01.112/50.007 Machine Learning

Lecture 8

Support Vector Machines (Part 2)

Recap

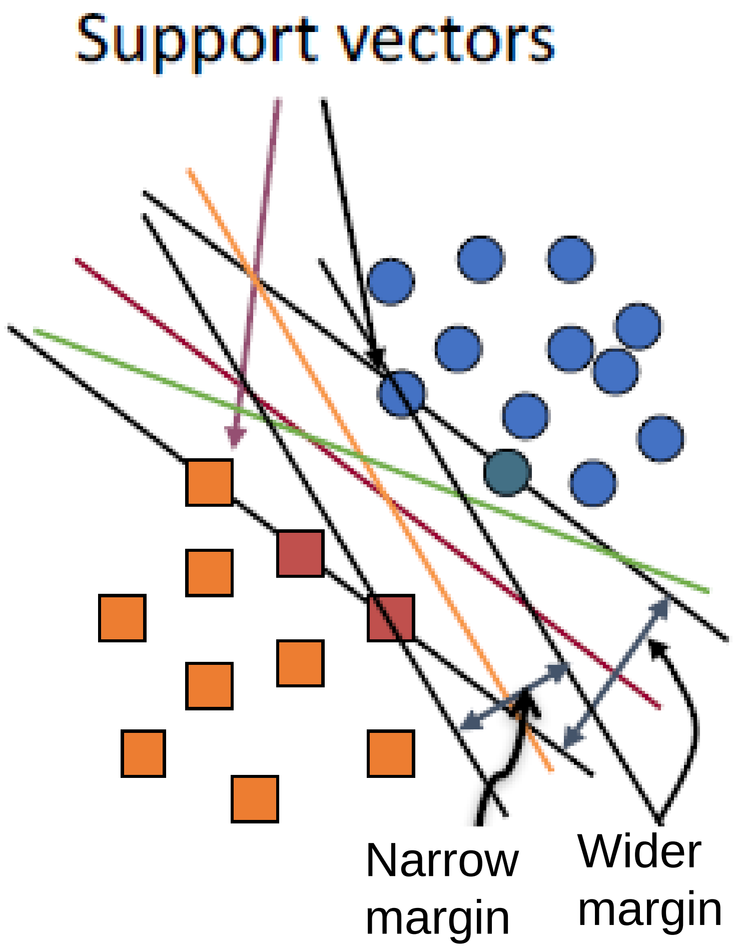
Support Vector Machine (SVM)

- SVMs **maximize the *margin*** around the separating hyperplane. A.k.a. **large margin classifiers.**

- The decision function is fully specified by a subset of training samples, ***the support vectors***.

- Solving SVMs is a ***quadratic programming* problem**.

- Seen by many as the most successful current text classification method*

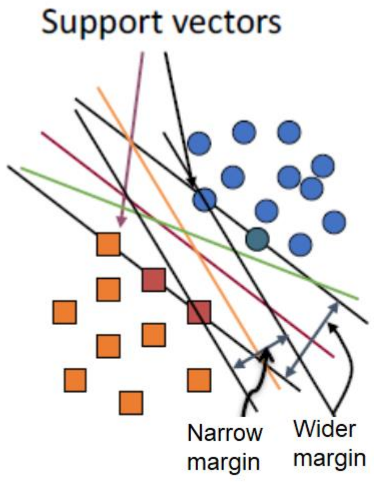


*but other discriminative methods often perform very similarly

Support Vector Machine (SVM)

- **Distance of each point from decision boundary**

$$\gamma^{(t)}(\theta, \theta_0) = \frac{y^{(t)}(\theta \cdot x^{(t)} + \theta_0)}{\|\theta\|}$$



- **Goal: Maximize minimum distance to the boundary**

$$\min_{t=1,\ldots,n} \gamma^{(t)}(\theta, \theta_0)$$

- Formulate the goal as **quadratic programming problem (SVM)**

$$\min \frac{1}{2}\|\theta\|^2 \text{ subject to } y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \geq 1, t = 1, \ldots, n$$

Constrained Optimization

Want to minimize some function $f(x)$, but
there are some *constraints* on the values of x.

Method 1 (Dual Problem)

Solve a *dual optimization problem* where the constraints are nicer, and where it is easier to implement gradient descent.

