MAC 0460 / 5832
Introduction to Machine Learning

18 — Support Vector Machines (SVM) – Part II

• hyperplane • margin • margin violation •
• QP problems • dual problems • kernel trick •

IME/USP (21/06/2021)

SVM roadmap

- Binary linear classification: The linearly separable case
- hard-margin SVM: Maximum margin formulation
- Binary linear classification: The non-linearly separable case
- soft-margin SVM: allows margin violation
- hard-margin/soft-margin SVM is a QP problem
- Dual of hard-margin/soft-margin SVM is also QP
- How to solve QP problems
- Non-linear classification: the kernel trick

Dual QP optimization

Recall primal and dual formulation

When we discussed regularization, we started with the following optimization problem

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & E_{in}(\mathbf{w}) \\ \text{subject to:} \quad & \mathbf{w}^T\mathbf{w} \leq C \end{aligned}$$

and we ended up solving the following problem

$$\underset{\mathbf{w}}{\text{minimize}} \quad E_{aug}(\mathbf{w}) = E_{in}(\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}$$

That is, we started with a problem with constraints on $\mathbf{w}$ and ended up with a problem without such constraints (only one constraint: $\lambda \geq 0$)

Introduction to the dual formulation, first with one constraint

QP with one constraint:

$$\begin{aligned} &\underset{\mathbf{u}}{\text{minimize}} \quad && \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} \\ &\text{subject to:} \quad && \mathbf{a}^T\mathbf{u} \geq c \quad \text{(one constraint)} \end{aligned}$$

Fact: If there is an optimal solution $\mathbf{u}^*$ for the above problem, then $\mathbf{u}^*$ is also an optimal solution of the following problem:

$$\underset{\mathbf{u}}{\text{minimize}} \quad \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} + \max_{\alpha \geq 0} \alpha(c - \mathbf{a}^T\mathbf{u})$$

Why? Since $\mathbf{a}^T\mathbf{u} \geq c \iff c - \mathbf{a}^T\mathbf{u} \leq 0$, then $\max_{\alpha \geq 0} \alpha(c - \mathbf{a}^T\mathbf{u}) = 0$

Introduction to the dual formulation, first with one constraint

Dual formulation

$$\underset{\mathbf{u}}{\text{minimize}} \quad \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} + \max_{\alpha \geq 0} \alpha(c - \mathbf{a}^T\mathbf{u})$$

The term $\alpha(c - \mathbf{a}^T\mathbf{u})$ forces $c - \mathbf{a}^T\mathbf{u}$ to stay negative – i.e., to satisfy the constraint $\mathbf{a}^T\mathbf{u} \geq c$ (because this helps to minimize the cost function). On the other hand, α is chosen so as to maximize $\alpha(c - \mathbf{a}^T\mathbf{u})$ (to avoid $c - \mathbf{a}^T\mathbf{u}$ going to $-\infty$)

There is no constraints
We have a min-max optimization problem

Lagrangean

$$\mathcal{L}(\mathbf{u}, \alpha) = \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} + \alpha(c - \mathbf{a}^T\mathbf{u})$$

Min-max optimization

$$\min_{\mathbf{u}} \max_{\alpha \geq 0} \mathcal{L}(\mathbf{u}, \alpha)$$

Strong duality

$$\min_{\mathbf{u}} \max_{\alpha \geq 0} \mathcal{L}(\mathbf{u}, \alpha) = \max_{\alpha \geq 0} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \alpha)$$

Dual formulation

Theorem 8.7 (KKT)**.** For a feasible convex QP-problem in *primal* form,

$$\begin{array}{lll} \underset{\mathbf{u}\in\mathbb{R}^L}{\text{minimize:}} & \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathrm{Q}\mathbf{u}+\mathbf{p}^{\mathrm{T}}\mathbf{u} & \\ \text{subject to:} & \mathbf{a}_m^{\mathrm{T}}\mathbf{u} \geq c_m & (m=1,\cdots,M), \end{array}$$

define the Lagrange function

$$\mathcal{L}(\mathbf{u},\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathrm{Q}\mathbf{u}+\mathbf{p}^{\mathrm{T}}\mathbf{u}+\sum_{m=1}^{M}\alpha_m\left(c_m-\mathbf{a}_m^{\mathrm{T}}\mathbf{u}\right).$$

