Support vector machine (SVM)

CE-477: Machine Learning - CS-828: Theory of Machine Learning
Sharif University of Technology
Fall 2024

Fatemeh Seyyedsalehi

References of the lecture are mentioned in the last slide

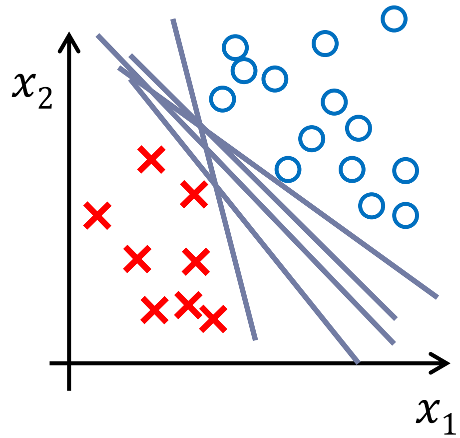
Outline

- Margin concept
- Hard-Margin SVM
  - Dual Problem of Hard-Margin SVM
- Soft-Margin SVM
  - Dual Problem of Soft-Margin SVM

Margin

- Which line is better to select as the boundary to provide more generalization capability?

Larger margin provides better generalization to unseen data

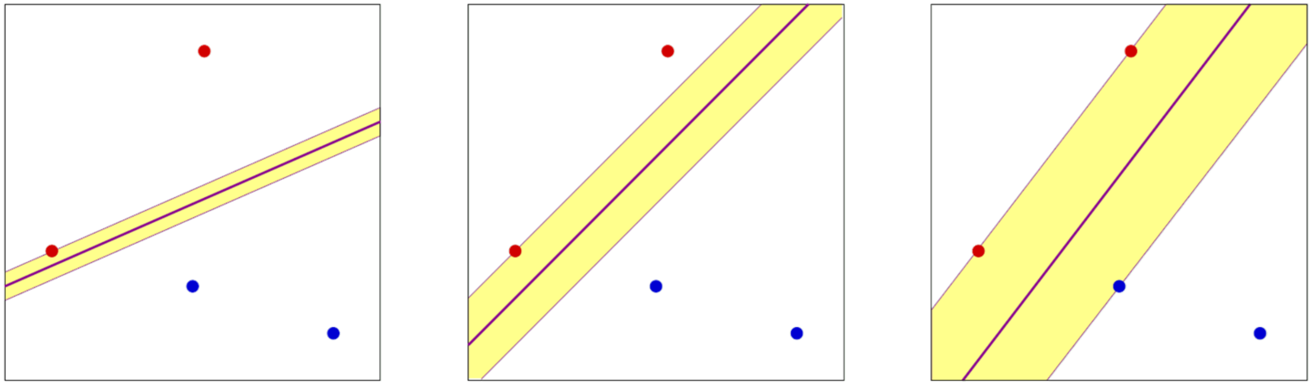


- **Margin** for a hyperplane that separates samples of two linearly separable classes is:
  - The smallest distance between the decision boundary and any of the training samples

What is better linear separation

- Linearly separable data

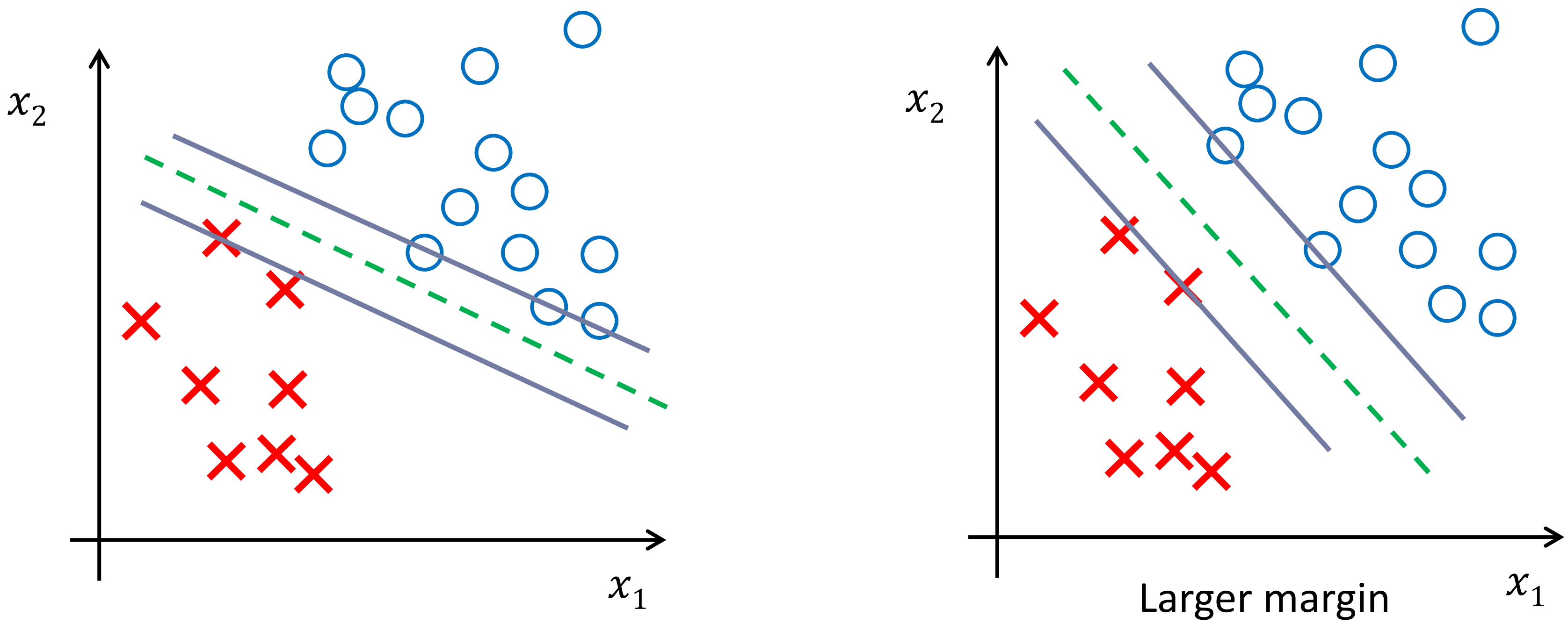
- Which line is better?



- Why the bigger margin?

Maximum margin

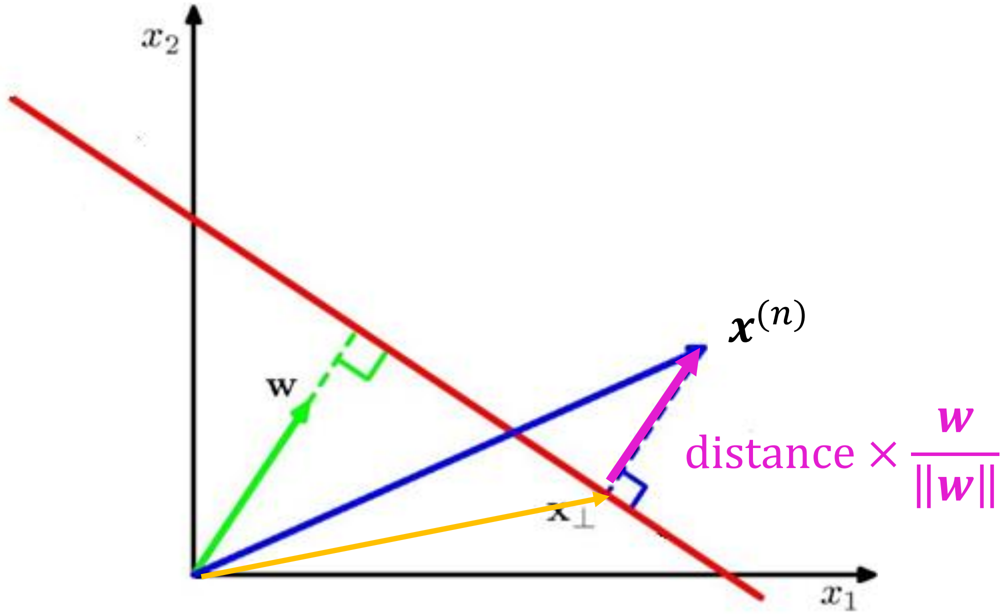
- SVM finds the solution with maximum margin



- The hyperplane with the largest margin has equal distances to the nearest sample of both classes

Distance between an $\boldsymbol{x}^{(n)}$ and the plane

$$\text{distance} = \frac{\left|\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0\right|}{\|\boldsymbol{w}\|}$$



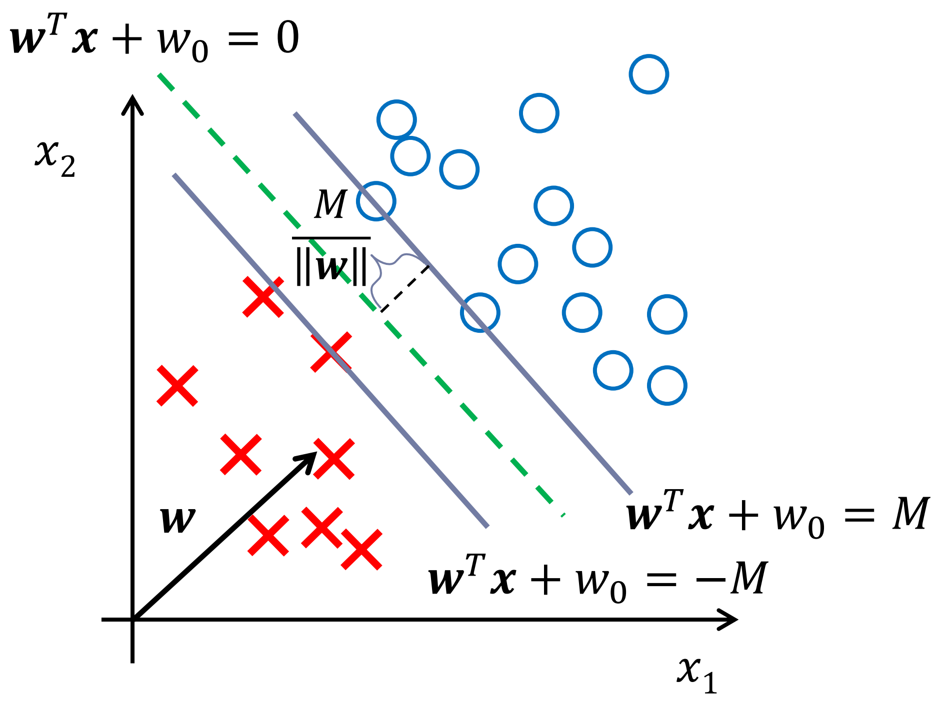
Hard-margin SVM: Optimization problem

$$\max_{M,\boldsymbol{w},w_0} \frac{2M}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq M \quad \forall \boldsymbol{x}^{(i)} \in C_1 \longrightarrow y^{(i)} = 1$$

$$\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \leq -M \quad \forall \boldsymbol{x}^{(i)} \in C_2 \longrightarrow y^{(i)} = -1$$

