- Using refined techniques we can improve the bound for the hard SVM rule for particular μ:

Theorem 4.5:

Let μ be a distribution on $\mathbb{R}^d \times \{-1, +1\}$ with the so-called (γ, ρ)-separability property, i.e., there exists $(\mathbf{w}^*, b^*) \in \mathbb{R}^{d+1}$ with $\|\mathbf{w}^*\| = 1$ and such that for $(\mathbf{X}, Y) \sim \mu$ almost surely

$$Y\left(\mathbf{w}^* \cdot \mathbf{X} + b^*\right) \geq \gamma > 0 \qquad \text{and} \qquad \|\mathbf{X}\| \leq \rho < \infty.$$

Then we have with probability at least $1 - \delta$ that

$$\mathcal{R}_\mu(\mathrm{SVM}_{\mathsf{hard}}(S)) \leq \frac{1}{\sqrt{m}}\left(\frac{2\rho}{\gamma} + \sqrt{2\ln\left(\frac{2}{\delta}\right)}\right).$$

Proof: See Chapter 26 in "Understanding Machine Learning" (2014)

- This yields, the error of the (hard) SVM rule is dimension independent for such distributions μ.

4.2 Soft SVM

- Let us now extend the procedure of the hard SVM rule to the case of arbitrary, in particular non-linearly separable samples s.
- I.e., we can no longer assume that there is $(\mathbf{w}, b) \in \mathbb{R}^{d+1}$ with

$$y_i\left(\mathbf{w} \cdot \mathbf{x}_i + b\right) > 0 \qquad \forall i = 1, \ldots, m$$

respectively, in terms of the constraint of the hard SVM rule,

$$y_i\left(\mathbf{w} \cdot \mathbf{x}_i + b\right) \geq 1 \qquad \forall i = 1, \ldots, m.$$

Slack Variables are used in the Soft Margin SVM to handle cases where data points are not perfectly separable.

- We therefore introduce non-negative slack variables $\xi_i \geq 0$, $i = 1, \ldots, m$, and replace the above constraint with

$$y_i\left(\mathbf{w} \cdot \mathbf{x}_i + b\right) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, m.$$

- Thus, for $\xi_i > 1$, $\mathbf{x}_i$ may be on the "wrong side" of the hyperplane.

- The slack variables ξ_i quantify the violation of the constraints of the hard SVM rule.

- So we are looking for a vector $(\mathbf{w}, b)$ with small norm $\|\mathbf{w}\|^2$ and small violation of the original hard SVM constraints.

- To balance both goals we choose a control parameter $\lambda > 0$ and consider

Soft SVM rule

number of slack vars coresponds to traing data

Given:

- Sample s with m data pairs $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, +1\}$
- Parameter $\lambda > 0$

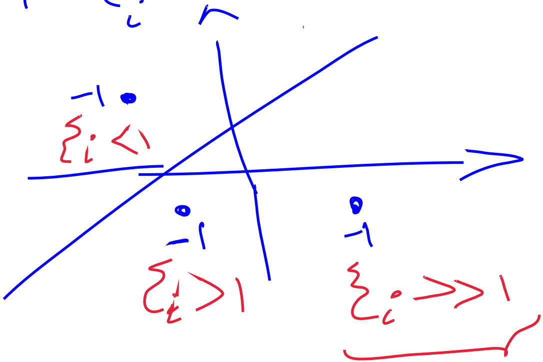
Compute: $h_{\mathbf{w}_s, b_s} = \mathrm{SVM}_{\mathsf{soft}}(s; \lambda) \in \mathcal{L}_d$ given by

$$(\mathbf{w}_s, b_s, \boldsymbol{\xi}_s) \in \underset{(\mathbf{w}, b, \boldsymbol{\xi}) \in \mathbb{R}^{d+1+m}}{\operatorname{argmin}} \lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i$$

subject to: $\quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0 \quad \forall i = 1, \ldots, m.$

$$y_i(w \cdot x + b) \geq 1 - \xi_i$$

$\Rightarrow$



too far from hyperplane

$$\Rightarrow \underbrace{\lambda \sum_{i=1}^{m} \xi_i + \|w\|^2}_{\min}$$

Notes

- The optimal weight vector $\mathbf{w}_s$ in the soft SVM rule is unique. But there can be an interval of optimal bias values b_s – unlike for the hard SVM rule.

