Supervised learning: SVM

Shikui Tu

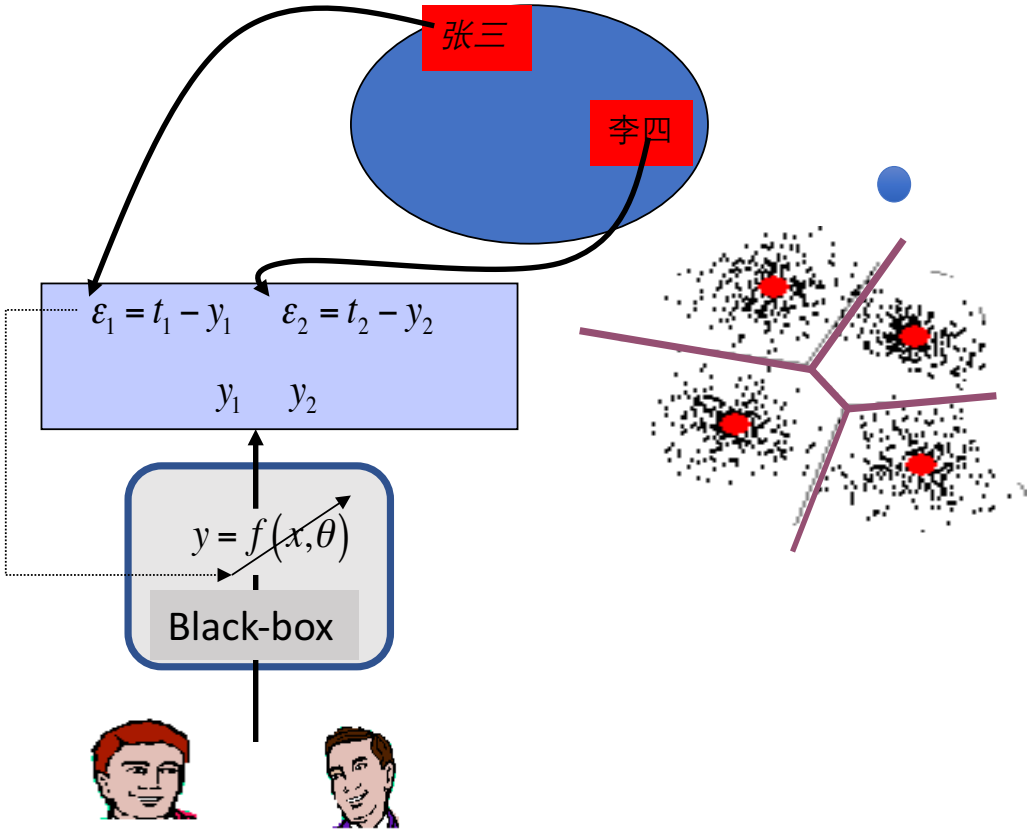
Department of Computer Science and Engineering, Shanghai Jiao Tong University

2021-05-25

Outline

- **Supervised learning**
- Preliminary on convex optimization
- Support Vector Machine (SVM)

Supervised learning



Outline

- Supervised learning

- **Preliminary on convex optimization**

- Support Vector Machine (SVM)

Convex optimization problem

standard form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, i = 1, \ldots, m \\ & a_i^T x = b_i, i = 1, \ldots, p \end{array}$$

or written as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, i = 1, \ldots, m \\ & Ax = b \end{array}$$

- $f_0, f_1, \ldots, f_m$ are convex, and equality constraints are affine.
- Feasible set of a convex optimization problem is convex, since it is the intersection of the domain of the problem, $\mathcal{D} = \cap_{i=0}^{m} \mathbf{dom}(f)$, which is a convex set, with m (convex) sublevel sets $\{x | f_i(x) \leq 0\}$ and p hyperplanes $\{x | a_i^T x = b_i\}$.

The Lagrangian

Standard form optimization problem (not necessarily convex)

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array} \tag{3}$$

with variable $x \in \mathbf{R}^n$, domain $\mathcal{D}$, optimal value p^*.

The Lagrangian

We define the Lagrangian $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$ associated with the problem by Eq.(3) as

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x). \tag{4}$$

with $\mathbf{dom}(L) = \mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$.