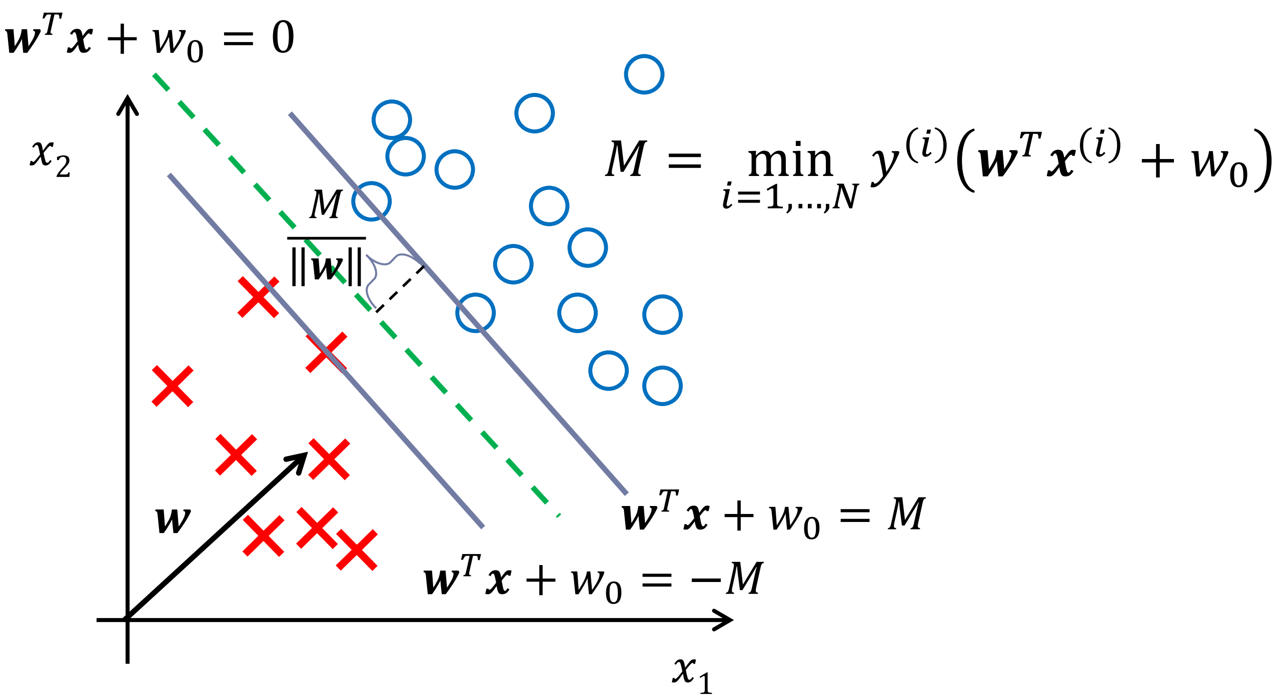
Margin: $2\frac{M}{\|\boldsymbol{w}\|}$



Hard-margin SVM: Optimization problem

$$\max_{M,\boldsymbol{w},w_0} \frac{2M}{\|\boldsymbol{w}\|}$$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq M \quad i = 1, \dots, N$$

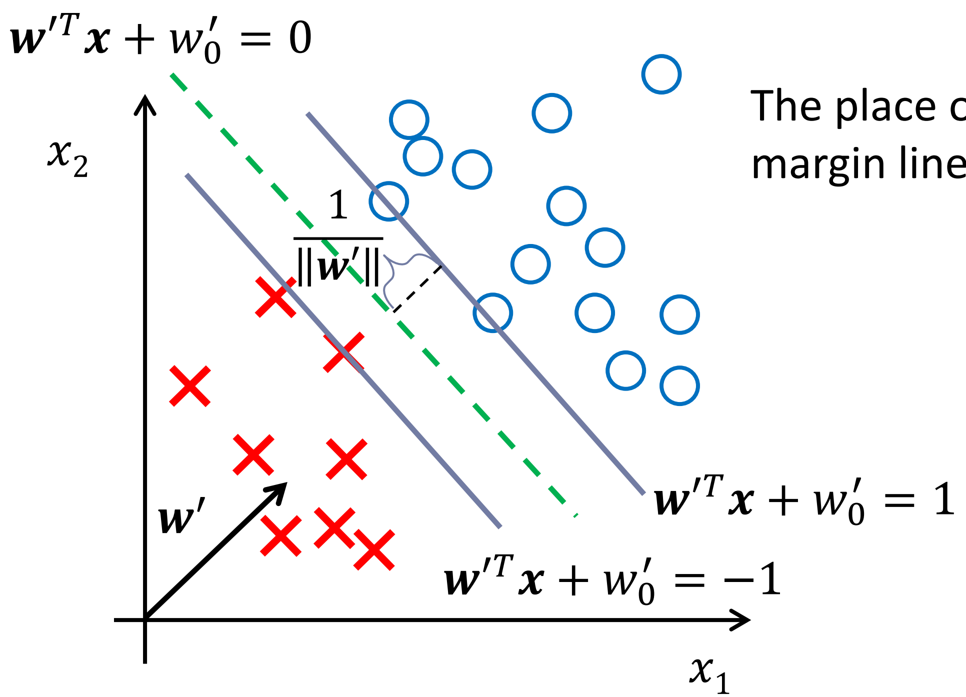


Hard-margin SVM: Optimization problem

Rescaling parameters, $\boldsymbol{w}' = \frac{\boldsymbol{w}}{M}$, $w_0' = \frac{w_0}{M}$:

$$\max_{\boldsymbol{w}', w_0'} \frac{2}{\|\boldsymbol{w}'\|}$$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{w}'^T \boldsymbol{x}^{(i)} + w_0'\right) \geq 1 \quad i = 1, \dots, N$$



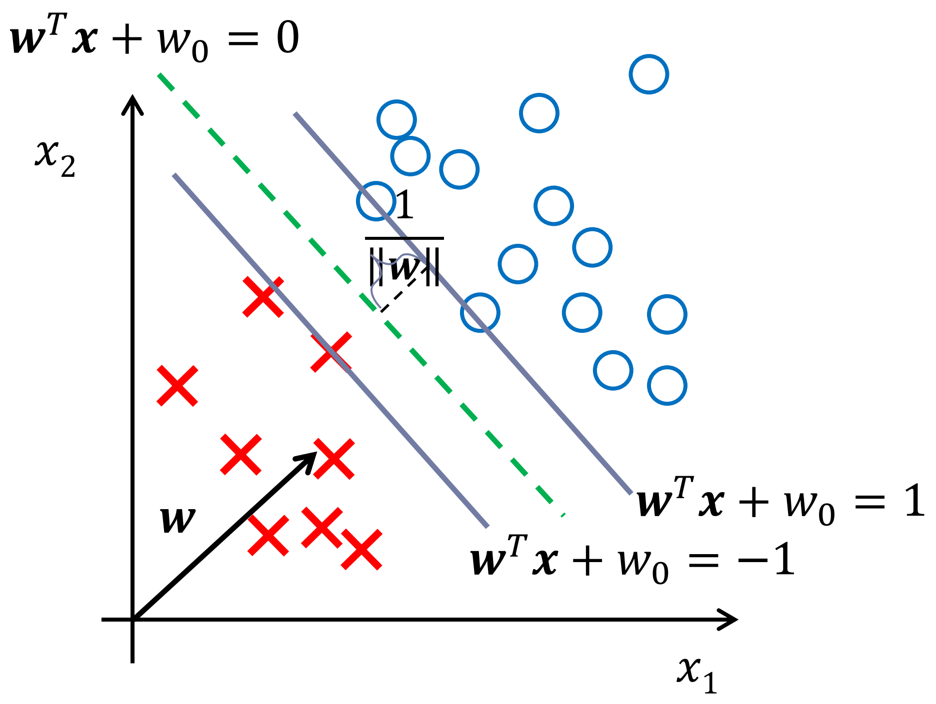
The place of boundary and margin lines do not change

Hard-margin SVM: Optimization problem

$$\max_{\boldsymbol{w},w_0} \frac{2}{\|\boldsymbol{w}\|}$$

$$\text{s.t.}\ y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+w_0\right)\geq 1\ ,n=1,\ldots,N$$

Margin: $\frac{2}{\|\boldsymbol{w}\|}$



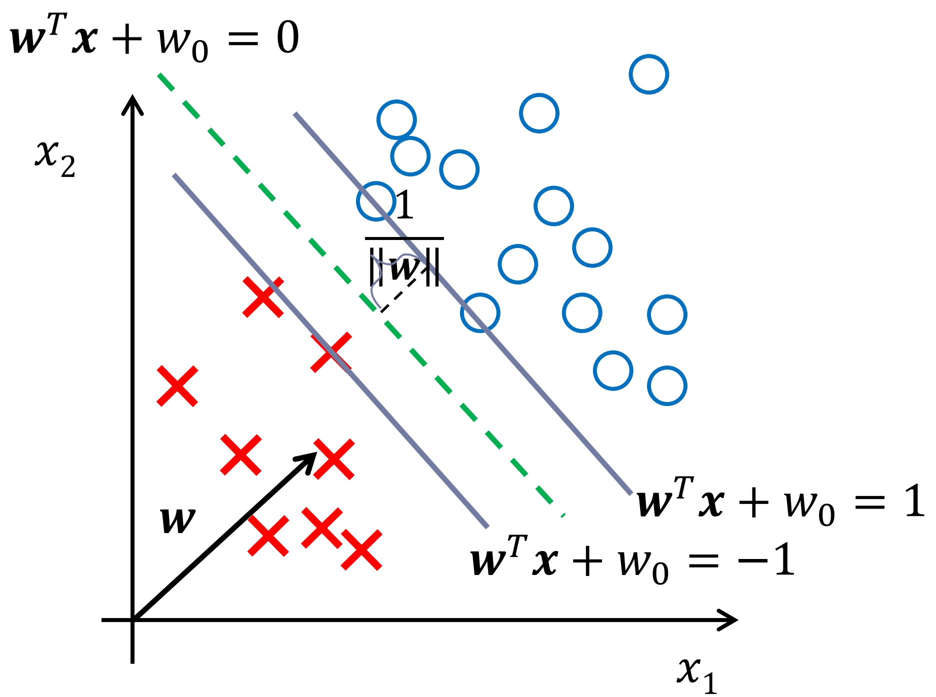
Hard-margin SVM: Optimization problem

$$\max_{\boldsymbol{w}, w_0} \frac{2}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \left(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0\right) \geq 1 \quad \forall y^{(n)} = 1$$

$$\left(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0\right) \leq -1 \quad \forall y^{(n)} = -1$$

Margin: $\frac{2}{\|\boldsymbol{w}\|}$



Hard-margin SVM: Optimization problem

We can equivalently optimize:

$$\min_{\boldsymbol{w},w_0} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{s.t.} \quad y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 \quad n = 1, \dots, N$$

- It is a convex Quadratic Programming (QP) problem
  - There are computationally efficient packages to solve it and find optimum $\boldsymbol{w}$ and w_0, i.e. the decision boundary.
  - It has a <u>global</u> minimum (if any).

