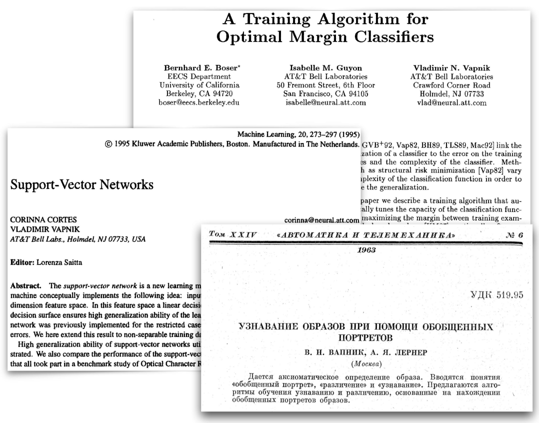
Support Vector Machines

Machine Learning Course - CS-433
Oct 24, 2023
Nicolas Flammarion

EPFL

Vapnik's invention



Binary classification

We observe some data $S = \{x_n, y_n\}_{n=1}^N \in \mathscr{X} \times \{-1, 1\}$

Goal: given a new observation x, we want to predict its label y

How:

$S = \{x_n, \underset{\{-1,1\}}{\underset{\Cap}{y_n}}\}_{n=1}^N$ $\qquad \mathscr{A} \qquad f_S = \mathscr{A}(S)$

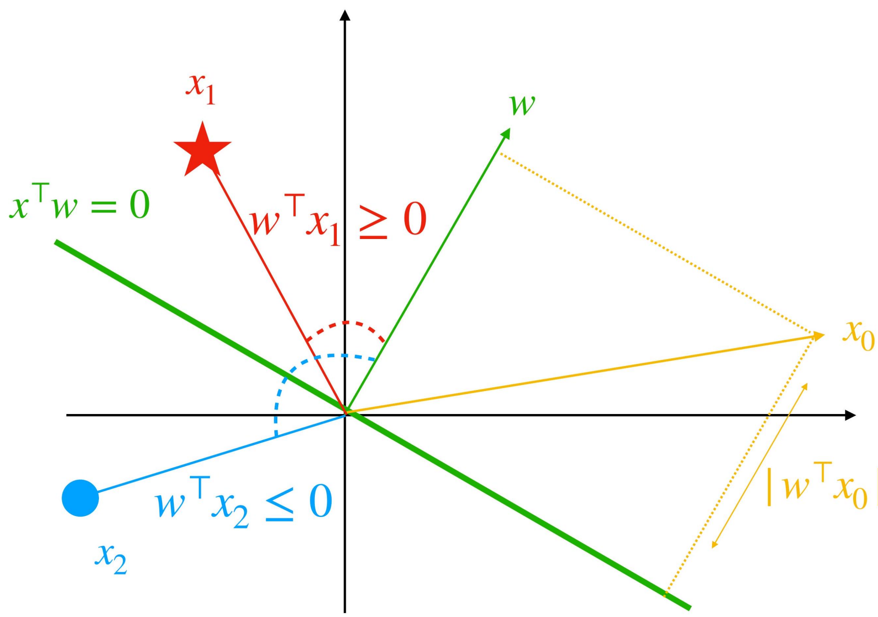
Linear Classifier

Define a hyperplane as $\{x : w^\top x = 0\}$ where $\|w\| = 1$

Prediction:

$$f(x) = \text{sign}\left(x^\top w\right)$$

Claim: The distance between a point x_0 and the hyperplane defined by w is $|w^\top x_0|$