- λ_i: the Lagrange multiplier associated with the i-th inequality constraint $f_i(x) \leq 0$;
- ν_i the Lagrange multiplier associated with the i-th equality constraint $h_i(x) = 0$;

The Lagrange dual function

Definition

We define the Lagrange dual function (or just dual function) $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$ as the minimum value of the Lagrangian over x: $\forall \lambda \in \mathbf{R}^m, \nu \in \mathbf{R}^p$,

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \tag{5}$$

Lower bound property

The dual function yields lower bounds on the optimal value p^* of the problem by Eq.(3), i.e., $\forall \lambda \succeq 0$ and $\forall \nu$, we have

$$g(\lambda, \nu) \leq p^*.$$

Least-squares solution of linear equations

The problem

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b, \quad A \in \mathbf{R}^{p \times n} \end{array}$$

- The Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$, with domain $\mathbf{R}^n \times \mathbf{R}^p$.
- Since $L(x, \nu)$ is a convex quadratic function of x, we can find the minimum from the optimality condition by setting gradient equal to zero:

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \implies x = -\frac{1}{2} A^T \nu.$$

- Plug in in $L(x, \nu)$ to obtain the dual function $g(\nu)$:

$$g(\nu) = L(-\frac{1}{2} A^T \nu, \nu) = -\frac{1}{4} \nu^T A A^T \nu - b^T \nu$$

  which is a concave quadratic function, with domain $\mathbf{R}^p$.
- lower bound property: $p^* \geq -\frac{1}{4} \nu^T A A^T \nu - b^T \nu, \ \forall \nu$.

The Lagrange dual problem

Dual problem

For each pair (λ, ν) with $\lambda \succeq 0$, the Lagrange dual function gives us a lower bound on the optimal value p^*. What is the best lower bound that can be obtained from the Lagrange dual function? This leads to:

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{8}$$

- The original problem by Eq.(3) is sometimes called the primal problem.
- The term dual feasible means, as the name implies, that (λ, ν) is feasible for the dual problem in Eq.(8).
- We refer to (λ^*, ν^*) as dual optimal or optimal Lagrange multipliers if they are optimal for the problem in Eq.(8).
- The dual problem optimum value is denoted as d^*.

Weak duality and strong duality

Weak duality: $d^* \leq p^*$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems
  For example, solving the following problem (SDP):

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T\nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0. \end{array}$$

  gives a lower bound for the two-way partitioning problem.

Strong duality: $d^* = p^*$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called constraint qualifications

KKT conditions for nonconvex problems

Assume $f_0, \ldots, f_m$, $h_1, \ldots, h_p$ are differentiable (no assumptions on convexity):

Karush-Kuhn-Tucker (KKT) conditions

Let x^* and (λ^*, ν^*) be any primal and dual optimal points with zero duality gap. Since x^* minimizes $L(x, \lambda^*, \nu^*)$ over x, it follows that its gradient must vanish at x^*, i.e.,

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0.$$

Thus we have the following Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned} f_i(x^*) &\leq 0, \quad i = 1, \ldots, m \\ h_i(x^*) &= 0, \quad i = 1, \ldots, p \\ \lambda_i^* &\geq 0, \quad i = 1, \ldots, m \\ \lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m \end{aligned} \tag{9}$$

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0,$$

For any optimization problem with differentiable objective and constraint functions for which strong duality obtains, any pair of primal and dual optimal points must satisfy the KKT conditions.

KKT conditions for convex problems

KKT conditions for convex problems

When the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal. In other words, if f_i are convex and h_i are affine, and $\widetilde{x}, \widetilde{\lambda}, \widetilde{\nu}$ are any points that satisfy the KKT conditions:

$$
\begin{aligned}
f_i(\widetilde{x}) &\leq 0, \quad i = 1, \ldots, m \\
h_i(\widetilde{x}) &= 0, \quad i = 1, \ldots, p \\
\widetilde{\lambda}_i &\geq 0, \quad i = 1, \ldots, m \\
\widetilde{\lambda}_i f_i(\widetilde{x}) &= 0, \quad i = 1, \ldots, m
\end{aligned}
$$