Dual formulation of the SVM

- We are going to introduce the *dual* SVM problem which is equivalent to the original *primal* problem
  - Gives us further insights into the optimal hyper-plane
  - Enable us to exploit the kernel trick

- Lagrangian multipliers technique
  - An optimization method useful for problems with equality or inequality constraints

Lagrangian multipliers technique

- Considering following convex optimization problem with convex constraints

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{s.t. } g_i(\boldsymbol{x}) \leq 0 \quad i = 1, \ldots, m$$

- We can construct the following Lagrangian function

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i \, g_i(\boldsymbol{x})$$

Lagrangian multipliers

- And optimize:

$$\min_{\boldsymbol{x}} \max_{\{\alpha_i \geq 0\}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha})$$
$$\max_{\{\alpha_i \geq 0\}} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha})$$

Hard-margin SVM: Dual problem

$$\min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \quad i = 1, \dots, N$$

- By incorporating the constraints through Lagrangian multipliers, we will have:

$$\min_{\boldsymbol{w},w_0} \max_{\{\alpha_n \geq 0\}} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)\right\}$$

Hard-margin SVM: Dual problem

$$\min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \quad i = 1, \dots, N$$

- By incorporating the constraints through Lagrangian multipliers, we will have:

$$\min_{\boldsymbol{w},w_0} \max_{\{\alpha_n \geq 0\}} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n\left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)\right\}$$

- Dual problem (changing the order of min and max in the above problem):

$$\max_{\{\alpha_n \geq 0\}} \min_{\boldsymbol{w},w_0} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n\left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)\right\}$$

Hard-margin SVM: Dual problem

$$\max_{\{\alpha_n \geq 0\}} \boxed{\min_{\boldsymbol{w}, w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}$$

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0)\right)$$

Hard-margin SVM: Dual problem

$$\max_{\{\alpha_n \geq 0\}} \min_{\boldsymbol{w}, w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$$

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0\right)\right)$$

$$\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = 0 \Rightarrow \boldsymbol{w} - \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

Hard-margin SVM: Dual problem

$$\max_{\{\alpha_n \geq 0\}} \min_{\boldsymbol{w}, w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$$

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0)\right)$$

$$\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = 0 \Rightarrow \boldsymbol{w} - \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial w_0} = 0 \Rightarrow \quad -\underbrace{\sum_{n=1}^{N} \alpha_n y^{(n)} = 0}$$

w_0 do not appear, instead, a “global” constraint on $\boldsymbol{\alpha}$ is created.

Substituting

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} \qquad \sum_{n=1}^{N} \alpha_n y^{(n)} = 0$$

In the Largrangian

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0)\right)$$

Substituting

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} \qquad \sum_{n=1}^{N} \alpha_n y^{(n)} = 0$$

In the Largrangian

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\mathbf{w}^T \boldsymbol{x}^{(n)} + w_0)\right)$$

We get

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(m)}$$

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

Hard-margin SVM: Dual problem

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(m)} \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$$\alpha_n \geq 0 \quad n = 1, \ldots, N$$

- The dual form is a convex QP too!

Solution

- Quadratic programming:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T \begin{bmatrix} y^{(1)}y^{(1)}\boldsymbol{x}^{(1)^T}\boldsymbol{x}^{(1)} & \cdots & y^{(1)}y^{(N)}\boldsymbol{x}^{(1)^T}\boldsymbol{x}^{(N)} \\ \vdots & \ddots & \vdots \\ y^{(N)}y^{(1)}\boldsymbol{x}^{(N)^T}\boldsymbol{x}^{(1)} & \cdots & y^{(N)}y^{(N)}\boldsymbol{x}^{(N)^T}\boldsymbol{x}^{(N)} \end{bmatrix}\boldsymbol{\alpha} + (-\mathbf{1})^T\boldsymbol{\alpha}$$

$$\begin{aligned} \text{s.t.} -\boldsymbol{\alpha} &\leq \mathbf{0} \\ \boldsymbol{y}^T\boldsymbol{\alpha} &= \mathbf{0} \end{aligned}$$

Finding the hyperplane

- After finding $\boldsymbol{\alpha}$ by QP, we find $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

- How to find w_0?
  - we discuss it after introducing support vectors

Karush-Kuhn-Tucker (KKT) conditions

- Necessary conditions for the solution $[\boldsymbol{w}^*, w_0^*, \boldsymbol{\alpha}^*]$:
  - $\alpha_n^* \geq 0 \quad n = 1, \dots, N$
  - $y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + w_0^*\right) \geq 1 \quad n = 1, \dots, N$
  - $\alpha_i^* \left(1 - y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + w_0^*\right)\right) = 0 \quad n = 1, \dots, N$

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{s.t. } g_i(\boldsymbol{x}) \leq 0 \quad i = 1, \dots, m$$

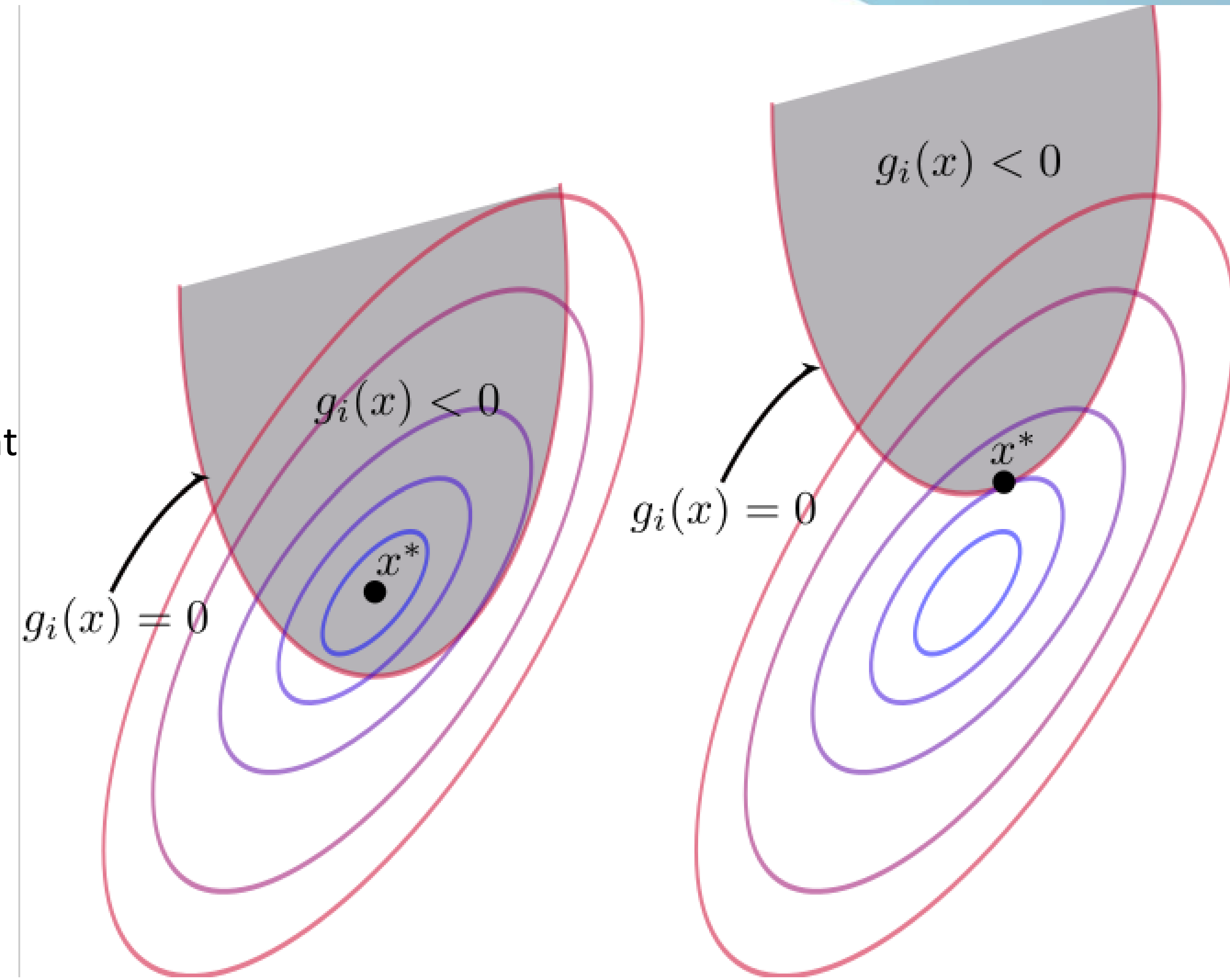
$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum \alpha_i \, g_i(\boldsymbol{x})$$

In general, the optimal $\boldsymbol{x}^*, \boldsymbol{\alpha}^*$ satisfies KKT conditions:

$$\alpha_i^* \geq 0 \quad i = 1, \dots, m$$
$$g_i(\boldsymbol{x}^*) \leq 0 \quad i = 1, \dots, m$$
$$\alpha_i^* g_i(\boldsymbol{x}^*) = 0 \quad i = 1, \dots, m$$

Karush-Kuhn-Tucker (KKT) conditions

Inactive constraint ($\alpha = 0$)