The solution $\mathbf{u}^*$ is optimal for the primal if and only if $(\mathbf{u}^*,\boldsymbol{\alpha}^*)$ is a solution to the dual optimization problem

$$\max_{\boldsymbol{\alpha}\geq\mathbf{0}}\min_{\mathbf{u}}\mathcal{L}(\mathbf{u},\boldsymbol{\alpha}).$$

KKT conditions

The optimal $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$ satisfies the Karush-Kühn-Tucker (KKT) conditions:

(i) *Primal and dual constraints:*

$$\mathbf{a}_m^{\mathrm{T}}\mathbf{u}^* \geq c_m \qquad \text{and} \qquad \alpha_m \geq 0 \qquad\qquad (m = 1, \cdots, M).$$

(ii) *Complementary slackness:*

$$\alpha_m^* \left(\mathbf{a}_m^{\mathrm{T}}\mathbf{u}^* - c_m\right) = 0.$$

(iii) *Stationarity with respect to* $\mathbf{u}$*:*

$$\nabla_{\mathbf{u}}\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha})\big|_{\mathbf{u}=\mathbf{u}^*, \boldsymbol{\alpha}=\boldsymbol{\alpha}^*} = \mathbf{0}.$$

Solving the dual QP optimization problems

Dual: characterized by the Lagrangean $\mathcal{L}$

It is a min-max problem

$$\min_{\mathbf{u}} \max_{\alpha \geq 0} \mathcal{L}(\mathbf{u}, \alpha) = \max_{\alpha \geq 0} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \alpha)$$

Two-step optimization:

1. We fix α and optimize $\min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \alpha)$
2. Then, we fix $\mathbf{u}$ and optimize $\max_{\alpha} \mathcal{L}(\mathbf{u}, \alpha)$

Dual formulation for the hard-margin SVM

Primal formulation:

$$\begin{array}{ll} \underset{\mathbf{w},b}{\text{minimize}} & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, n = 1, \ldots, N \end{array}$$

Lagrangean function

$$\begin{aligned} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)\right) \\ &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^{\mathrm{T}}\mathbf{x}_n - b\sum_{n=1}^{N} \alpha_n y_n + \sum_{n=1}^{N} \alpha_n. \end{aligned}$$

Solution of the dual hard-margin SVM

$$\begin{aligned}\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)\right) \\ &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^{\mathrm{T}}\mathbf{x}_n - b\sum_{n=1}^{N} \alpha_n y_n + \sum_{n=1}^{N} \alpha_n.\end{aligned}$$

1. Minimize $\mathcal{L}(b, \mathbf{w}, \alpha)$ with respect to $(b, \mathbf{w})$:

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n \qquad \text{and} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n.$$

Computing the zero:

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

Solution of the dual hard-margin SVM

$$\begin{aligned}\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)\right) \\ &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^{\mathrm{T}}\mathbf{x}_n - b\sum_{n=1}^{N} \alpha_n y_n + \sum_{n=1}^{N} \alpha_n.\end{aligned}$$

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

The Lagrangean is reduced to a function on α only:

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

Solution of the dual hard-margin SVM

The Lagrangean is reduced to a function on α only:

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

2. Now we would like to maximize $\mathcal{L}$: Minimize $-\mathcal{L}$ with respect to α

$$\begin{aligned}
\underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{minimize:}} \quad & \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m - \sum_{n=1}^{N}\alpha_n \\
\text{subject to:} \quad & \sum_{n=1}^{N} y_n \alpha_n = 0 \\
& \alpha_n \geq 0 \qquad\qquad (n = 1, \cdots, N).
\end{aligned}$$

This is a standard QP and we can solve it using Solvers!