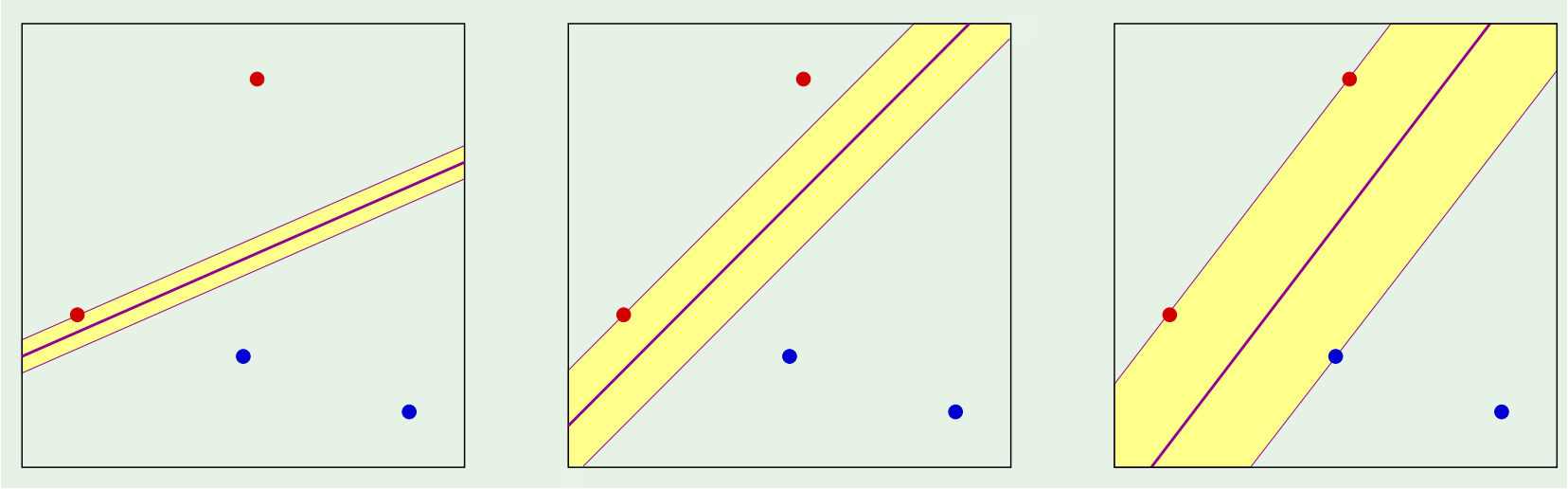
$$
\nabla f_0(\widetilde{x}) + \sum_{i=1}^{m} \widetilde{\lambda}_i \nabla f_i(\widetilde{x}) + \sum_{i=1}^{p} \widetilde{\nu}_i \nabla h_i(\widetilde{x}) = 0,
$$

then $\widetilde{x}$ and $(\widetilde{\lambda}, \widetilde{\nu})$ are primal and dual optimal, with zero duality gap.

Outline

- Supervised learning
- Preliminary on convex optimization
- Support Vector Machine (SVM)

Which is a better linear separation



Two questions:

1. Why is bigger margin better?

2. Which $\mathbf{w}$ maximizes the margin?

Finding w with large margin

Let $\mathbf{x}_n$ be the nearest data point to the plane $\mathbf{w}^T\mathbf{x} = 0$. How far is it?

2 preliminary technicalities:

1. **Normalize $\mathbf{w}$:**

   $$|\mathbf{w}^T\mathbf{x}_n| = 1$$

2. **Pull out w_0:**

   $\mathbf{w} = (w_1, \cdots, w_d)$ apart from b

   The plane is now $\boxed{\mathbf{w}^T\mathbf{x} + b = 0}$ (no x_0)