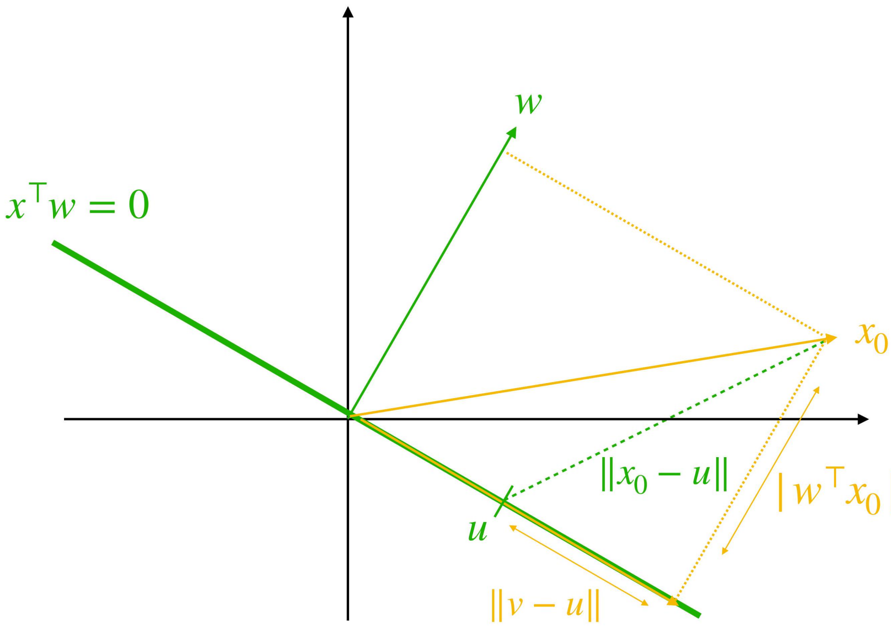
Linear Classifier

Proof: The distance between x_0 and the hyperplane is given by $\min_{u:w^\top u=0} \|x_0 - u\|$

Let $v = x_0 - w^\top x_0 w$ then by the Pythagorean theorem for any u s.t. $w^\top u = 0$

$\|x_0 - u\|^2 = \left(w^\top x_0\right)^2 + \|v - u\|^2 \geq \left(w^\top x_0\right)^2$

Claim: The distance between a point x_0 and the hyperplane defined by w is $|w^\top x_0|$



Hard-SVM rule: max-margin separating hyperplane

First assume the dataset $(x_n, y_n)_{n=1}^N$ is linearly separable

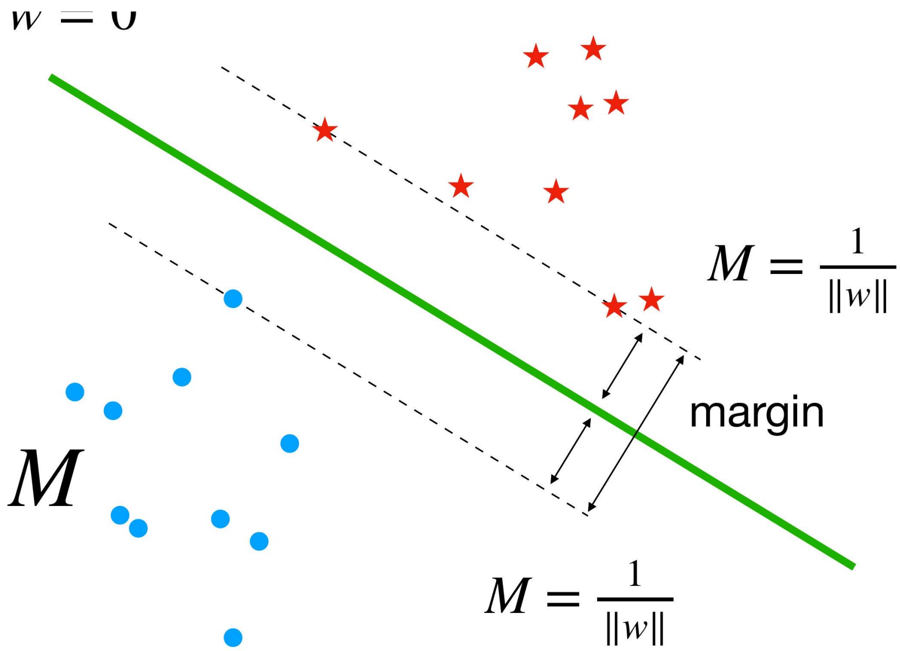
Margin of a hyperplane: $\min_{n \le N} \left| w^\top x_n \right|$

$$n \le N \quad x^\top w = 0$$

Max-margin separating hyperplane:

$$\max_{w, \|w\|=1} \min_{n \le N} \left| w^\top x_n \right| \text{ such that } \forall n, y_n x_n^\top w \ge 0$$

Equivalent to $\max_{M \in \mathbb{R}, w, \|w\|=1} M$ such that $\forall n, y_n x_n^\top w \ge M$
also equivalent to:



$$\min_w \frac{1}{2}\|w\|^2 \text{ such that } \forall n, y_n x_n^\top w \geq 1$$

Proof of the equivalent formulations

Claim: The following optimization problems are equivalent $\max\min \left|w^\top x_n\right|$ $w, \|w\| = 1 \quad n \leq N$ s.t. $\forall n, y_n x_n^\top w \geq 0$ $\max_{M\in\mathbb{R}, w, \|w\|=1} M$ s.t. $\forall n, y_n x_n^\top w \geq M$