Recall: Standard form of QP problems

M inequality constraints and Q positive semi-definite

$$\begin{aligned} \underset{\mathbf{u}}{\text{minimize}} \quad & \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} \\ \text{subject to:} \quad & \mathbf{a}_m^T\mathbf{u} \geq c_m \quad (m = 1, \ldots, M) \end{aligned}$$

In matrix form

$$\begin{aligned} \underset{\mathbf{u}}{\text{minimize}} \quad & \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} \\ \text{subject to:} \quad & A\mathbf{u} \geq \mathbf{c} \end{aligned}$$

QP solvers can be used to compute the optimal solution $\mathbf{u}^*$:

$$\mathbf{u}^* \leftarrow \mathrm{QP}(Q, \mathbf{p}, A, \mathbf{c})$$

Solution of the dual hard-margin SVM

Minimization of $-\mathcal{L}$ with respect to α is a standard QP:

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{minimize:}} \quad \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{Q}_{\mathrm{D}}\boldsymbol{\alpha} - \mathbf{1}_N^{\mathrm{T}}\boldsymbol{\alpha} \qquad (8.$$

$$\text{subject to:} \quad \mathrm{A}_{\mathrm{D}}\boldsymbol{\alpha} \geq \mathbf{0}_{N+2},$$

where Q_{D} and A_{D} (D for dual) are given by:

$$\mathrm{Q}_{\mathrm{D}} = \begin{bmatrix} y_1 y_1 \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1 & \ldots & y_1 y_N \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_N \\ y_2 y_1 \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_1 & \ldots & y_2 y_N \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_N \\ \vdots & \vdots & \vdots \\ y_N y_1 \mathbf{x}_N^{\mathrm{T}}\mathbf{x}_1 & \ldots & y_N y_N \mathbf{x}_N^{\mathrm{T}}\mathbf{x}_N \end{bmatrix} \text{ and } \mathrm{A}_{\mathrm{D}} = \begin{bmatrix} \mathbf{y}^{\mathrm{T}} \\ -\mathbf{y}^{\mathrm{T}} \\ \mathrm{I}_{N\times N} \end{bmatrix}$$

$$\alpha^* \leftarrow QP(Q_D, -\mathbf{1}, A_D, \mathbf{0})$$

Computing $(b^*, \mathbf{w}^*)$ from α^*

From the stationarity condition, we derived $$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n.$$

$\Longrightarrow$ Thus, $\mathbf{w}^* = \sum_{n=1}^{N} y_n \alpha_n^* \mathbf{x}_n$

At least one α_s^* is strictly positive (suposing we have examples from both classes). From the complementary slackness, we know that

$$\alpha_m^* \left(\mathbf{a}_m^{\mathrm{T}} \mathbf{u}^* - c_m\right) = 0.$$

$\Longrightarrow$ Thus equality holds for sample s:

$$y_s({\mathbf{w}^*}^T \mathbf{x}_s + b^*) = 1 \Longrightarrow b^* = y_s - \sum_{n=1}^{N} y_n \alpha_n^* \mathbf{x}_n^T \mathbf{x}_s$$

Hard-Margin SVM with Dual QP

1: Construct Q_D and A_D as in Exercise 8.11

$$Q_D = \begin{bmatrix} y_1 y_1 \mathbf{x}_1^T \mathbf{x}_1 & \ldots & y_1 y_N \mathbf{x}_1^T \mathbf{x}_N \\ y_2 y_1 \mathbf{x}_2^T \mathbf{x}_1 & \ldots & y_2 y_N \mathbf{x}_2^T \mathbf{x}_N \\ \vdots & \vdots & \vdots \\ y_N y_1 \mathbf{x}_N^T \mathbf{x}_1 & \ldots & y_N y_N \mathbf{x}_N^T \mathbf{x}_N \end{bmatrix} \text{ and } A_D = \begin{bmatrix} \mathbf{y}^T \\ -\mathbf{y}^T \\ I_{N \times N} \end{bmatrix}.$$

2: Use a QP-solver to optimize the dual problem:

$$\boldsymbol{\alpha}^* \leftarrow \mathsf{QP}(Q_D, -\mathbf{1}_N, A_D, \mathbf{0}_{N+2}).$$

3: Let s be a support vector for which $\alpha_s^* > 0$. Compute b^*,

$$b^* = y_s - \sum_{\alpha_n^* > 0} y_n \alpha_n^* \mathbf{x}_n^T \mathbf{x}_s.$$