Computing the distance

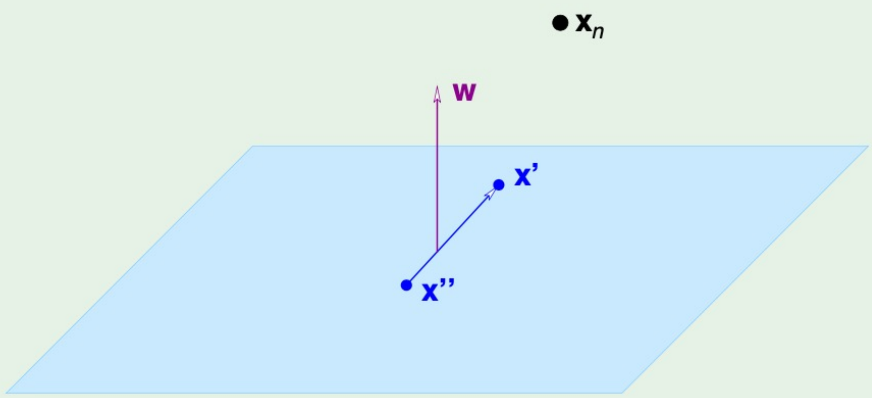
The distance between $\mathbf{x}_n$ and the plane $\mathbf{w}^\mathsf{T}\mathbf{x} + b = 0$ where $\left|\mathbf{w}^\mathsf{T}\mathbf{x}_n + b\right| = 1$

The vector $\mathbf{w}$ is $\perp$ to the plane in the $\mathcal{X}$ space:

Take $\mathbf{x}'$ and $\mathbf{x}''$ on the plane

$\mathbf{w}^\mathsf{T}\mathbf{x}' + b = 0$ and $\mathbf{w}^\mathsf{T}\mathbf{x}'' + b = 0$

$\Longrightarrow \quad \mathbf{w}^\mathsf{T}(\mathbf{x}' - \mathbf{x}'') = 0$

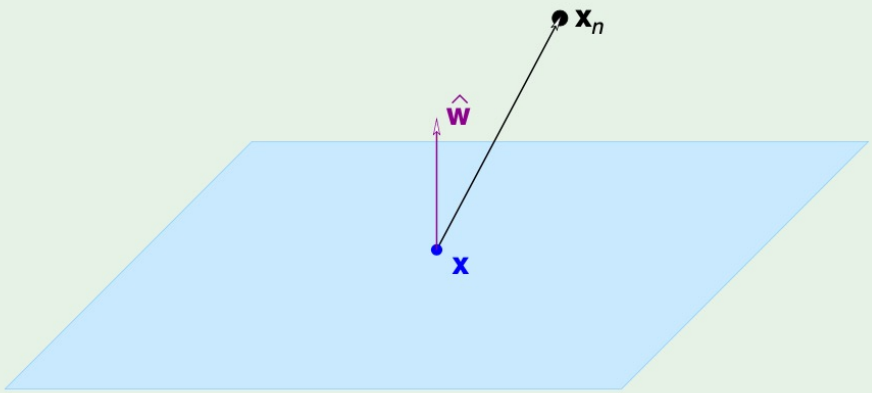


The distance from x_n to the plane

Distance between $\mathbf{x}_n$ and the plane:

Take any point $\mathbf{x}$ on the plane

Projection of $\mathbf{x}_n - \mathbf{x}$ on $\mathbf{w}$



$$\hat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \implies \text{distance} = \left|\hat{\mathbf{w}}^{\mathsf{T}}(\mathbf{x}_n - \mathbf{x})\right|$$

$$\text{distance} = \frac{1}{\|\mathbf{w}\|}\left|\mathbf{w}^{\mathsf{T}}\mathbf{x}_n - \mathbf{w}^{\mathsf{T}}\mathbf{x}\right| = \frac{1}{\|\mathbf{w}\|}\left|\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b - \mathbf{w}^{\mathsf{T}}\mathbf{x} - b\right| = \frac{1}{\|\mathbf{w}\|}$$

The SVM optimization problem

Maximize $\dfrac{1}{\|\mathbf{w}\|}$

subject to $\min_{n=1,2,\ldots,N} |\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b| = 1$

Notice: $|\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b| = y_n(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b)$

Minimize $\dfrac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w}$

subject to $y_n(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b) \geq 1$ for $n = 1, 2, \ldots, N$

Constrained optimization

Minimize $\frac{1}{2}\,\mathbf{w}^{\mathsf{T}}\mathbf{w}$

subject to $y_n\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b\right) \geq 1$ for $n = 1, 2, \ldots, N$