Proof: Let w_1 be a solution of (I) and $M_1 = \min_{n\leq N} \left|w_1^\top x_n\right|$ and let w_2 and M_2 be solutions of (II)

- (w_1, M_1) is admissible for (II) so $M_1 \leq M_2$
- w_2 is admissible for (I) so $\min_{n\leq N} \left|w_2^\top x_n\right| \leq \min_{n\leq N} \left|w_1^\top x_n\right|$
- $\forall n, y_n x_n^\top w_2 \geq M_2$ implies that $\forall n, \left|x_n^\top w_2\right| \geq M_2$ and $\min_{n\leq N} \left|x_n^\top w_2\right| \geq M_2$

Therefore $M_1 = \min_{n\leq N} \left|w_1^\top x_n\right| \geq \min_{n\leq N} \left|w_2^\top x_n\right| \geq M_2 \geq M_1$
And the two problems are equivalent

Proof of the equivalent formulations

Claim: The following optimization problems are equivalent

$$
\begin{aligned}
&\max_{M\in\mathbb{R},w,\|w\|=1} M \\
&\text{s.t. } \forall n, y_n x_n^\top w \geq M \\
&\min_w \frac{1}{2}\|w\|^2 \\
&\text{s.t. } \forall n, y_n x_n^\top w \geq 1
\end{aligned}
$$

Proof:

$$
\begin{aligned}
&\max_{M\in\mathbb{R},w,\|w\|=1} M \text{ such that } \forall n, y_n x_n^\top w \geq M \\
\iff &\max_{M\in\mathbb{R},w} M \text{ such that } \forall n, y_n x_n^\top \frac{w}{\|w\|} \geq M
\end{aligned}
$$

The constraints are independent of the scale of w. Set $\|w\| = 1/M$:
$\iff \max 1/\|w\|$ such that $\forall n, y_n x_n^\top w \geq 1$
$\iff \min_w^w \frac{1}{2}\|w\|^2$ such that $\forall n, y_n x_n^\top w \geq 1$

Soft SVM: a relaxation of the Hard-SVM rule that can be applied even if the training set is not linearly separable

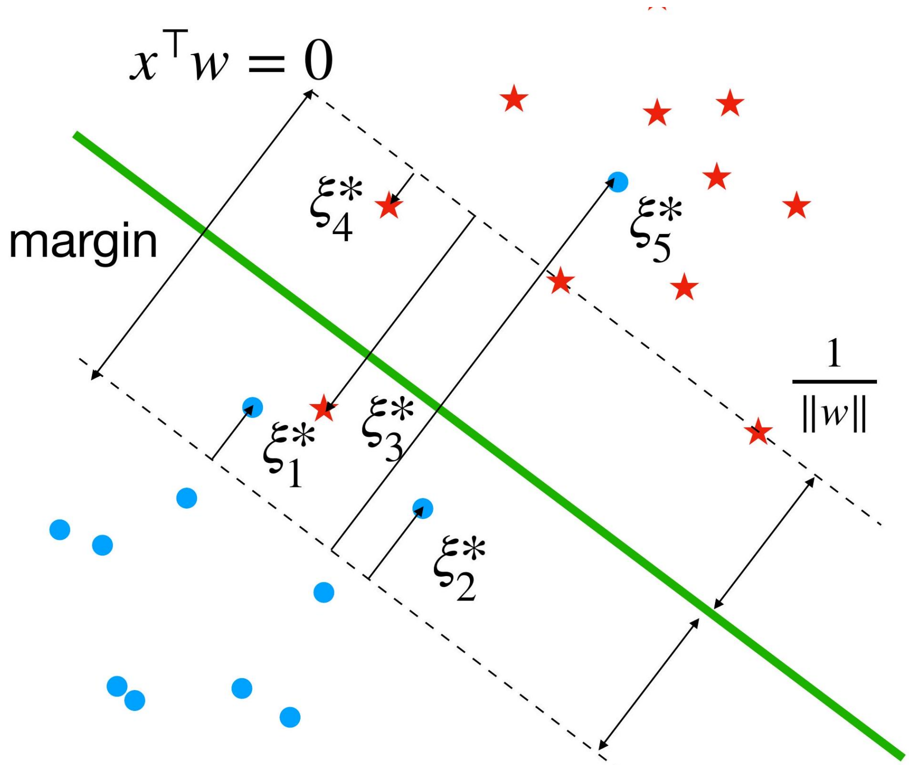
Idea: Maximize the margin while allowing some constraints to be violated