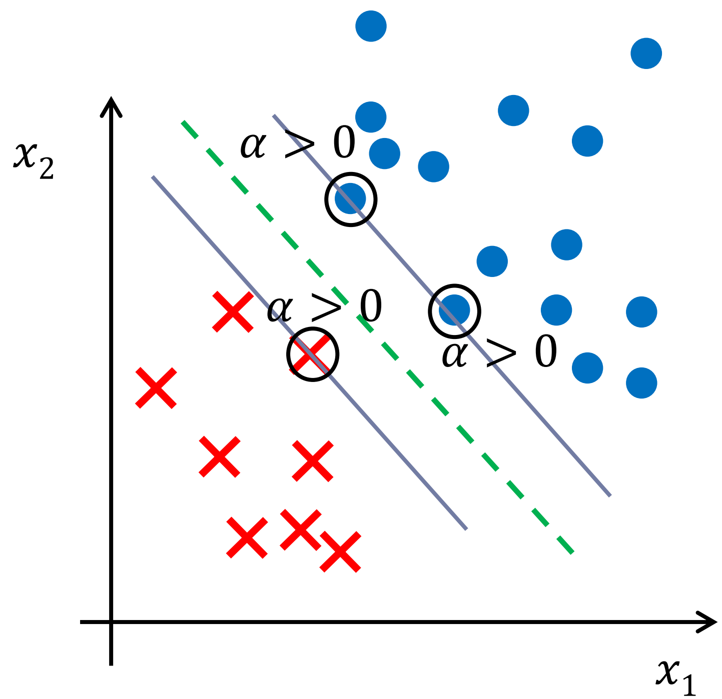


Active constraint

[wikipedia]

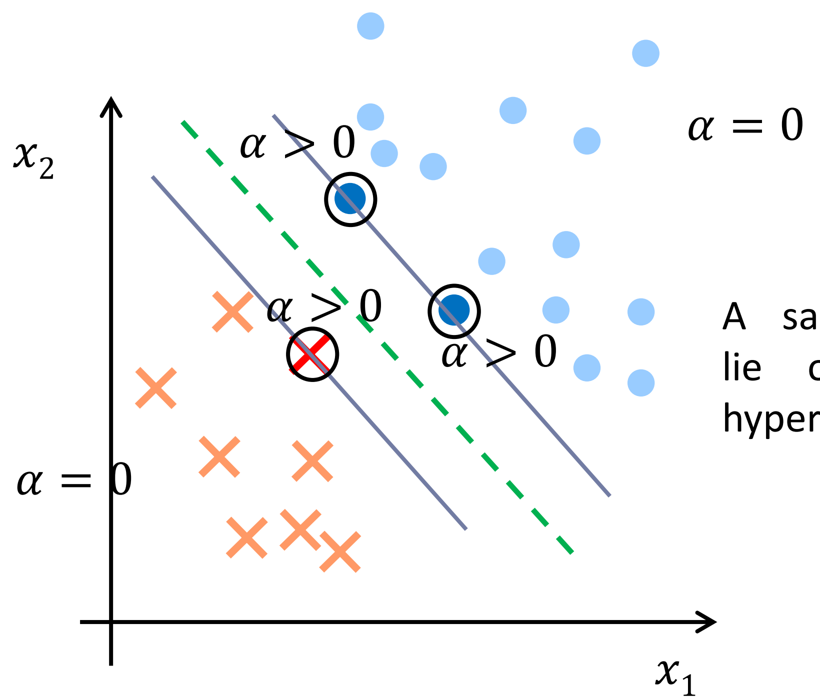
Hard-margin SVM: Support vectors

- **Inactive** constraint: $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) > 1$
  - $\Rightarrow \alpha_n = 0$ and thus $\boldsymbol{x}^{(n)}$ is not a support vector.
- **Active** constraint: $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) = 1$
  - $\Rightarrow \alpha_n$ can be greater than 0 and thus $\boldsymbol{x}^{(i)}$ can be a support vector.



Hard-margin SVM: Support vectors

- **Inactive** constraint: $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) > 1$
  - $\Rightarrow \alpha_n = 0$ and thus $\boldsymbol{x}^{(n)}$ is not a support vector.
- **Active** constraint: $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) = 1$



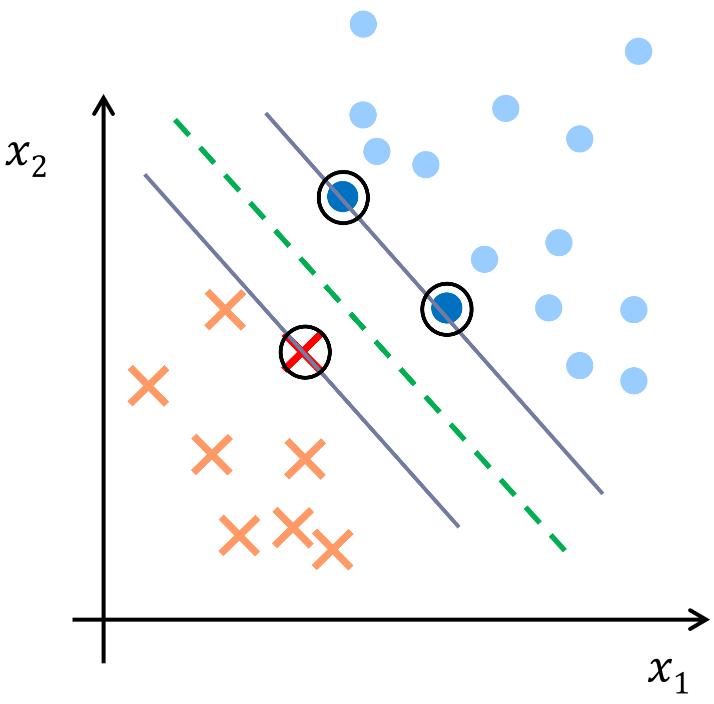
A sample with $\alpha_n > 0$ can also lie on one of the margin hyperplanes

Hard-margin SVM: Support vectors

- Support Vectors (SVs)= $\{\boldsymbol{x}^{(n)} | \alpha_n > 0\}$

- The **direction** of hyper-plane can be found only based on support vectors:

$$\boldsymbol{w} = \sum_{\alpha_n > 0} \alpha_n \, y^{(n)} \boldsymbol{x}^{(n)}$$



Finding the hyperplane

- After finding $\boldsymbol{\alpha}$ by QP, we find $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

- How to find w_0?
  - Each of the samples that has $\alpha_s > 0$ is on the margin, thus we solve for w_0 using any of SVs:

$$y^{(s)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(s)} + w_0\right) = 1$$

$$\Rightarrow w_0 = y^{(s)} - \boldsymbol{w}^T \boldsymbol{x}^{(s)}$$

Hard-margin SVM: Dual problem
Classifying new samples using only SVs

- Classification of a new sample $\boldsymbol{x}$:

$$\hat{y} = \text{sign}(w_0 + \boldsymbol{w}^T \boldsymbol{x})$$

$$\hat{y} = \text{sign}\left(w_0 + \left(\sum_{\alpha_n > 0} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} \right)^T \boldsymbol{x} \right)$$

$$\hat{y} = \text{sign}(\underbrace{y^{(s)} - \sum_{\alpha_n > 0} \alpha_n y^{(n)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(s)}}_{w_0} + \sum_{\alpha_n > 0} \alpha_n y^{(n)} \boldsymbol{x}^{(n)^T} \boldsymbol{x})$$

Support vectors are sufficient to predict labels of new samples

- The classifier is based on the expansion in terms of dot products of $\boldsymbol{x}$ with support vectors.

Hard-margin SVM dual problem: An important property

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(m)} \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$$\alpha_n \geq 0 \ \ n = 1, \dots, N$$

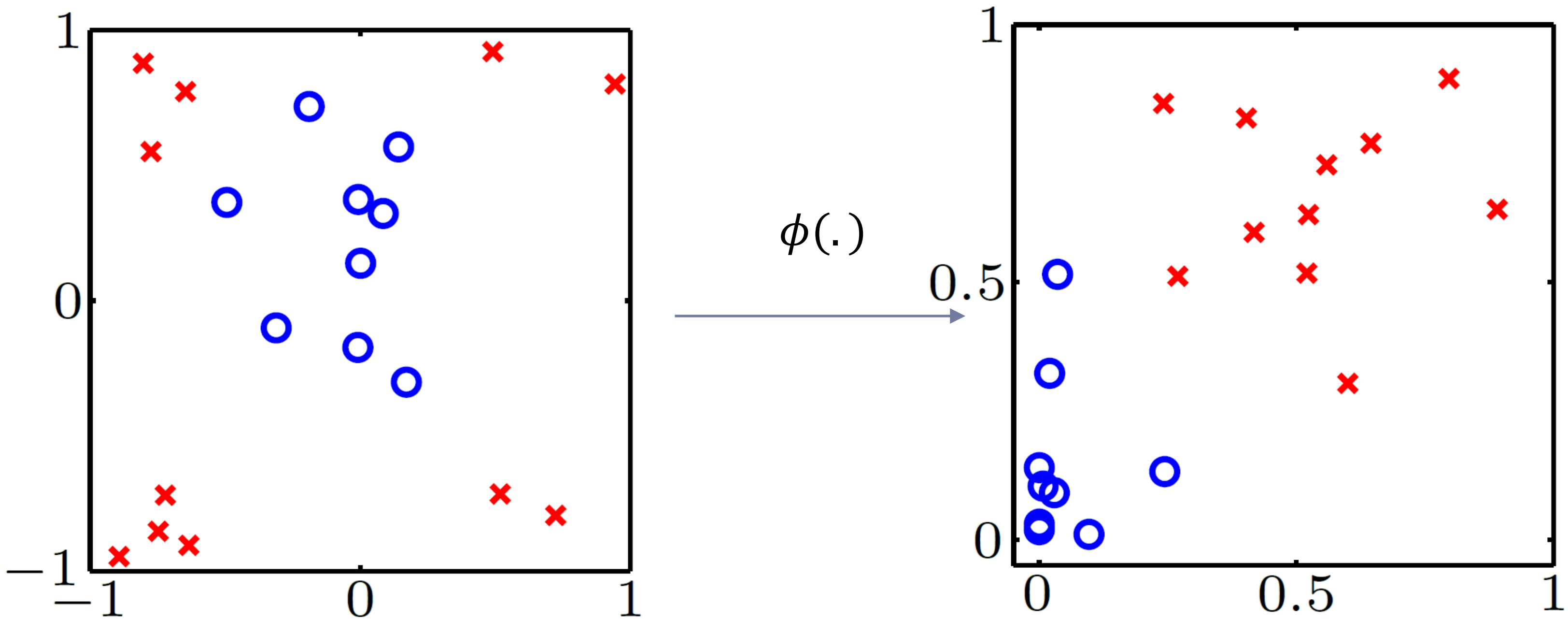
- Only the dot product of each pair of training data appears in the optimization problem
  - An important property that is helpful to extend to non-linear SVM
  - We will talk about it later (kernel-based methods)