$\mathbf{w} \in \mathbb{R}^d,\ b \in \mathbb{R}$

Lagrange? inequality constraints $\Longrightarrow$ KKT

Lagrange formulation

Minimize $$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n (y_n (\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b) - 1)$$

w.r.t. $\mathbf{w}$ and b and maximize w.r.t. each $\alpha_n \geq 0$

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = \mathbf{0}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0$$

Dual formulation

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \qquad \text{and} \qquad \sum_{n=1}^{N} \alpha_n y_n = 0$$

in the Lagrangian $$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n \left(y_n \left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b\right) - 1\right)$$

we get $$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \, \alpha_n \alpha_m \, \mathbf{x}_n^{\mathsf{T}} \mathbf{x}_m$$

Maximize w.r.t. to $\boldsymbol{\alpha}$ subject to $\alpha_n \geq 0$ for $n = 1, \cdots, N$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$

Quadratic programming (QP)

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^{\mathsf{T}} \underbrace{\begin{bmatrix} y_1 y_1\, \mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1 & y_1 y_2\, \mathbf{x}_1^{\mathsf{T}}\mathbf{x}_2 & \dots & y_1 y_N\, \mathbf{x}_1^{\mathsf{T}}\mathbf{x}_N \\ y_2 y_1\, \mathbf{x}_2^{\mathsf{T}}\mathbf{x}_1 & y_2 y_2\, \mathbf{x}_2^{\mathsf{T}}\mathbf{x}_2 & \dots & y_2 y_N\, \mathbf{x}_2^{\mathsf{T}}\mathbf{x}_N \\ \dots & \dots & \dots & \dots \\ y_N y_1\, \mathbf{x}_N^{\mathsf{T}}\mathbf{x}_1 & y_N y_2\, \mathbf{x}_N^{\mathsf{T}}\mathbf{x}_2 & \dots & y_N y_N\, \mathbf{x}_N^{\mathsf{T}}\mathbf{x}_N \end{bmatrix}}_{\text{quadratic coefficients}} \boldsymbol{\alpha} + \underbrace{(-\mathbf{1}^{\mathsf{T}})}_{\text{linear}}\boldsymbol{\alpha}$$

subject to $\underbrace{\mathbf{y}^{\mathsf{T}}\boldsymbol{\alpha} = 0}_{\text{linear constraint}}$

$$\underbrace{\mathbf{0}}_{\text{lower bounds}} \leq \boldsymbol{\alpha} \leq \underbrace{\infty}_{\text{upper bounds}}$$

Solution

Solution: $\boldsymbol{\alpha} = \alpha_1, \cdots, \alpha_N$

$$\Longrightarrow \quad \mathbf{w} \;=\; \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

KKT condition: For $n = 1, \cdots, N$

$$\alpha_n \left(y_n \left(\mathbf{w}^{\mathsf{T}} \mathbf{x}_n + b \right) - 1 \right) \;=\; 0$$

$\alpha_n > 0 \implies \mathbf{x}_n$ is a **support vector**

Support vectors

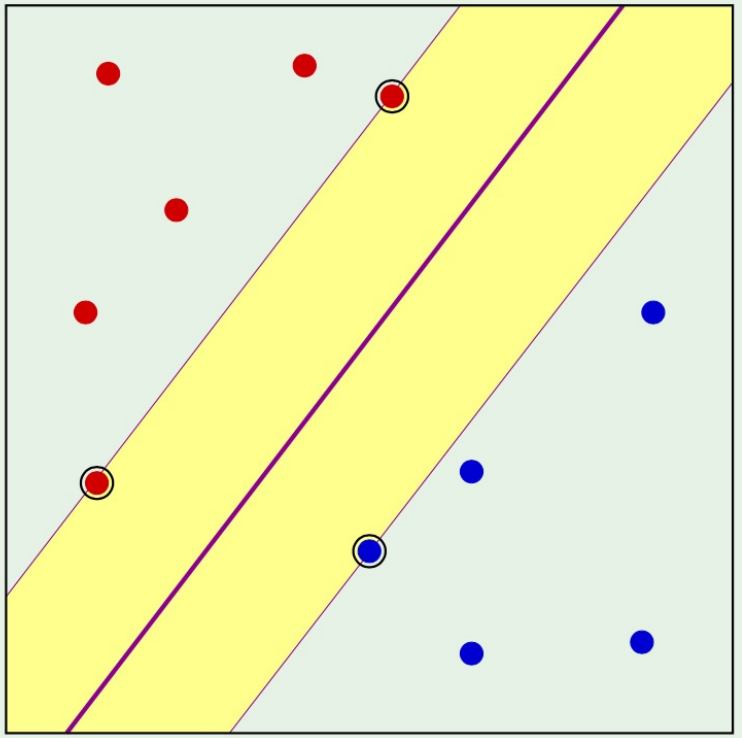
Closest $\mathbf{x}_n$'s to the plane: achieve the margin