How: Introduce positive slack variables $\xi_1, \cdots, \xi_N$ and replace the constraints with $y_n x_n^\top w \geq 1 - \xi_n$ Soft SVM:

$$
\begin{aligned}
&\min_{w,\xi} \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n \\
&\text{s.t. } \forall n, y_n x_n^\top w \geq 1 - \xi_n \text{ and } \xi_n \geq 0
\end{aligned}
$$

which is equivalent to

$$
\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N} \left[1 - y_n x_n^\top w\right]_+
$$



$$\xi_i^* = \frac{\xi_i}{\|w\|}$$

-

Soft SVM: a relaxation of the Hard-SVM rule that can be

andied even if the trainina set is not linearly separableProof: Fix w and consider the minimization over ξ :

- If $y_n x_n^\top w \geq 1$, then $\xi_n = 0$
- If $y_n x_n^\top w < 1, \xi_n = 1 - y_n x_n^\top w$

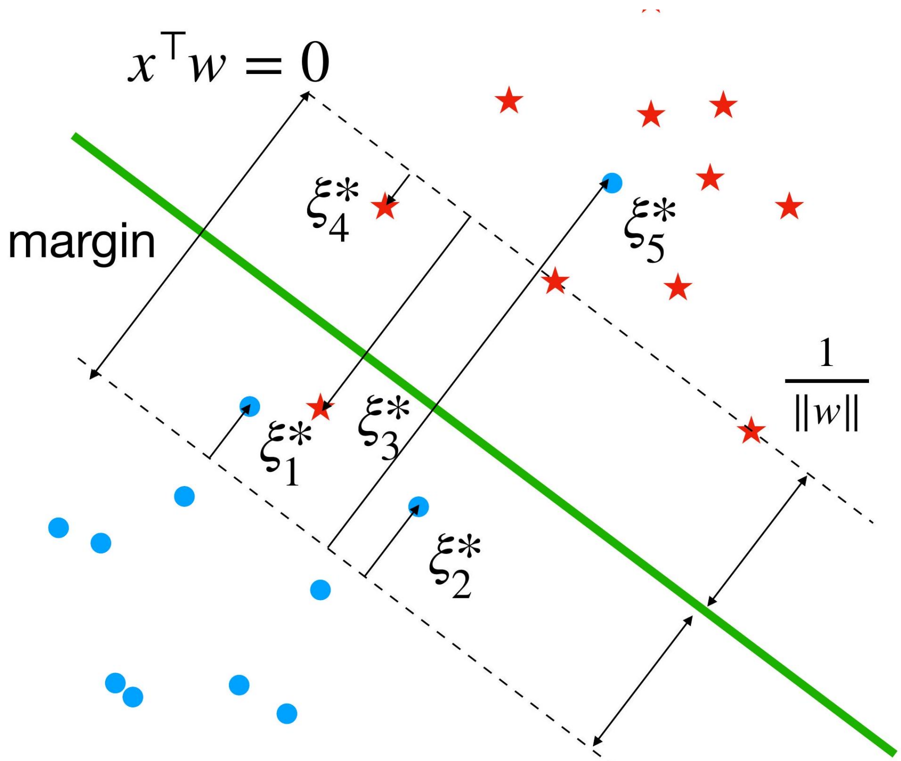
Therefore $\xi_n = \left[1 - y_n x_n^\top w\right]_+$

$$\min_{w,\xi} \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n$$
$$\text{s.t. } \forall n, y_n x_n^\top w \geq 1 - \xi_n \text{ and } \xi_n \geq 0$$

which is equivalent to

$$\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N}\left[1 - y_n x_n^\top w\right]_+$$

raints to



$$\xi_i^* = \frac{\xi_i}{\|w\|}$$

Classification by risk minimization

Setting: $(X, Y) \sim \mathscr{D}$ with ranges $\mathscr{X}$ and $\mathscr{Y} = \{-1, 1\}$
Goal: Find a classifier $f : \mathscr{X} \to \mathcal{Y}$ that minimizes the true risk

$$L(f) = \mathbb{E}_{\mathscr{D}}\left(1_{Y \neq f(X)}\right)$$

How: Through Empirical Risk Minimization (ERM):

$$\min_w L_{\text{train}}(w) = \frac{1}{N}\sum_{n=1}^{N}\phi\left(y_n w^\top x_n\right)$$