4: Return the final hypothesis

$$g(\mathbf{x}) = \text{sign}\left(\sum_{\alpha_n^* > 0} y_n \alpha_n^* \mathbf{x}_n^T \mathbf{x} + b^* \right).$$

Solvers

There are software packages for solving optimization problems

Python Notebook `svm_cvxpy.ipynb`

- uses CVXPY (https://www.cvxpy.org/)
- presents solutions for the toy example
    - using a solver for standard optimization
    - using a solver for the standard QP formulation
      (Linear Hard-Margin SVM with QP)
    - using a solver for the dual QP formulation
      (Hard Margin SVM with Dual QP)

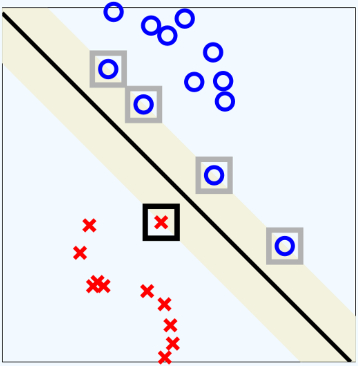
Interpretation of the solution

Support vectors: $\alpha_s > 0 \Longrightarrow y_s({\mathbf{w}^*}^T\mathbf{x}_s + b^*) = 1$

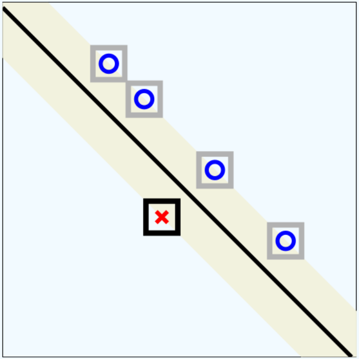
Weights: $\mathbf{w}^* = \sum_{n=1}^{N} y_n \alpha_n^* \mathbf{x}_n$

Bias: $b^* = y_s - {\mathbf{w}^*}^T\mathbf{x}_s = y_s - \sum_{n=1}^{N} y_n \alpha_n^* \mathbf{x}_n$

Hypothesis: $g(\mathbf{x}) = \mathrm{sign}\Big(\sum_{\alpha_n^* > 0} y_n \alpha_n^* \mathbf{x}_n^T \mathbf{x} + b^* \Big)$



(a) All data



(b) Only support vectors

Solution of the soft-margin SVM

The soft-margin SVM is also a QP

Only with more constraints

Thus the same discussion on QP and dual QP holds

The new optimization

Minimize $\quad \frac{1}{2}\,\mathbf{w}^{\intercal}\mathbf{w} + C\sum_{n=1}^{N}\xi_n$

subject to $\quad y_n\left(\mathbf{w}^{\intercal}\mathbf{x}_n + b\right) \geq 1 - \xi_n \quad$ for $\quad n = 1, \ldots, N$

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

$\mathbf{w} \in \mathbb{R}^d \;,\; b \in \mathbb{R} \;,\; \boldsymbol{\xi} \in \mathbb{R}^N$

Lagrange formulation

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\, \mathbf{w}^{\intercal}\mathbf{w} + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \alpha_n (y_n \left(\mathbf{w}^{\intercal}\mathbf{x}_n + b\right) - 1 + \xi_n) - \sum_{n=1}^{N} \beta_n\, \xi_n$$

Minimize w.r.t. $\mathbf{w}$, b, and $\boldsymbol{\xi}$ and maximize w.r.t. each $\alpha_n \geq 0$ and $\beta_n \geq 0$

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

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

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \beta_n = 0$$

Hard × soft margin

Optimization of $\mathcal{L}$ with respect to α:

Hard-margin

$$\begin{aligned}
\underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{minimize:}} \quad & \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \\
\text{subject to:} \quad & \sum_{n=1}^{N} y_n \alpha_n = 0 \\
& \alpha_n \geq 0 \qquad\qquad (n = 1, \cdots, N).
\end{aligned}$$

Soft-margin (it is also a QP problem)