- Furthermore, due to the convexity of the optimization task, any minimum found is a global minimum – both SVM rules thus have no local minima, unlike (deep) neural networks.

- Likewise, the weight vector $\mathbf{w}_s$ learned by the soft SVM rule is again in the span of certain support vectors $\mathbf{x}_i$:

$$\mathbf{w}_s = \sum_{i:\ g_i(\mathbf{w}_S, b_S, \boldsymbol{\xi}_S)=0} \alpha_i \mathbf{x}_i.$$

- For $\lambda \to 0$, the violation of the constraints is increasingly penalized. If s is linearly separable, then for sufficiently small $\lambda \ll 1$ we have

$$\mathrm{SVM}_{\mathsf{soft}}(s;\lambda) \approx \mathrm{SVM}_{\mathsf{hard}}(s)$$

The Hinge loss

The soft SVM rule is a regularized ERM rule based on a new loss function:

$\xi_i \geq 1 - y(wx + b)$

(sub differentiable)

The hinge loss

For $\mathbf{w}' = (\mathbf{w}, b) \in \mathbb{R}^{d+1}$ and $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{-1, +1\}$ let

$$\ell_{\text{hinge}}(\mathbf{w}', (\mathbf{x}, y)) := \max\{0, 1 - y(\mathbf{w}' \cdot \mathbf{x}')\}, \qquad \mathbf{x}' := (\mathbf{x}, 1).$$

The smallest slack vars, the better because of violating original hard svm constraint, so it has to be maximized

Proposition 4.6:

The soft SVM learning rule is equivalent to

$$(\mathbf{w}_s, b_s) \in \operatorname*{argmin}_{(\mathbf{w}, b) \in \mathbb{R}^{d+1}} \mathcal{R}_s^{\text{hinge}}((\mathbf{w}, b)) + \lambda \|\mathbf{w}\|^2$$

with the empirical risk $\mathcal{R}_s^{\text{hinge}}((\mathbf{w}, b)) := \frac{1}{m} \sum_{i=1}^m \ell_{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}_i, y_i))$.

Notes

- The proposition follows from the fact that in a minimum of the soft SVM rule we have for the slip variables $\xi_1 \geq 0$, i.e.,

$$\xi_i = \max\{0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i + b)\} = \ell_{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}_i, y_i)) \qquad \forall i = 1, \ldots, m.$$

penalities in logistic regression

- Learning rules of the type

$$\mathcal{R}_s(h) + \lambda R(h) \to \min_h$$

are called regularized ERM rules with regularization parameter λ and regularization or penalty functional R.

L1, L2

- The hinge loss is again convex with respect to the parameters $\mathbf{w}, b$ and thus also the objective function to be minimized

$$f(\mathbf{w}, b) = \mathcal{R}_s((\mathbf{w}, b)) + \lambda \|\mathbf{w}\|^2.$$ L2, L1 = $\|w\|$

However, ℓ_{hinge} and hence f is no longer differentiable with respect to $\mathbf{w}, b$.