ϕ represents the loss function of the functional margin $y_n x_n^\top w$
ϕ also serves as a convex surrogate for the $0-1$ loss

Losses for Classification

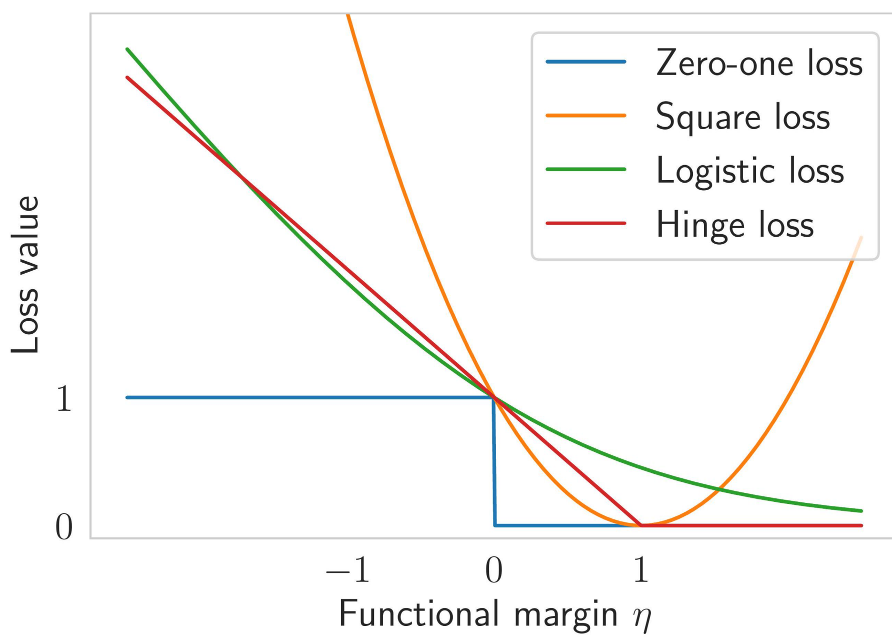
Examples of margin-based losses $\left(\eta = yx^\top w\right)$:

- Quadratic loss: $\mathrm{MSE}(\eta) = (1-\eta)^2$
- Logistic loss: $\mathrm{Logistic}(\eta) = \frac{\log(1+\exp(-\eta))}{\log(2)}$
- Hinge loss: $\mathrm{Hinge}(\eta) = [1-\eta]_+$

Common features: these losses are convex and provide an upper bound for the zero-one loss

Behavioral differences:

- MSE: Penalizes any deviation from 1
- Logistic Loss: Asymmetric cost - a penalty is always incurred.