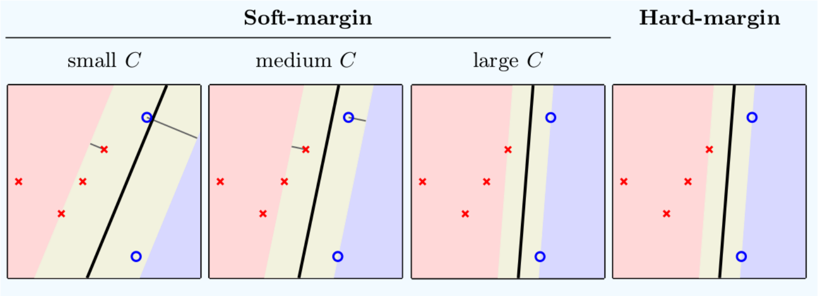
$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}} \mathrm{Q_D} \boldsymbol{\alpha} - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} \\
\text{subject to} \quad & \mathbf{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0; . \\
& \mathbf{0} \leq \boldsymbol{\alpha} \leq C \cdot \mathbf{1}.
\end{aligned}$$

Interpretation of C

$$\text{minimize } \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

$$O \leq \alpha_n \leq C$$

Interpretation of C



$0 < \alpha_n^* < C \implies \mathbf{x}_n$ is a support vector
$\alpha_n^* = 0 \implies \mathbf{x}_n$ is beyond the margin on the right side
$\alpha_n^* = C \implies \mathbf{x}_n$ is inside the margin or in the wrong side

When the data is linearly separable, there exists C such that the soft-margin SVM solution is exactly the same solution of the hard-margin SVM

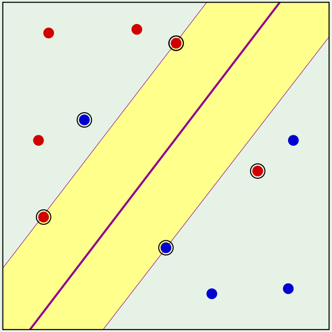
Types of support vectors

margin support vectors $(0 < \alpha_n < C)$

$$y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) = 1 \qquad (\xi_n \ = \ 0)$$

non-margin support vectors $(\alpha_n = C)$

$$y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) < 1 \qquad (\xi_n \ > \ 0)$$

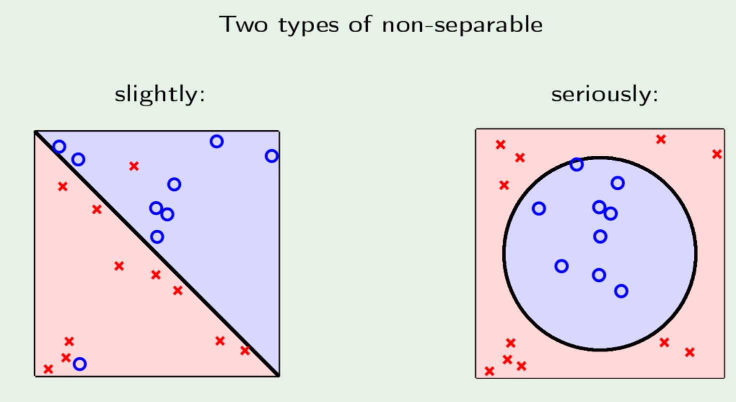


The Kernel trick

Motivation

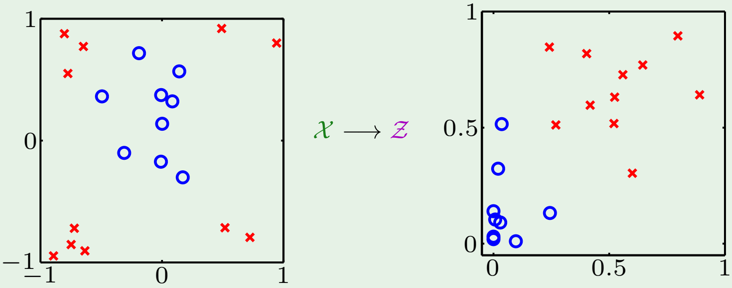
Soft-margin SVM could be used to solve non-linear cases

Would we get good solutions for both examples below ?



z instead of x

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



What is the problem?