Method 2 (Exact Solution)

Solve the *Lagrangian* system of equations.

Equality Constraints

Problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_l(x) = 0 \end{array}$$

Lagrangian.

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Example.

$$\begin{array}{ll} \text{minimize} & f(x) = n_1 \log x_1 + \cdots + n_d \log x_d \\ \text{subject to} & h(x) = x_1 + \cdots + x_d - 1 = 0 \end{array}$$

$$L(x, \lambda) = n_1 \log x_1 + \cdots + n_d \log x_d + \lambda(x_1 + \cdots + x_d - 1)$$

Two-Player Game

$$L(x,\lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Rules.

- You get to choose the value of x.
  Your goal is to minimize $L(x,\lambda)$.
- Your adversary gets to choose the value of λ.
  His goal is to maximize $L(x,\lambda)$.

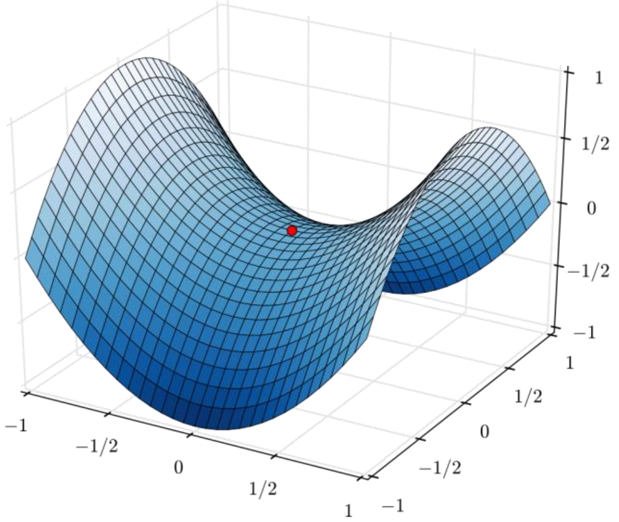
Primal Game

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Primal Game. You go first.

Your Strategy.

- Ensure that $h_1(x) = 0, \ldots, h_l(x) = 0$.
- Find x that minimizes $f(x)$.



Final Score. $p^* = \min_x \max_\lambda L(x, \lambda)$

The optimal x^*, λ^* are *saddle points* of $L(x, \lambda)$.

Dual Game

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Dual Game. You go second.

Adversary's Strategy.

- For each λ, compute $\ell(\lambda) = \min_x L(x, \lambda)$
- Find λ that maximizes $\ell(\lambda)$.

Final Score. $d^* = \max_\lambda \min_x L(x, \lambda)$

Max-Min Inequality

Primal. $p^* = \min_x \max_\lambda L(x, \lambda)$

Dual. $d^* = \max_\lambda \min_x L(x, \lambda)$

"you do better if you have the last say"

Weak Duality

$$p^* = \min_x \max_\lambda L(x, \lambda)$$
$$\geq \max_\lambda \min_x L(x, \lambda) = d^*$$

If $p^* = d^*$, we can solve the primal by solving the dual.

Strong duality

Max-Min Inequality

Example.

	$x = 1$	$x = 2$
$\lambda = 1$	1	4
$\lambda = 2$	3	2

Primal. $$p^* = \min_x \max_\lambda L(x, \lambda) = 3$$

Dual. $$d^* = \max_\lambda \min_x L(x, \lambda) = 2$$

Exact Solution

Problem.

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && h_1(x) = 0, \dots, h_l(x) = 0 \end{aligned}$$

Lagrange multipliers.

1. Write down the Lagrangian.

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

2. Solve for critical points x, λ.

$$\nabla_x L(x, \lambda) = 0, \quad h_1(x) = 0, \dots, h_l(x) = 0$$

3. Pick critical point which gives global minimum.

Example

$$\begin{array}{ll} \text{minimize} & f(x) = n_1 \log x_1 + \cdots + n_d \log x_d \\ \text{subject to} & h(x) = x_1 + \cdots + x_d - 1 = 0 \end{array}$$

Lagrangian

$$L(x, \lambda) = n_1 \log x_1 + \cdots + n_d \log x_d + \lambda(x_1 + \cdots + x_d - 1)$$

Critical points

$$0 = n_i / x_i + \lambda \qquad\qquad x_i = n_i / (-\lambda)$$

$$0 = x_1 + \cdots + x_d - 1 \qquad\qquad (-\lambda) = n_1 + \cdots + n_d$$

Inequality Constraints (Primal-Dual)

Primal Problem.