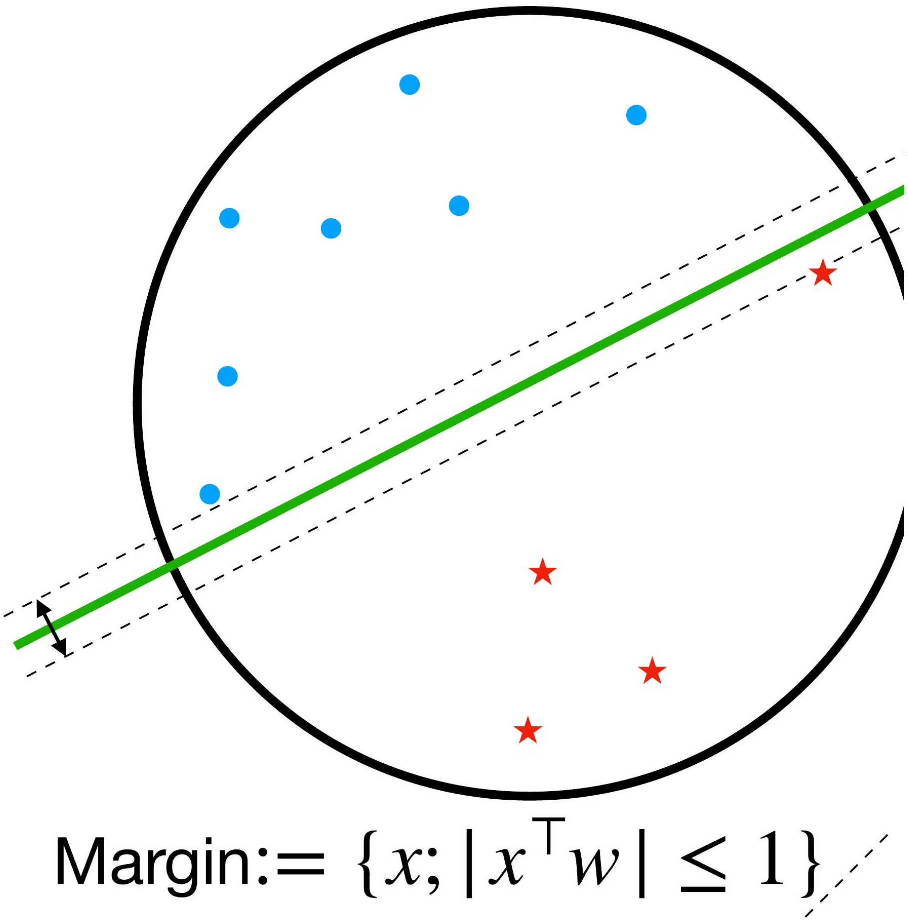


- Hinge Loss: A penalty is applied if the prediction is incorrect or lacks confidence

Summary

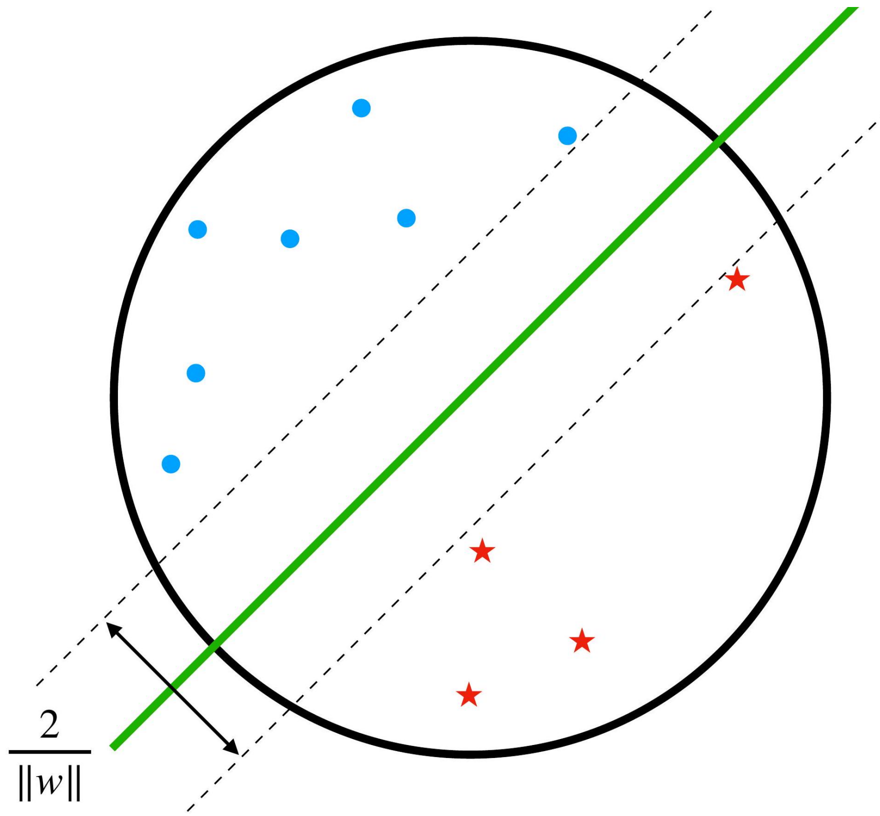
$$\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N}\left[1 - y_n x_n^\top w\right]_+$$

ERM for the hinge loss with ridge regularization



Interpretation for separable data with small λ :

1. Choose the direction of w such that $w^\perp$ acts as a separating hyperplane
2. Adjust the scale of w to ensure that no point lies with the margin
3. Select the hyperplane with the largest margin



Optimization: How to get w ?

$$\min_w \frac{1}{N} \sum_{n=1}^{N} \left[1 - y_n x_n^\top w\right]_+ + \frac{\lambda}{2} \|w\|^2$$

Convex (but non-smooth) objective which can be minimized with:

- Subgradient method
- Stochastic Subgradient method

Convex duality

Assume you can define an auxiliary function $G(w, \alpha)$ such that

$$\min_w L(w) = \min_w \max_\alpha G(w, \alpha)$$

Primal problem: $\min \max G(w, \alpha)$
w $\quad \alpha$

Dual problem: $\max \min G(w, \alpha)$
$\alpha \quad w$
$\Rightarrow$ Sometimes, the dual problem is easier to solve than the primal problem.
Questions:

1. How do we identify a suitable $G(w, \alpha)$?
2. Under what conditions can the min and max be interchanged?
3. When is the dual problem more tractable than the primal problem?