When we map data $\mathbf{x} \in \mathbb{R}^d$ to $\mathbf{z} \in \mathbb{R}^{\tilde{d}}$, $\tilde{d} >> d$

huge $\tilde{d}$

What do we need from the $\mathcal{Z}$ space?

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

Constraints: $\alpha_n \geq 0$ for $n = 1, \cdots, N$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$

$$\boxed{g(\mathbf{x}) = \operatorname{sign}\left(\mathbf{w}^{\mathsf{T}} \mathbf{z} + b\right)}$$ need $\mathbf{z}_n^{\mathsf{T}} \mathbf{z}$

where $\mathbf{w} = \sum_{\mathbf{z}_n \text{ is SV}} \alpha_n y_n \mathbf{z}_n$

and b: $y_m \left(\mathbf{w}^{\mathsf{T}} \mathbf{z}_m + b\right) = 1$ need $\mathbf{z}_n^{\mathsf{T}} \mathbf{z}_m$

Kernel trick

Is there any kernel function $K()$ satisfying

$$K_\Phi(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}')$$

and such that computation is more efficient than computing $\mathbf{z}^T \mathbf{z}' = \Phi(\mathbf{x})^T \Phi(\mathbf{x}')$?

If there is suck $K()$, then

$$\mathrm{Q_D} = \begin{bmatrix} y_1 y_1 \mathbf{x}_1^{\mathrm{T}} \mathbf{x}_1 & \dots & y_1 y_N \mathbf{x}_1^{\mathrm{T}} \mathbf{x}_N \\ y_2 y_1 \mathbf{x}_2^{\mathrm{T}} \mathbf{x}_1 & \dots & y_2 y_N \mathbf{x}_2^{\mathrm{T}} \mathbf{x}_N \\ \vdots & \vdots & \vdots \\ y_N y_1 \mathbf{x}_N^{\mathrm{T}} \mathbf{x}_1 & \dots & y_N y_N \mathbf{x}_N^{\mathrm{T}} \mathbf{x}_N \end{bmatrix} \qquad \mathrm{Q_D} = \begin{bmatrix} y_1 y_1 \mathrm{K}_{11} & \dots & y_1 y_N \mathrm{K}_{1N} \\ y_2 y_1 \mathrm{K}_{21} & \dots & y_2 y_N \mathrm{K}_{2N} \\ \vdots & \vdots & \vdots \\ y_N y_1 \mathrm{K}_{N1} & \dots & y_N y_N \mathrm{K}_{NN} \end{bmatrix}$$

Kernel K would be equivalent to mapping $\mathbf{x}$ to $\mathbf{z}$ and applying dual SVM on $\mathbf{z}$, but without explicitly computing $\mathbf{z}$!

Hard-Margin SVM with Kernel

1: Construct $\mathrm{Q_D}$ from the kernel K, and $\mathrm{A_D}$:

$$\mathrm{Q_D} = \begin{bmatrix} y_1y_1\mathrm{K}_{11} & \dots & y_1y_N\mathrm{K}_{1N} \\ y_2y_1\mathrm{K}_{21} & \dots & y_2y_N\mathrm{K}_{2N} \\ \vdots & \vdots & \vdots \\ y_Ny_1\mathrm{K}_{N1} & \dots & y_Ny_N\mathrm{K}_{NN} \end{bmatrix} \quad \text{and} \quad \mathrm{A_D} = \begin{bmatrix} \mathbf{y}^{\mathrm{T}} \\ -\mathbf{y}^{\mathrm{T}} \\ \mathrm{I}_{N\times N} \end{bmatrix},$$

where $\mathrm{K}_{mn} = K(\mathbf{x}_m, \mathbf{x}_n)$. (K is called the *Gram* matrix.)