In the transformed space

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \phi\left(\boldsymbol{x}^{(n)}\right)^T \phi\left(\boldsymbol{x}^{(m)}\right) \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$$\alpha_n \geq 0 \;\; n = 1, \dots, N$$



Beyond linear separability

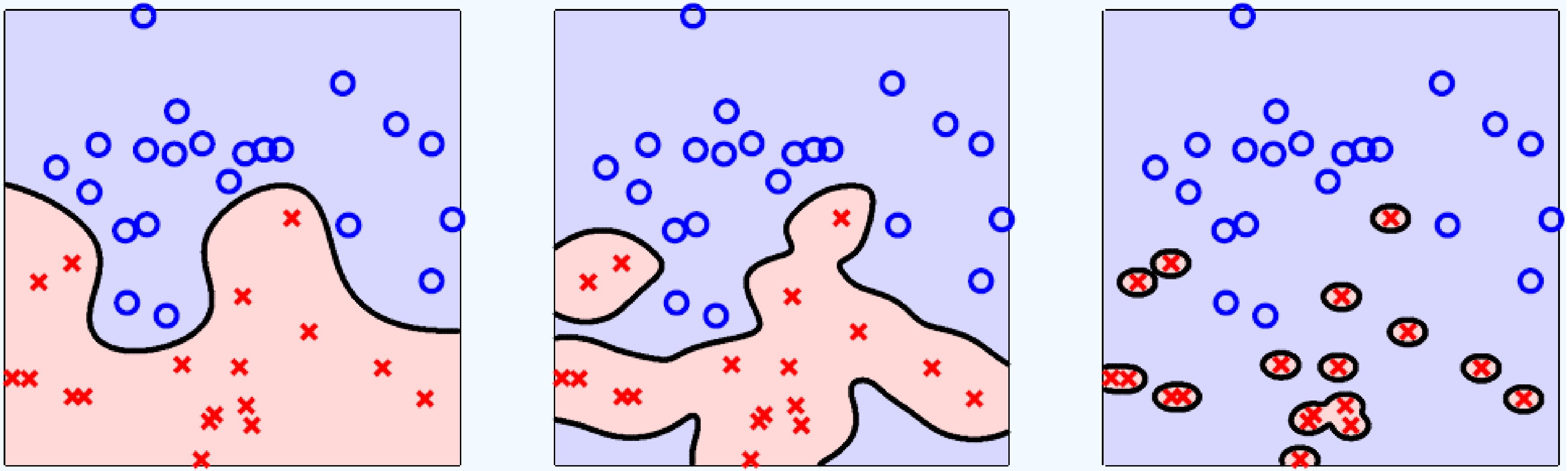
- When training samples are not linearly separable, it has no solution.

- How to extend it to find a solution even though the classes are not exactly linearly separable.

Gaussian kernel

- Example: SVM boundary for a Gaussian kernel
  - Considers a Gaussian function around each data point.
  - $w_0 + \sum_{\alpha_i > 0} \alpha_i y^{(i)} \exp(-\frac{\left\|\boldsymbol{x} - \boldsymbol{x}^{(i)}\right\|^2}{\sigma}) = 0$
  - SVM + Gaussian Kernel can classify any arbitrary training set
    - Training error is zero when $\sigma \to 0$
      - All samples become support vectors (likely overfiting)

Hard margin Example
Gaussian kernel

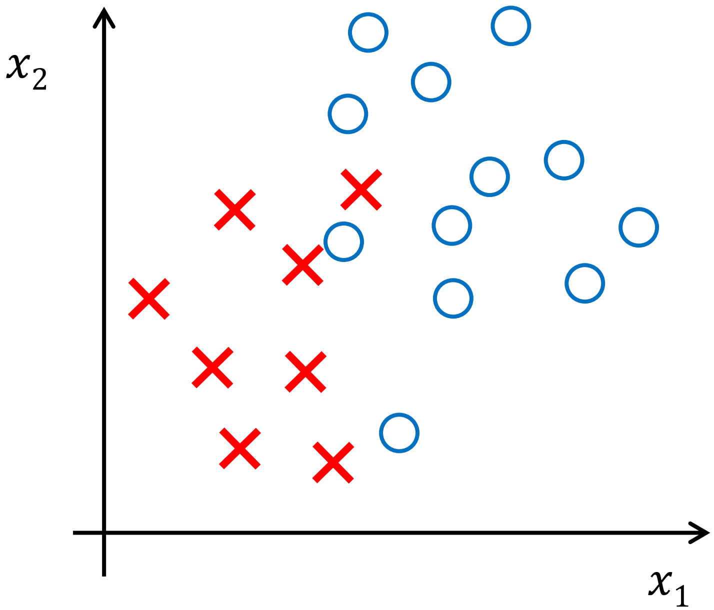


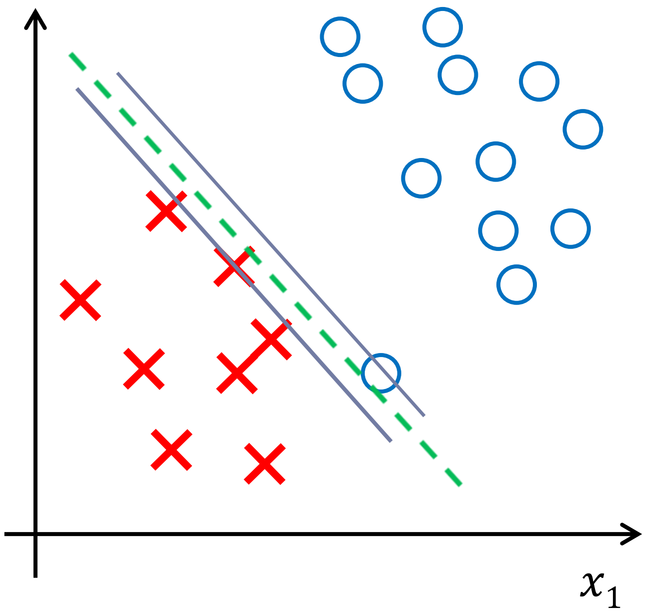
$\exp(-1\|\mathbf{x}-\mathbf{x}'\|^2)$ $\exp(-10\|\mathbf{x}-\mathbf{x}'\|^2)$ $\exp(-100\|\mathbf{x}-\mathbf{x}'\|^2)$

Y. Abu-Mostafa et. Al, 2012

Near linear separability

- How to extend the hard-margin SVM to allow classification error
  - Overlapping classes that can be approximately separated by a linear boundary
  - Noise in the linearly separable classes





Near linear separability: Soft-margin SVM

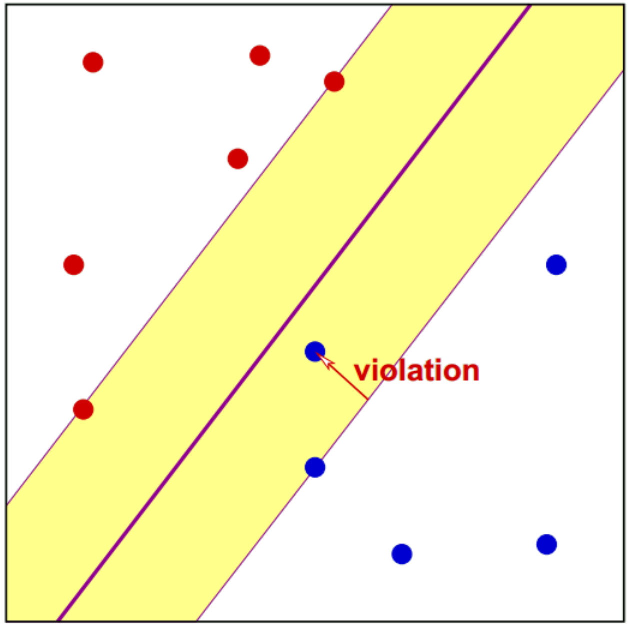
- Minimizing the number of misclassified points?!
  - NP-complete

- Soft margin:
  - Maximizing a margin while trying to minimize the *distance* between misclassified points and their correct margin plane

Error measure

- Margin violation amount ξ_n ($\xi_n \geq 0$):
  - $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n$

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



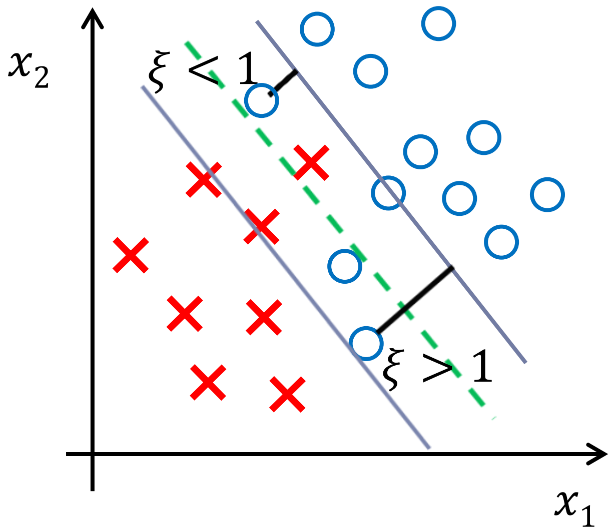
Soft-margin SVM: Optimization problem

- SVM with slack variables: allows samples to fall within the margin, but penalizes them

$$\min_{\boldsymbol{w},w_0,\{\xi_n\}_{n=1}^N} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.} \quad y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n \quad n = 1, \dots, N$$

$$\xi_n \geq 0$$



ξ_n: **slack** variables

$0 < \xi_n < 1$: if $\boldsymbol{x}^{(n)}$ is correctly classified but inside margin

$\xi_n > 1$: if $\boldsymbol{x}^{(n)}$ is misclassifed

Soft-margin SVM

- linear penalty (hinge loss) for a sample if it is misclassified or lied in the margin
  - tries to maintain ξ_n small while maximizing the margin.
  - always finds a solution (as opposed to hard-margin SVM)
  - more robust to the outliers

- Soft margin problem is still a convex QP

Soft-margin SVM: Parameter C

- C is a tradeoff parameter:
  - small C allows margin constraints to be easily ignored
    - large margin
  - large C makes constraints hard to ignore
    - narrow margin