Q1: How do we find a suitable $G(w, \alpha)$?

$$[z]_+ = \max(0, z) = \max_{\alpha \in [0,1]} \alpha z$$

Therefore $\left[1 - y_n x_n^\top w\right]_+ = \max_{\alpha_n \in [0,1]} \alpha_n \left(1 - y_n x_n^\top w\right)$
The SVM problem is equivalent to:

$$\min_w L(w) = \min_w \max_{\alpha \in [0,1]^n} \underbrace{\frac{1}{N} \sum_{n=1}^{N} \alpha_n \left(1 - y_n x_n^\top w\right) + \frac{\lambda}{2} \|w\|_2^2}_{G(w,\alpha)}$$

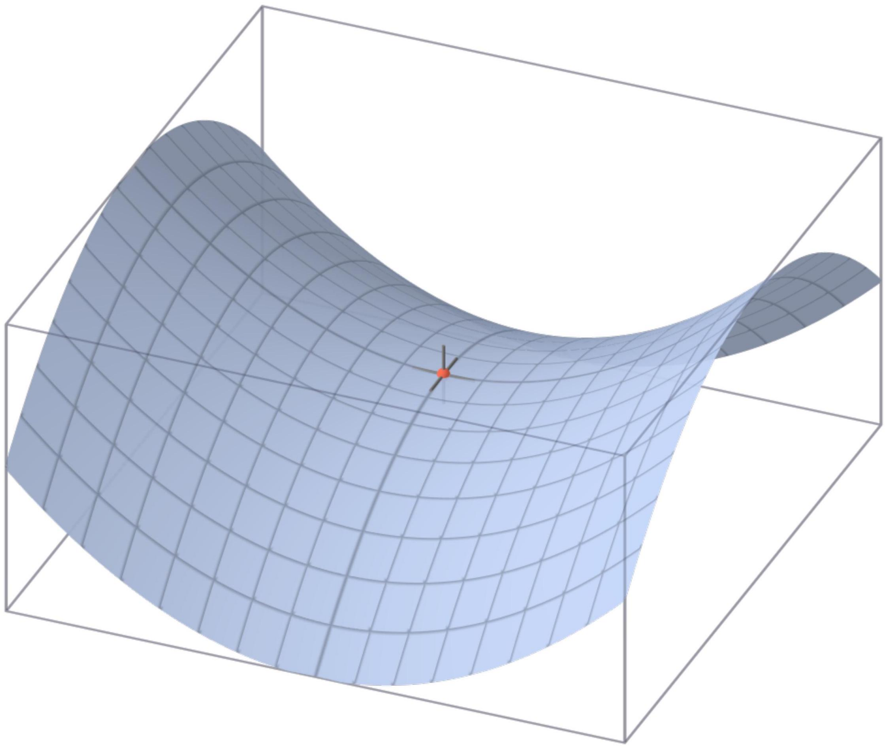
The function G is convex in w and concave in α

Q2: Can the min and max be interchanged?

Always true:

$$\max_\alpha \min_w G(w, \alpha) \leq \min_w \max_\alpha G(w, \alpha)$$

Equality if G is convex in w, concave in α and the domains of w and α are convex and compact:
$\max \min G(w, \alpha) = \min \max G(w, \alpha)$



Q2: Can the min and max be interchanged?

Always true:

$$\max_{\alpha} \min_{\mathrm{w}} G(w, \alpha) \leq \min_{w} \max_{\alpha} G(w, \alpha)$$

Proof:

$$\min_{w} G(\alpha, w) \leq G\left(\alpha, w'\right) \text{ for any } w'$$

$$\max_{\alpha} \min_{w} G(\alpha, w) \leq \max_{\alpha} G\left(\alpha, w'\right) \text{ for any } w'$$

$$\max_{\alpha} \min_{w} G(\alpha, w) \leq \min_{w'} \max_{\alpha} G\left(\alpha, w'\right)$$

Application to SVM

For SVM, the condition is met, allowing us to interchange min and max:

$$\min_w L(w) = \max_{\alpha\in[0,1]^n} \min_w \frac{1}{N}\sum_{n=1}^{N} \alpha_n \left(1 - y_n x_n^\top w\right) + \frac{\lambda}{2}\|w\|_2^2$$

