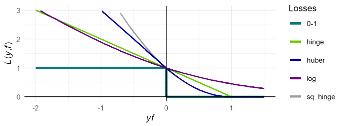
Introduction to Machine Learning

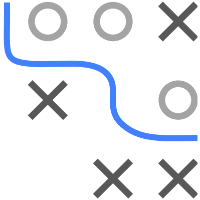
Linear Support Vector Machines
SVMs and Empirical Risk Minimization



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:

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

In the optimum, the inequalities will hold with equality (as we minimize the slacks), so $\zeta^{(i)} = 1 - y^{(i)} f\left(\mathbf{x}^{(i)}\right)$, but the lowest value $\zeta^{(i)}$ can take is 0 (we do no get a bonus for points beyond the margin on the correct side). So we can rewrite the above:

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

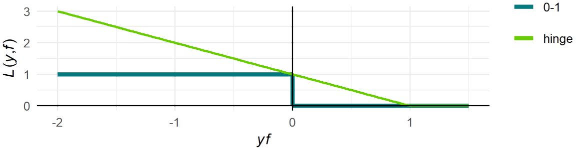
We can also write $L(y, f) = \max(1 - yf, 0)$.

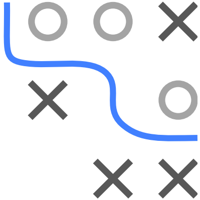
SOFT-MARGIN SVM WITH ERM AND HINGE LOSS / 2

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \; ; \; L\left(y, f\right) = \max(1 - yf, 0)$$

- This now obviously L2-regularized empirical risk minimization.
- Actually, a lot of ERM theory was established when Vapnik (co-)invented the SVM in the beginning of the 90s.
- L is called hinge loss – as it looks like a door hinge.
- It is a continuous, convex, upper bound on the zero-one loss. In a certain sense it is the best upper convex relaxation of the 0-1.

SOFT-MARGIN SVM WITH ERM AND HINGE LOSS / 3



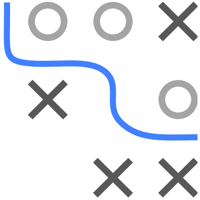


SOFT-MARGIN SVM WITH ERM AND HINGE LOSS / 4

$$\frac{1}{2}\|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right); \; L(y, f) = \max(1 - yf, 0)$$

- The ERM interpretation does not require any of the terms – the loss or the regularizer – to be geometrically meaningful.
- The above form is a very compact form to define the convex optimization problem of the SVM.
- It is "well-behaved" due to convexity, every minimum is global.
- The above is convex, without constraints! We might see this as "easier to optimize" than the QP from before. But note it is non-differentiable due to the hinge. So specialized techniques (e.g. sub-gradient) would have to be used.
- Some literature claims this primal cannot be easily kernelized - which is not really true.

OTHER LOSSES



SVMs can easily be generalized by changing the loss function.

- Squared hinge loss / Least Squares SVM:
  $L(y, f) = \max(0, (1 - yf)^2)$
- 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".