- $C \to \infty$ enforces all constraints: hard margin

- C can be determined using a technique like cross-validation

Soft-margin SVM: Cost function

$$\min_{\boldsymbol{w}, w_0, \{\xi_n\}_{n=1}^{N}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \xi_n$$

$$\text{s.t.} \quad y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n \quad n = 1, \dots, N$$

$$\xi_n \geq 0$$

Lagrange formulation

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$
$$= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \xi_n$$
$$+ \sum_{n=1}^{N} \alpha_n\left(1 - \xi_n - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right) - \sum_{n=1}^{N} \beta_n \xi_n$$

- Minimize w.r.t. $\boldsymbol{w}$, w_0, $\boldsymbol{\xi}$ and maximize w.r.t. $\alpha_n \geq 0$ and $\beta_n \geq 0$

$$\min_{\boldsymbol{w}, w_0, \{\xi_n\}_{n=1}^N} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \xi_n$$
$$\text{s.t.} \quad y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n \quad n = 1, \dots, N$$
$$\xi_n \geq 0$$

Lagrange formulation

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n + \sum_{n=1}^{N}\alpha_n\Big(1 - \xi_n - y^{(n)}\big(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\big)\Big) - \sum_{n=1}^{N}\beta_n\xi_n$$

$$\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0 \Rightarrow \boldsymbol{w} - \sum_{n=1}^{N}\alpha_n y^{(n)}\boldsymbol{x}^{(n)} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{w} = \sum_{n=1}^{N}\alpha_n y^{(n)}\boldsymbol{x}^{(n)}$$

$$\frac{\partial\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial w_{\boldsymbol{0}}} = 0 \Rightarrow \quad -\sum_{n=1}^{N}\alpha_n y^{(n)} = 0$$

$$\frac{\partial\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \xi_n} = 0 \Rightarrow C - \alpha_n - \beta_n = 0$$

Soft-margin SVM: Dual problem

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(m)} \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$$0 \leq \alpha_n \leq C \quad n = 1, \ldots, N$$

- After solving the above quadratic problem, $\boldsymbol{w}$ is find as:

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \, y^{(n)} \boldsymbol{x}^{(n)}$$

Karush-Kuhn-Tucker (KKT) conditions

- Necessary conditions for the solution $[\boldsymbol{w}^*, w_0^*, \xi^*, \boldsymbol{\alpha}^*, \beta^*]$:
  - $\alpha_n^* \geq 0 \quad n = 1, \dots, N$
  - $y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + w_0^*\right) \geq 1 - \xi_n^* \quad n = 1, \dots, N$
  - $\alpha_i^*\left(1 - y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + w_0^*\right) - \xi_n^*\right) = 0 \quad n = 1, \dots, N$
  - $\beta_n^* \geq 0 \quad n = 1, \dots, N$
  - $\xi_n^* \geq 0$
  - $\xi_n^* \beta_n^* = 0$

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{s.t. } g_i(\boldsymbol{x}) \leq 0 \quad i = 1, \dots, m$$

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum \alpha_i \, g_i(\boldsymbol{x})$$

In general, the optimal $\boldsymbol{x}^*, \boldsymbol{\alpha}^*$ satisfies KKT conditions:

$$\alpha_i^* \geq 0 \quad i = 1, \dots, m$$
$$g_i(\boldsymbol{x}^*) \leq 0 \quad i = 1, \dots, m$$
$$\alpha_i^* g_i(\boldsymbol{x}^*) = 0 \quad i = 1, \dots, m$$

Soft-margin SVM: Support vectors

- Support Vectors: $\alpha_n > 0$
  - If $0 < \alpha_n < C$ (**margin** support vector)

    SVs on the margin

$$y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) = 1 \qquad (\xi_n = 0)$$

  - If $\alpha \;=\; C$ (**non-margin** support vector)

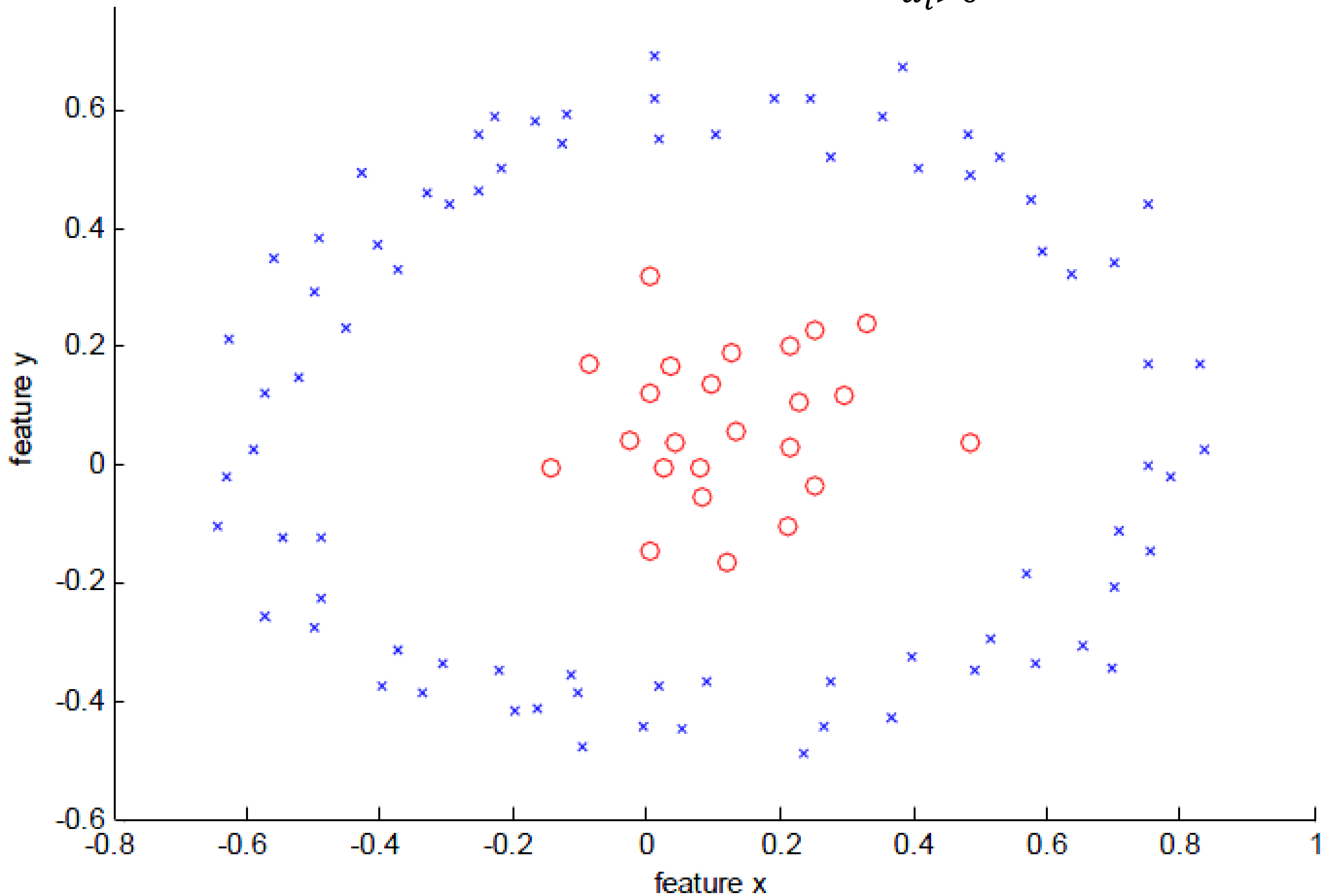
    SVs on or over the margin

$$y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) < 1 \qquad (\xi_n > 0)$$

$$C - \alpha_n - \beta_n = 0$$

SVM Gaussian kernel: Example
Soft margin

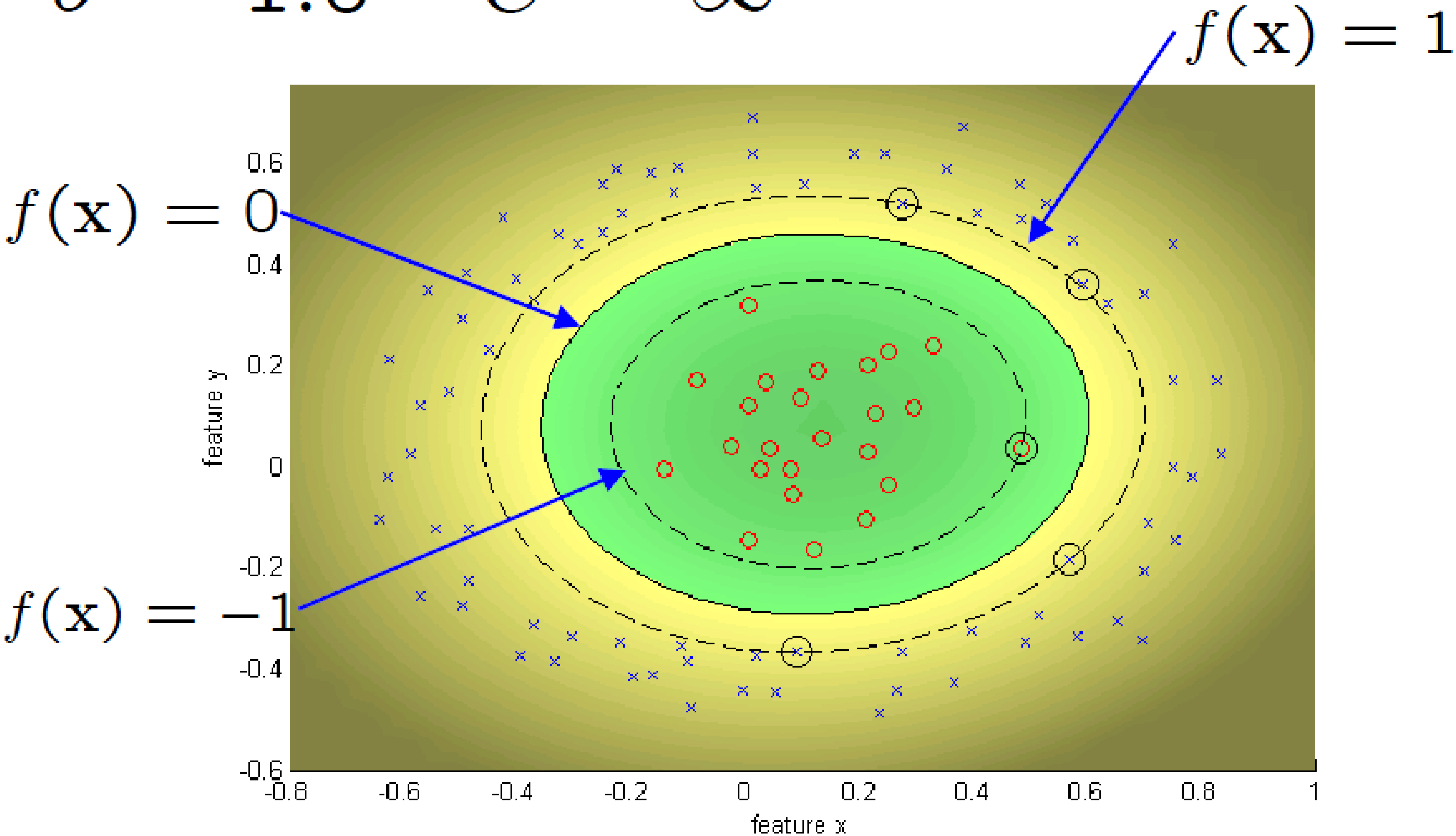
$$f(\boldsymbol{x}) = w_0 + \sum_{\alpha_i > 0} \alpha_i y^{(i)} \exp\left(-\frac{\left\|\boldsymbol{x} - \boldsymbol{x}^{(i)}\right\|^2}{2\sigma^2}\right)$$