Minimizer computation:

$$\mathbf{Y} = \mathrm{diag}(\mathbf{y})$$

$\nabla_w G(w,\alpha) = -\frac{1}{N}\sum_{n=1}^{N} \alpha_n y_n x_n + \lambda w = 0 \Longrightarrow w(\alpha) = \frac{1}{\lambda N}\sum_{n=1}^{N} \alpha_n y_n x_n = \frac{1}{\lambda N}\mathbf{X}^\top \mathbf{Y}\alpha$

Dual optimization problem:

$$\begin{aligned}
\min_w L(w) &= \max_{\alpha\in[0,1]^n} \frac{1}{N}\sum_{n=1}^{N} \alpha_n \left(1 - \frac{1}{\lambda N} y_n x_n^\top \mathbf{X}^\top \mathbf{Y}\alpha\right) + \frac{1}{2\lambda N^2}\left\|\mathbf{X}^\top \mathbf{Y}\alpha\right\|_2^2 \\
&= \max_{\alpha\in[0,1]^n} \frac{1^\top \alpha}{N} - \frac{1}{\lambda N^2}\alpha^\top \mathbf{Y}\mathbf{X}\mathbf{X}^\top \mathbf{Y}\alpha + \frac{1}{2\lambda N^2}\left\|\mathbf{X}^\top \mathbf{Y}\alpha\right\|_2^2 \\
&= \max_{\alpha\in[0,1]^n} \frac{1^\top \alpha}{N} - \frac{1}{2\lambda N^2}\alpha^\top \underbrace{\mathbf{Y}\mathbf{X}\mathbf{X}^\top \mathbf{Y}}_{\text{PSD matrix}}\alpha
\end{aligned}$$

Q3: Why?

$$\max_{\alpha\in[0,1]^n} \alpha^\top 1 - \frac{1}{2\lambda N}\alpha^\top \underbrace{\mathbf{Y}\mathbf{X}\mathbf{X}^\top \mathbf{Y}}_{\text{PSD matrix}}\alpha$$

1. Differentiable Concave Problem: Efficient solutions can be achieved using

- Quadratic programming solvers
- Coordinate ascent

2. Kernel Matrix Dependency: The cost function only depends on the data via the kernel matrix $K = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{N\times N}$ - no dependency on d

3. Dual Formulation Insight: α is typically sparse and non-zero exclusively for the training examples that are crucial in determining the decision boundary

Interpretation of the dual formulation

For any (x_n, y_n), there is a corresponding α_n given by

$$\max_{\alpha_n\in[0,1]} \alpha_n \left(1 - y_n x_n^\top w\right)$$

- If x_n is on the correct side and outside the margin, $1 - y_n x_n^\top w < 0$, then $\alpha_n = 0$
- If x_n is on the correct side and on the margin, $1 - y_n x_n^\top w = 0$, then $\alpha_n \in [0, 1]$
- If x_n is strictly inside the margin or or the incorrect side, $1 - y_n x_n^\top w > 0$, then $\alpha_n = 1$

$\rightarrow$ The points for which $\alpha_n > 0$ are referred to as support vectors

$$(\alpha_n = 0 \text{ and } y_n = -1) \text{ or } (\alpha_n = 1 \text{ and } y_n = 1) \quad {}_{w^\top x = -1} w^\top x = 0$$

The SVM hyperplane is supported by

the support vectors

$$(\alpha_n = 0 \text{ and } y_n = 1) \text{ or } (\alpha_n = 1 \text{ and } y_n = -1)$$

$$w = \frac{1}{\lambda N} \sum_{n=1}^{N} \alpha_n y_n x_n$$

$\Rightarrow$ w does not depend on the observation (x_n, y_n) if $\alpha_n = 0$

$$(\alpha_n = 0 \text{ and } y_n = -1) \text{ or } (\alpha_n = 1 \text{ and } y_n = 1)$$

$$w^\top x = -1 \quad w^\top x = 0$$

Recap

- Hard SVM - finds max-margin separating hyperplane $\min_w \frac{1}{2} \|w\|^2$ such that $\forall n, y_n x_n^\top w \geq 1$
- Soft SVM - relax the constraint for non-separable data

$$\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{n=1}^{N} \left[1 - y_n x_n^\top w\right]_+$$

- Hinge loss can be optimized with (stochastic) sub-gradient method
- Duality: min max problem is equivalent to max min (convex-concave objective)
- Efficient solutions with quadratic programming and coordinate ascent
- The cost depends on the data via the kernel matrix (no dependency on d)