argmin $\frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i(wx_i + b_i)\}$

Comparison of loss functions

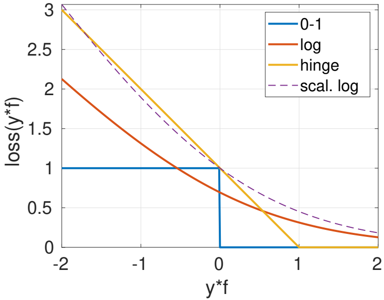
- The 0-1 loss, log loss, and hinge loss can be defined as follows using the affine mapping $f_{\mathbf{w},b}(\mathbf{x}) := \mathbf{w} \cdot \mathbf{x} + b$:

$$\ell_{\text{0-1}}(f_{\mathbf{w},b}, \mathbf{x}, y) := \mathbb{1}_{(-\infty,0)}\left(y f_{\mathbf{w},b}(\mathbf{x})\right),$$
$$\ell_{\text{log}}(f_{\mathbf{w},b}, \mathbf{x}, y) := \ln\left(1 + \exp(-y f_{\mathbf{w},b}(\mathbf{x}))\right),$$
$$\ell_{\text{hinge}}(f_{\mathbf{w},b}, \mathbf{x}, y) := \max\left\{0, 1 - y f_{\mathbf{w},b}(\mathbf{x})\right\}$$

$\Rightarrow$ All three losses are real functions of the scalar value $y f_{\mathbf{w},b}(\mathbf{x})$.

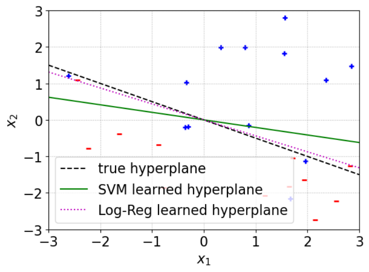
- The hinge loss is always greater or equal to the 0-1 loss.
- To achieve the same for the log loss you can scale it:

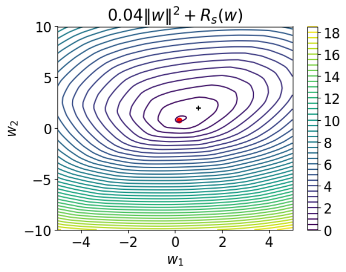
$$\frac{\ln\left(1 + \exp(-y f_{\mathbf{w},b}(\mathbf{x}))\right)}{\ln 2}.$$



Example: Synthetic dataset

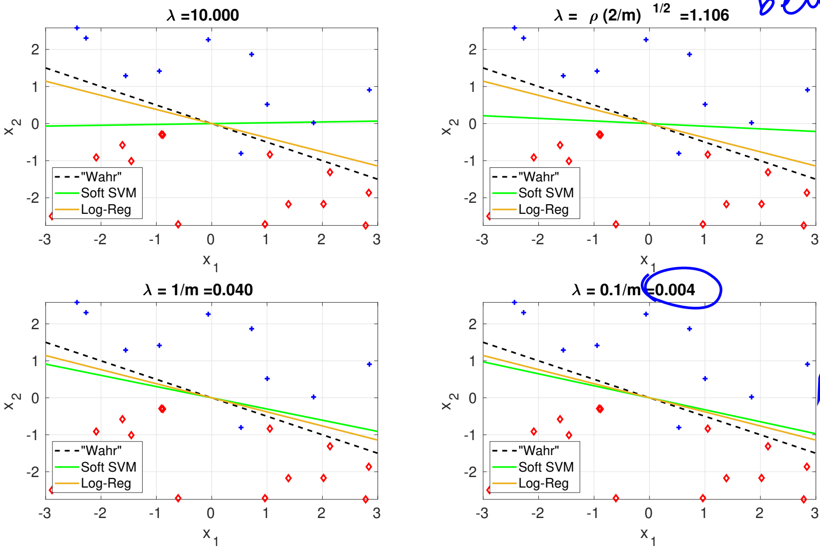
- For $\mathcal{X} = \mathbb{R}^2$ we want to learn $h_{\mathbf{w},0} \in \mathcal{L}_d$ by the soft SVM rule.
- $m = 25$ training data is randomly generated with random labels corresponding to a Bernoulli distribution as in Section 3.2 with $\mathbf{w}^\dagger = (1,2)^\top$, $b^\dagger = 0$.
- For the soft SVM rule, we choose $\lambda = \frac{1}{m}$ and obtain $\mathbf{w}_s \approx (0.17, 0.81)^\top$.
- Logistic regression yields $\mathbf{w}_s^{\mathsf{LR}} \approx (0.70, 1.60)^\top$.
- The example can be reproduced by a provided Jupyter notebook