2: Use a QP-solver to optimize the dual problem:

$$\boldsymbol{\alpha}^* \leftarrow \mathsf{QP}(\mathrm{Q_D}, -\mathbf{1}_N, \mathrm{A_D}, \mathbf{0}_{N+2}).$$

3: Let s be any support vector for which $\alpha_s^* > 0$. Compute

$$b^* = y_s - \sum_{\alpha_n^*>0} y_n\alpha_n^* K(\mathbf{x}_n, \mathbf{x}_s).$$

4: Return the final hypothesis

$$g(\mathbf{x}) = \operatorname{sign}\left(\sum_{\alpha_n^*>0} y_n\alpha_n^* K(\mathbf{x}_n, \mathbf{x}) + b^*\right).$$

The final hypothesis

Express $\quad g(\mathbf{x}) = \text{sign}\left(\mathbf{w}^{\intercal}\mathbf{z} + b\right) \quad$ in terms of $K(-,-)$

$$\mathbf{w} \;=\; \sum_{\mathbf{z}_n \text{ is SV}} \alpha_n y_n \mathbf{z}_n \quad \Longrightarrow \quad g(\mathbf{x}) \;=\; \text{sign}\left(\sum_{\alpha_n > 0} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + b\right)$$

where $\;b \;=\; y_m - \sum_{\alpha_n > 0} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}_m)$

for any support vector $(\alpha_m > 0)$

Examples of kernel

- Linear: $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$

- Polynomial of order Q: $K(\mathbf{x}, \mathbf{x}') = (\zeta + \gamma \mathbf{x}^T \mathbf{x}')^Q$, $\zeta, \gamma > 0$

- Gaussian RBF: $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma ||\mathbf{x} - \mathbf{x}'||^2)$, $\gamma > 0$

Polynomial kernel

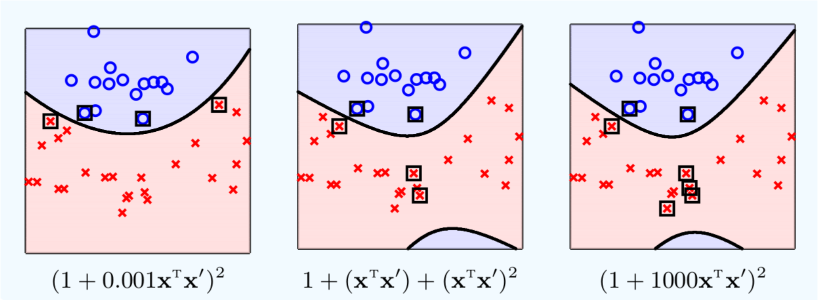
$\mathbf{x} = (x_1, \ldots, x_d)$
$\mathbf{z} = \Phi(\mathbf{x}) = (1, x_1, \ldots, x_d, x_1\,x_1, x_1\,x_2, \ldots, x_2\,x_1, \ldots, \ldots, x_d\,x_d)$

Dimension of $\mathbf{z}$: $\tilde{d} = 1 + d + d^2$

$$\begin{aligned}
\Phi(\mathbf{x})^T\Phi(\mathbf{x}') &= 1 + \sum_{i=1}^{d} x_i x_i' + \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j\, x_i' x_j' \\
&= 1 + \mathbf{x}^T\mathbf{x}' + (\sum_{i=1}^{d} x_i x_i')(\sum_{j=1}^{d} x_j x_j') \\
&= 1 + \mathbf{x}^T\mathbf{x}' + (\mathbf{x}^T\mathbf{x}')^2 = (1 + \mathbf{x}^T\mathbf{x})^2
\end{aligned}$$

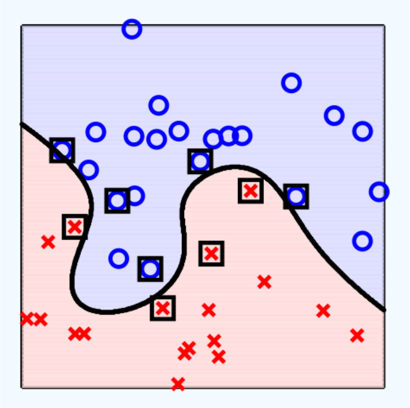
Computational complexity: from $\mathcal{O}(\tilde{d})$ to $\mathcal{O}(d)$

Example: polynomial kernel (degree 2)



$$K(\mathbf{x}, \mathbf{x}') = (\zeta + \gamma \mathbf{x}^T \mathbf{x}')^Q$$

Example: polynomial kernel (degree 10)



$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2)$ ($\gamma > 0$)