This example has been adopted from Zisserman's slides

SVM Gaussian kernel: Example

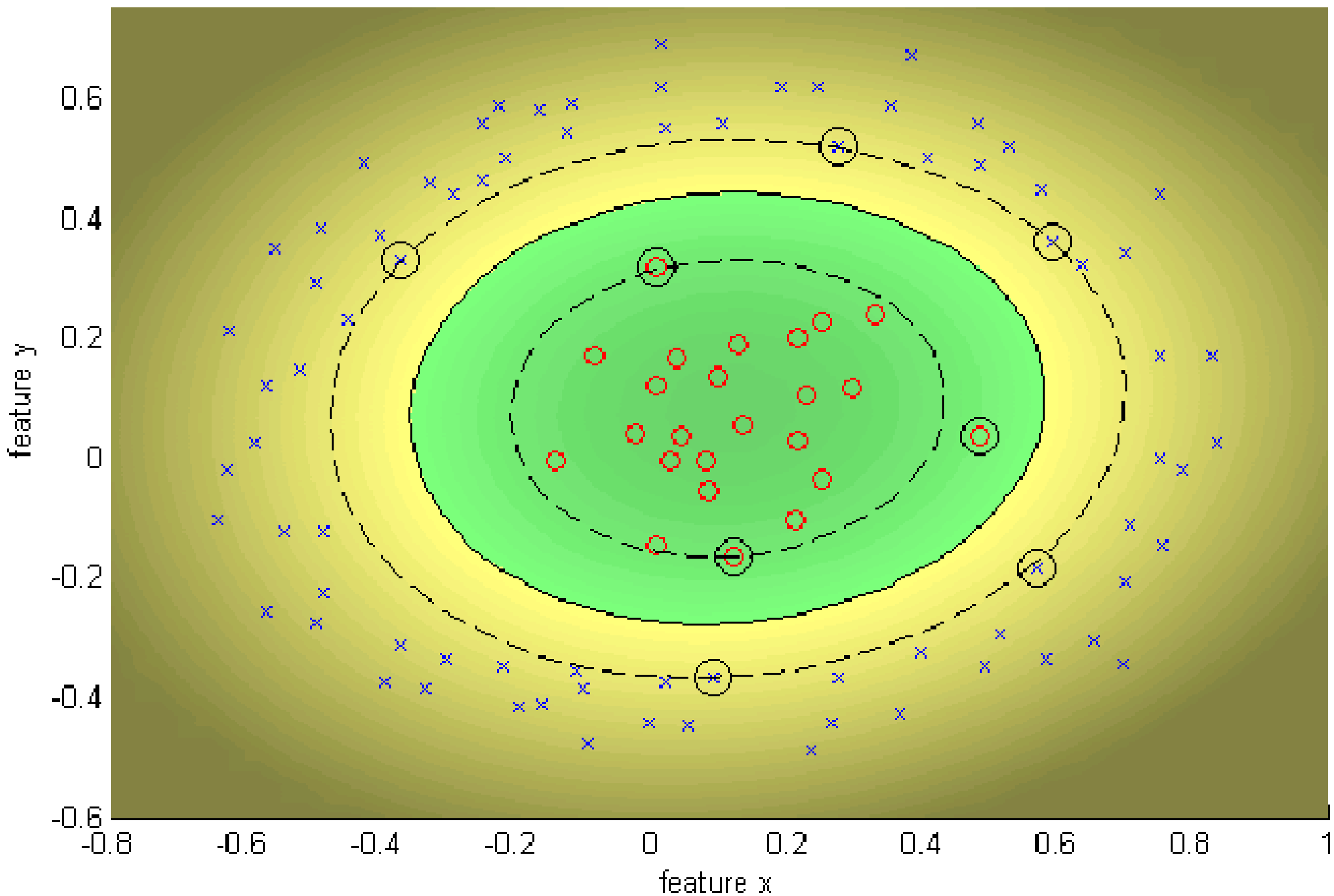
$$\sigma = 1.0 \quad C = \infty$$



This example has been adopted from Zisserman's slides

SVM Gaussian kernel: Example

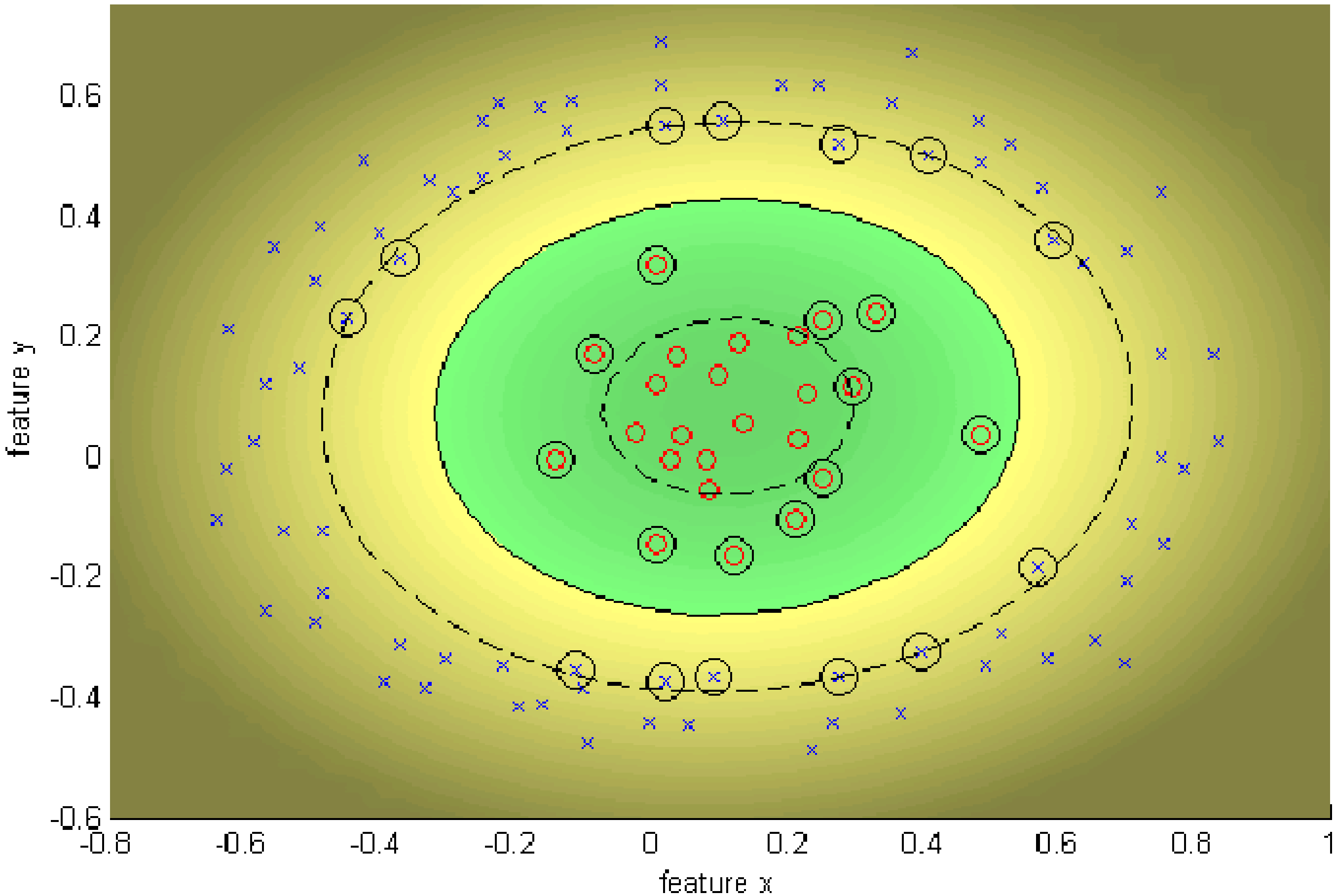
$$\sigma = 1.0 \qquad C = 100$$



This example has been adopted from Zisserman's slides

SVM Gaussian kernel: Example

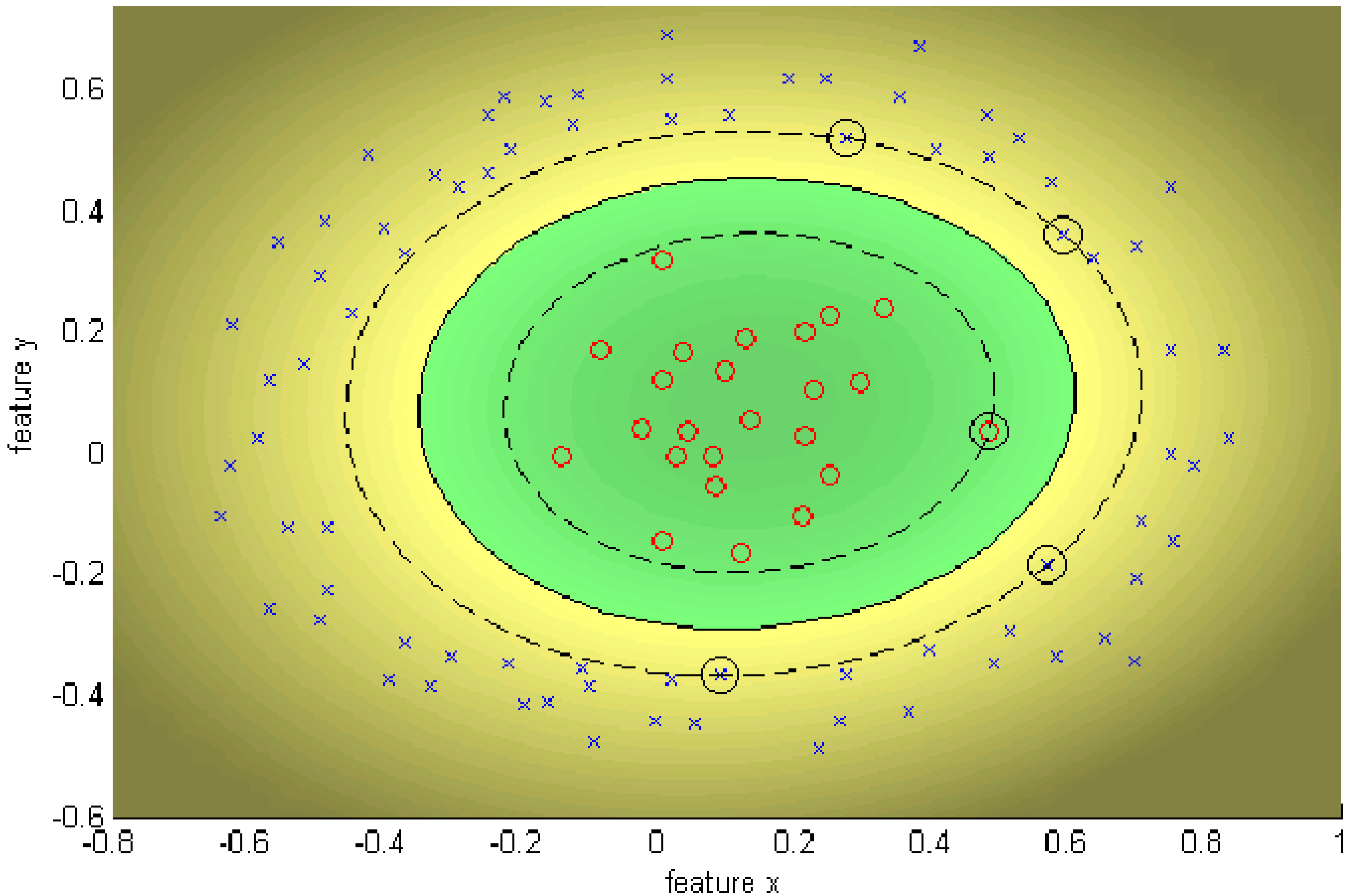
$$\sigma = 1.0 \quad C = 10$$



This example has been adopted from Zisserman's slides