$$\implies \quad y_n \left(\mathbf{w}^\mathsf{T}\mathbf{x}_n + b\right) = 1$$

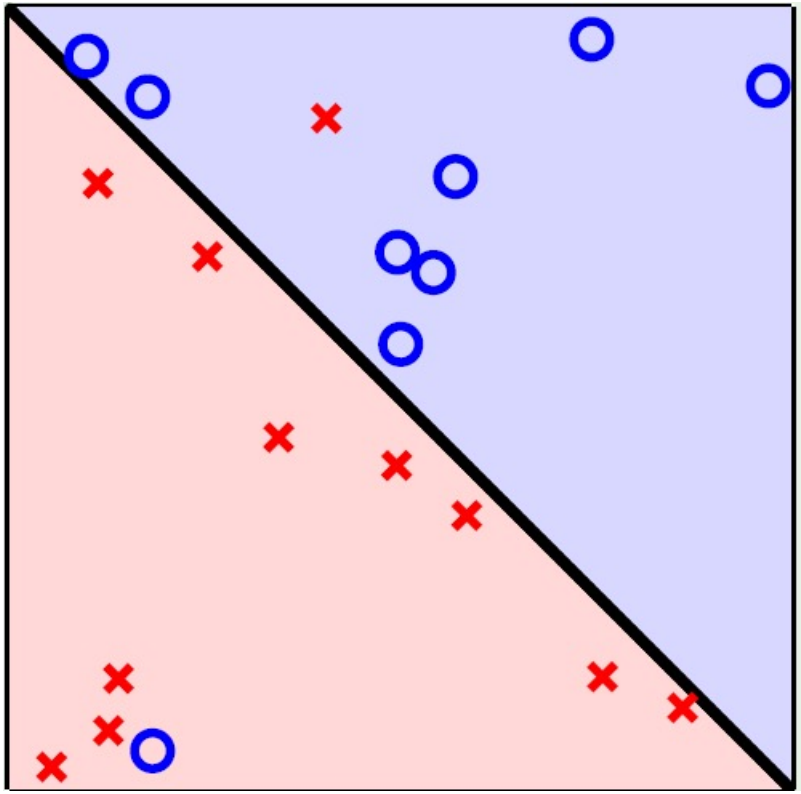
$$\mathbf{w} \;=\; \sum_{\mathbf{x}_n \text{ is SV}} \alpha_n y_n \mathbf{x}_n$$

Solve for b using any SV:

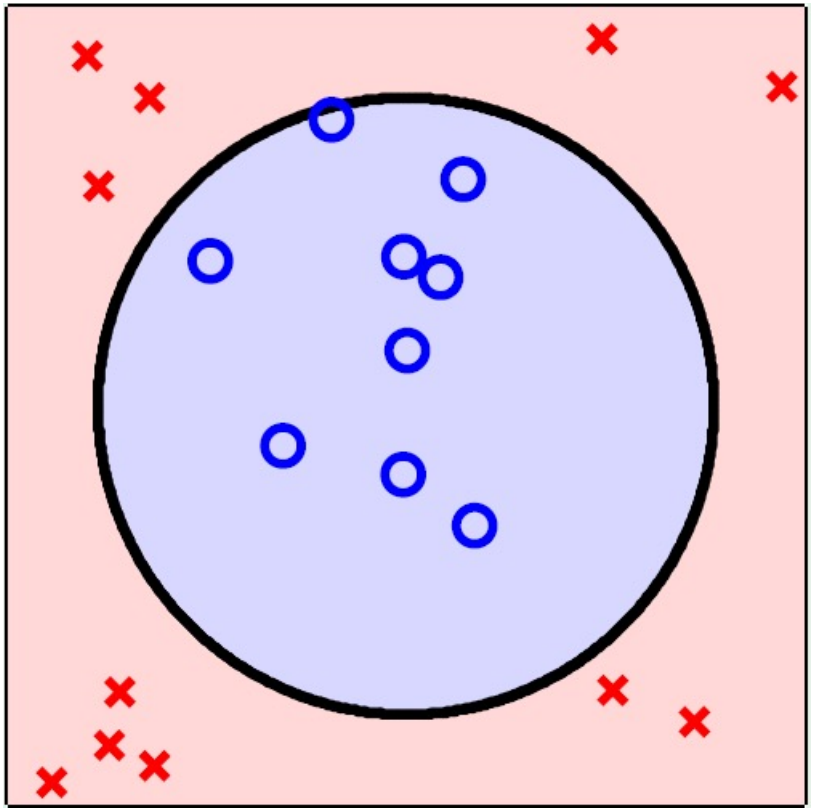
$$y_n \left(\mathbf{w}^\mathsf{T}\mathbf{x}_n + b\right) = 1$$



Non-separable Data



Slightly
non-separable

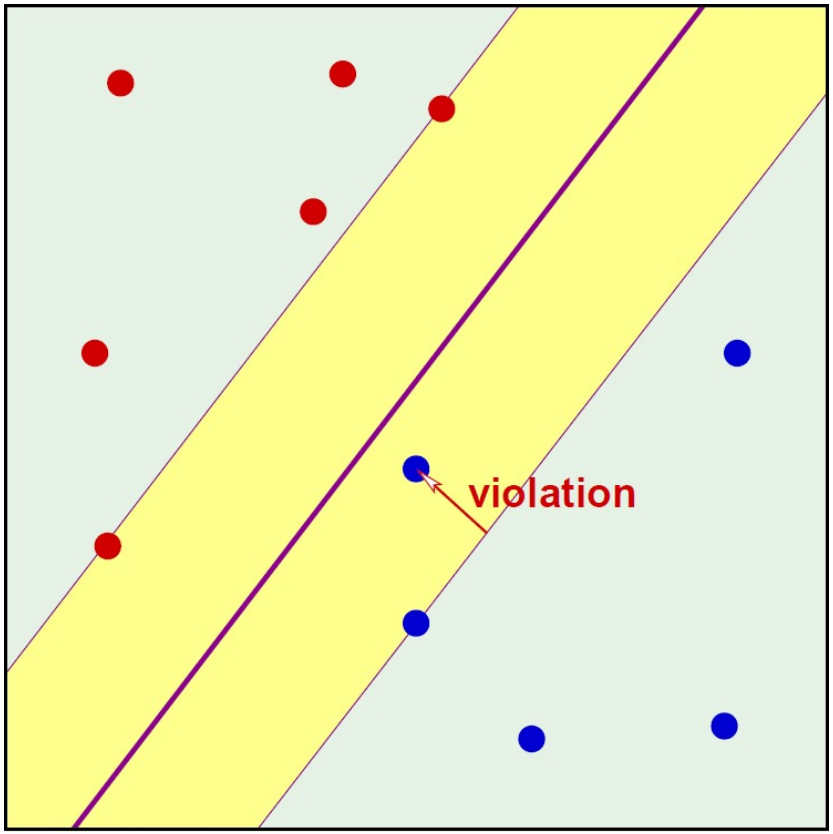


Seriously
non-separable

Soft-margin SVM

Margin violation: $y_n \left(\mathbf{w}^\intercal \mathbf{x}_n + b\right) \geq 1$ fails

Quantify: $y_n \left(\mathbf{w}^\intercal \mathbf{x}_n + b\right) \geq 1 - \xi_n$



$$\text{Total violation} = \sum_{n=1}^{N} \xi_n$$

ξ_n = ksi = "amount" of margin violation

$$\boxed{\xi_n \geq 0}$$

Soft-margin SVM optimization

- Similar to hard-margin optimization, but get compromise between maximizing the margin and allowing for violations