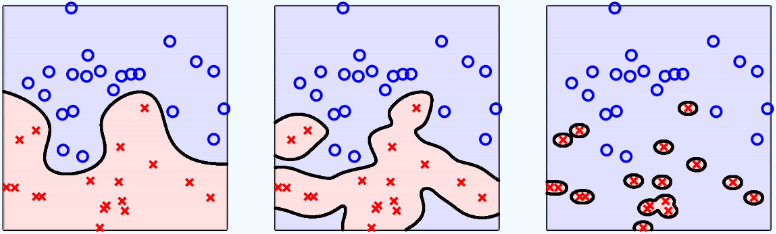
Expanding it for the case when $d = 1$

$$
\begin{aligned}
K(x, x') &= \exp\left(-\|x - x'\|^2\right) \\
&= \exp\left(-(x)^2\right) \cdot \exp(2xx') \cdot \exp\left(-(x')^2\right) \\
&= \exp\left(-(x)^2\right) \cdot \left(\sum_{k=0}^{\infty} \frac{2^k (x)^k (x')^k}{k!}\right) \cdot \exp\left(-(x')^2\right),
\end{aligned}
$$

$$
\Phi(x) = \exp(-x^2) \cdot \left(1, \sqrt{\frac{2^1}{1!}}x, \sqrt{\frac{2^2}{2!}}x^2, \sqrt{\frac{2^3}{3!}}x^3, \ldots\right)
$$

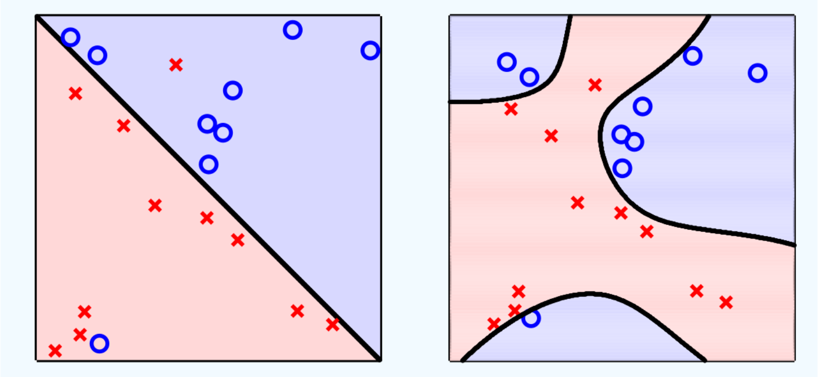
That means $\tilde{d} = \infty$!

Example: Gaussian-RBF kernel



$\exp(-1\|\mathbf{x} - \mathbf{x}'\|^2)$ $\exp(-10\|\mathbf{x} - \mathbf{x}'\|^2)$ $\exp(-100\|\mathbf{x} - \mathbf{x}'\|^2)$

Example: linear $\times$ Gaussian-RBF kernels



(a) linear classifier (b) Gaussian-RBF kernel

How do we know that $\mathcal{Z}$ exists ...

... for a given $K(\mathbf{x}, \mathbf{x}')$? valid kernel

Three approaches:

1. By construction
2. Math properties (*Mercer's condition*)
3. Who cares? ☺

Design your own kernel

$K(\mathbf{x}, \mathbf{x}')$ is a valid kernel iff

1. It is symmetric and **2.** The matrix:

$$\begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \dots & K(\mathbf{x}_1, \mathbf{x}_N) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \dots & K(\mathbf{x}_2, \mathbf{x}_N) \\ \dots & \dots & \dots & \dots \\ K(\mathbf{x}_N, \mathbf{x}_1) & K(\mathbf{x}_N, \mathbf{x}_2) & \dots & K(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

is **positive semi-definite**

for any $\mathbf{x}_1, \cdots, \mathbf{x}_N$ (Mercer's condition)

Summary

- SVM solves the maximum margin problem
- The optimization problem, in its primal form, can be written as a QP problem
- Its dual formulation (Lagrangean) also reduces to a QP problem
- Primal: d variables and N constraints
- Dual: N variables (α_n) and $N+1$ constraints
- Dual is more interesting when we use the kernel trick (very large $\tilde{d}$)
- Dual finds the support vectors

The long journey ends here ...

If you got to this point, you are a survivor :-)