Choice of λ

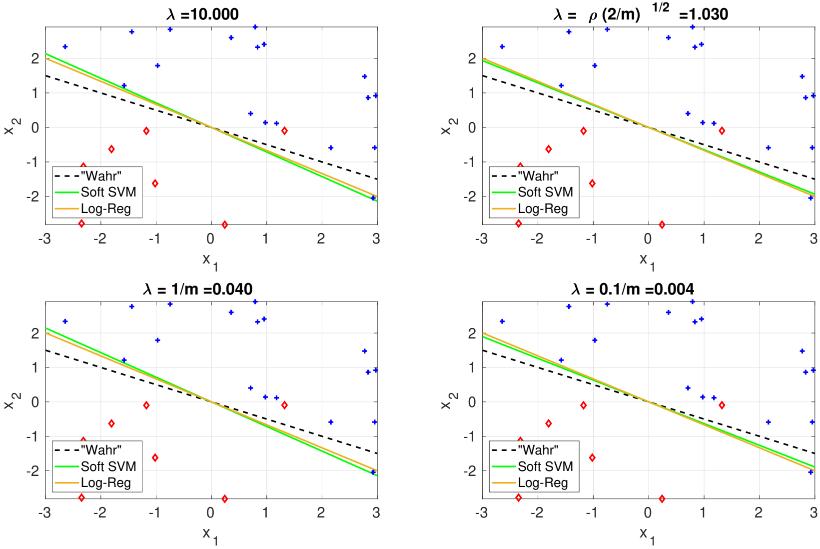
The parameter λ can have a significant effect on the learning outcome:



The less the λ the better not to getting results for the correct hypothesis & traing data

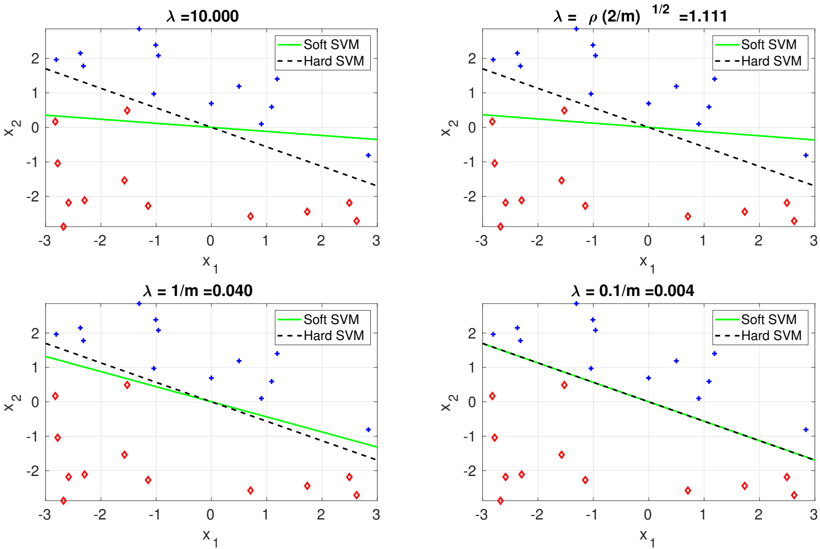
But it depends on traing data

Its effect depends in general on the data:



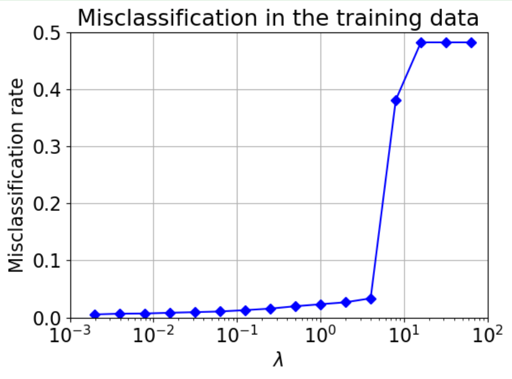
Comparison of hard and soft SVM

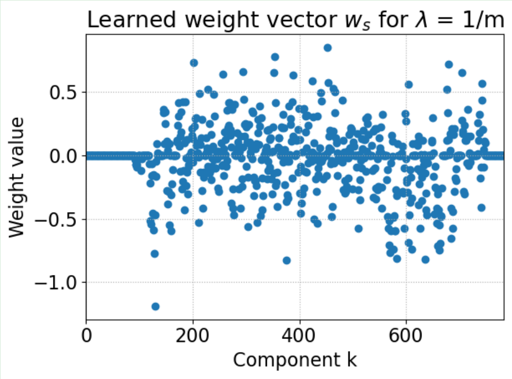
For linearly separable data we recover for small λ the result of the hard SVM rule:



Example: MNIST dataset

- We apply the soft SVM learning rule to the MNIST dataset to learn to distinguish handwritten sevens and eights.
- We thereby consider the influence of λ on the misclassification obtained.





- Small λ leads to an improved fit. For $\lambda = 1/m \approx 8 \cdot 10^{-5}$ we get 0.21% misclassifications.

Advantages of the soft SVM rule

- We cannot apply the quantitative fundamental theorem from Chapter 2 to obtain bounds $C(m,\delta) < \infty$ with

$$\mathbb{P}_{\mu^m}\left(\mathcal{R}_\mu^{\text{hinge}}(h_{\mathbf{w}_S,b_S}) \leq C(m,\delta)\right) \geq 1-\delta$$

  for the output $(\mathbf{w}_S, b_S)$ of the soft SVM rule, because:
  - the soft SVM rule yields for $\lambda > 0$ no ERM hypothesis,
  - the hinge loss is unbounded.

- However, for regularized ERM rules

$$h_s = \mathrm{ERM}_{\mathcal{H},R}(s;\lambda) \in \operatorname*{argmin}_{h\in\mathcal{H}} \mathcal{R}_s(h) + \lambda R(h)$$

  one can show bounds for the mean generalization error

$$\mathbb{E}_{\mu^m}[\mathcal{R}_\mu(h_S)] \leq C(m).$$

Theorem 4.7:

Let μ be a distribution on $\mathbb{R}^d \times \{-1,+1\}$ such that for $(\mathbf{X}, Y) \sim \mu$ we have almost surely $\|\mathbf{X}\| \leq \rho < \infty$. Then for

$$\mathbf{w}_s := \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|^2 + \mathcal{R}_s^{\text{hinge}}(f_{\mathbf{w},0}),$$

we have

$$\mathbb{E}_{\mu^m}\left[\mathcal{R}_\mu^{\text{hinge}}(f_{\mathbf{w}_S,0})\right] \leq \min_{\mathbf{v}\in\mathbb{R}^d}\left(\mathcal{R}_\mu^{\text{hinge}}(f_{\mathbf{v},0}) + \lambda\|\mathbf{v}\|^2\right) + \frac{2\rho^2}{\lambda m}.$$

- The term $\frac{2\rho^2}{\lambda m}$ bounds the (mean) estimation error $\mathbb{E}_{\mu^m}[\varepsilon_{\text{est}}(S)]$ and the green highlighted text the approximation error.
- Again, the bound for the generalization error does not depend on the feature dimension $d = \mathrm{VCD}(\mathcal{L}_d^0)$. This has some advantages in practice, e.g., in text classification where $d \gg 10^4$ but $\|\mathbf{x}\| \leq 1 = \rho$.