Still get large margin after term minimization

Allow for small violations by minimizing

$$\text{Minimize} \quad \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} + C\sum_{n=1}^{N}\xi_n$$

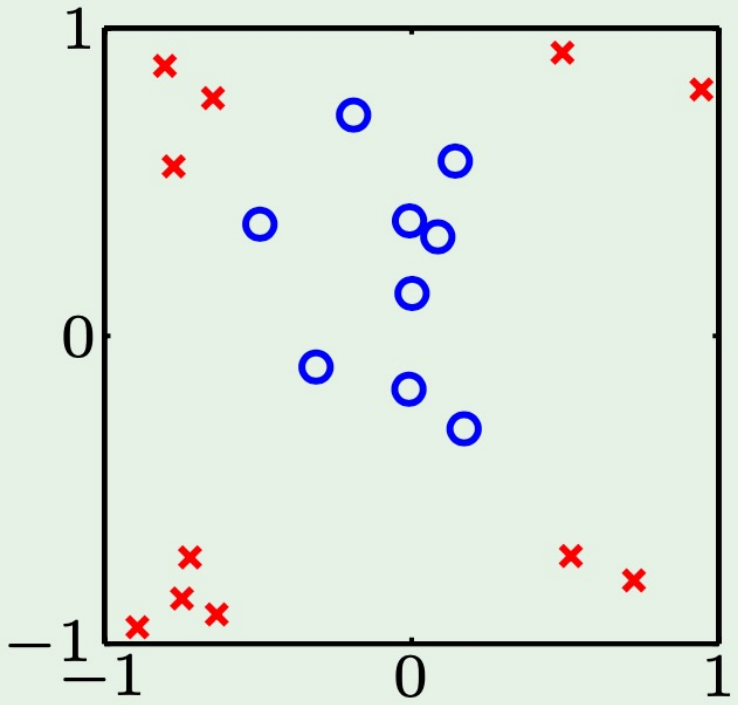
$$\text{subject to} \quad y_n\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b\right) \geq 1 - \xi_n$$

$$\text{for} \quad n = 1, \ldots, N$$

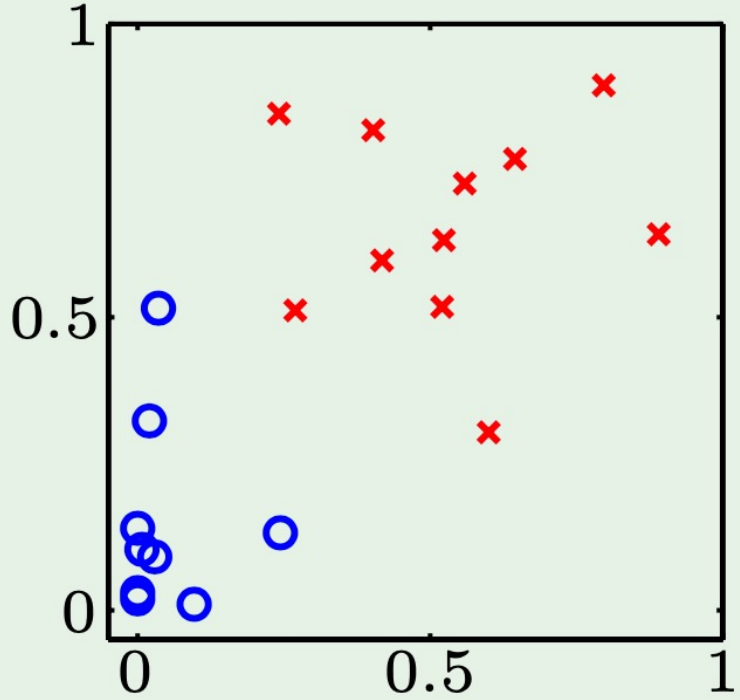
$$\text{and} \quad \xi_n \geq 0 \quad \text{for} \quad n = 1, \ldots, N$$

Kernel transform

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \, \alpha_n \alpha_m \, \mathbf{z}_n^{\mathsf{T}} \mathbf{z}_m$$



$\mathcal{X} \longrightarrow \mathcal{Z}$



Thank you!