SVM Gaussian kernel: Example

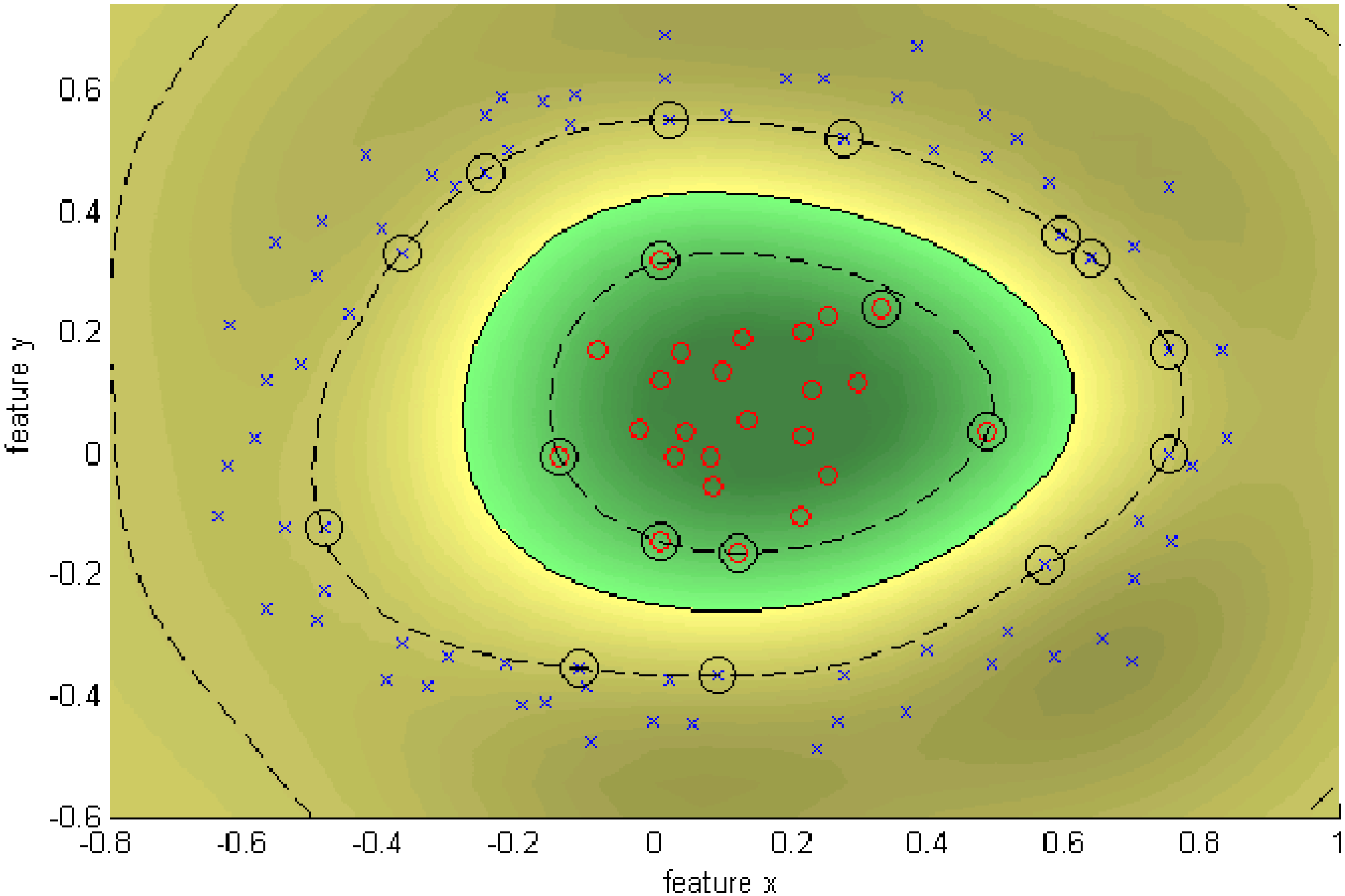
$$\sigma = 1.0 \qquad C = \infty$$



This example has been adopted from Zisserman's slides

SVM Gaussian kernel: Example

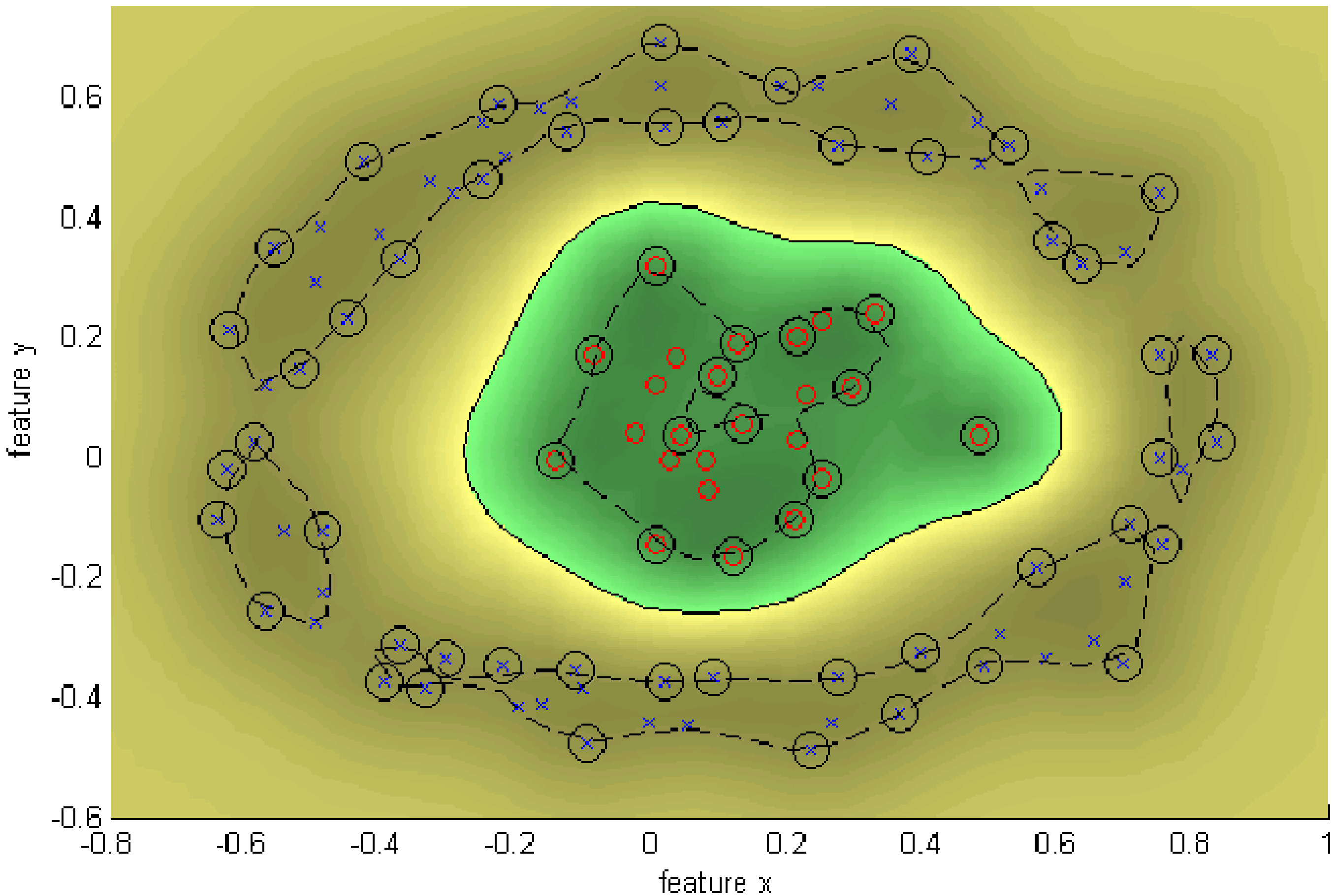
$$\sigma = 0.25 \qquad C = \infty$$



This example has been adopted from Zisserman's slides

SVM Gaussian kernel: Example

$$\sigma = 0.1 \quad C = \infty$$



This example has been adopted from Zisserman's slides

References

- Mahdieh Soleymani, Machine learning course, Sharif univ. of tech.