  $\zeta^{(i)} = 1 - y^{(i)} f(\mathbf{x}^{(i)})$ , but the lowest value  $\zeta^{(i)}$  can take is 0 (we do not get a bonus for points beyond the margin on the correct side). So we can rewrite the above:

$$\frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n L(y^{(i)}, f(\mathbf{x}^{(i)})); \quad L(y, f(\mathbf{x})) = \begin{cases} 1 - yf & \text{if } yf \leq 1 \\ 0 & \text{if } yf > 1 \end{cases}$$

We can also write  $L(y, f(\mathbf{x})) = \max(1 - yf, 0)$ .

# OTHER LOSSES

SVMs can easily be generalized by changing the loss function.

- Squared hinge loss / Least Squares SVM:

$$L(y, f(\mathbf{x})) = \max(0, (1 - yf)^2)$$

- Huber loss (smoothed hinge loss)
- Bernoulli/Log loss. This is L2-regularized logistic regression!
- NB: These other losses usually do not generate sparse solutions in terms of data weights and hence have no "support vectors".