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_1(x) \le 0, \dots, g_m(x) \le 0 \end{aligned}$$

Lagrangian.

$$L(x, \alpha) = f(x) + \alpha_1 g_1(x) + \cdots + \alpha_m g_m(x)$$

Dual Problem.

$$\begin{aligned} &\text{maximize} \quad \ell(\alpha) \qquad \text{where } \ell(\alpha) = \min_{x \in \mathbb{R}^d} L(x, \alpha) \\ &\text{subject to} \quad \alpha_1 \ge 0, \dots, \alpha_m \ge 0 \end{aligned}$$

Box constraints are
easier to work with!

Inequality Constraints (Exact Solution)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_1(x) \leq 0, \dots, g_m(x) \leq 0 \end{array}$$

Lagrangian.

$$L(x, \alpha) = f(x) + \alpha_1 g_1(x) + \cdots + \alpha_m g_m(x)$$

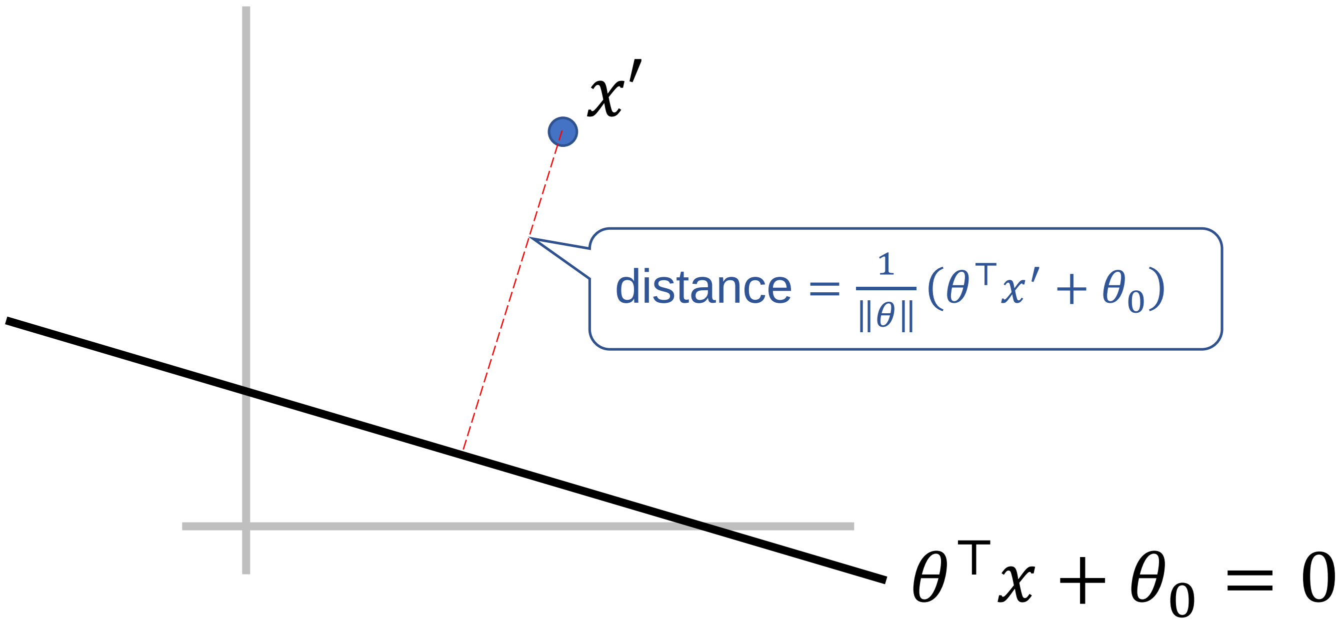
Solve for x, α satisfying

1. $\nabla_x L(x, \alpha) = 0$
2. $g_1(x) \leq 0, \dots, g_m(x) \leq 0$
3. $\alpha_1 \geq 0, \dots, \alpha_m \geq 0$
4. $\alpha_1 g_1(x) = 0, \dots, \alpha_m g_m(x) = 0$

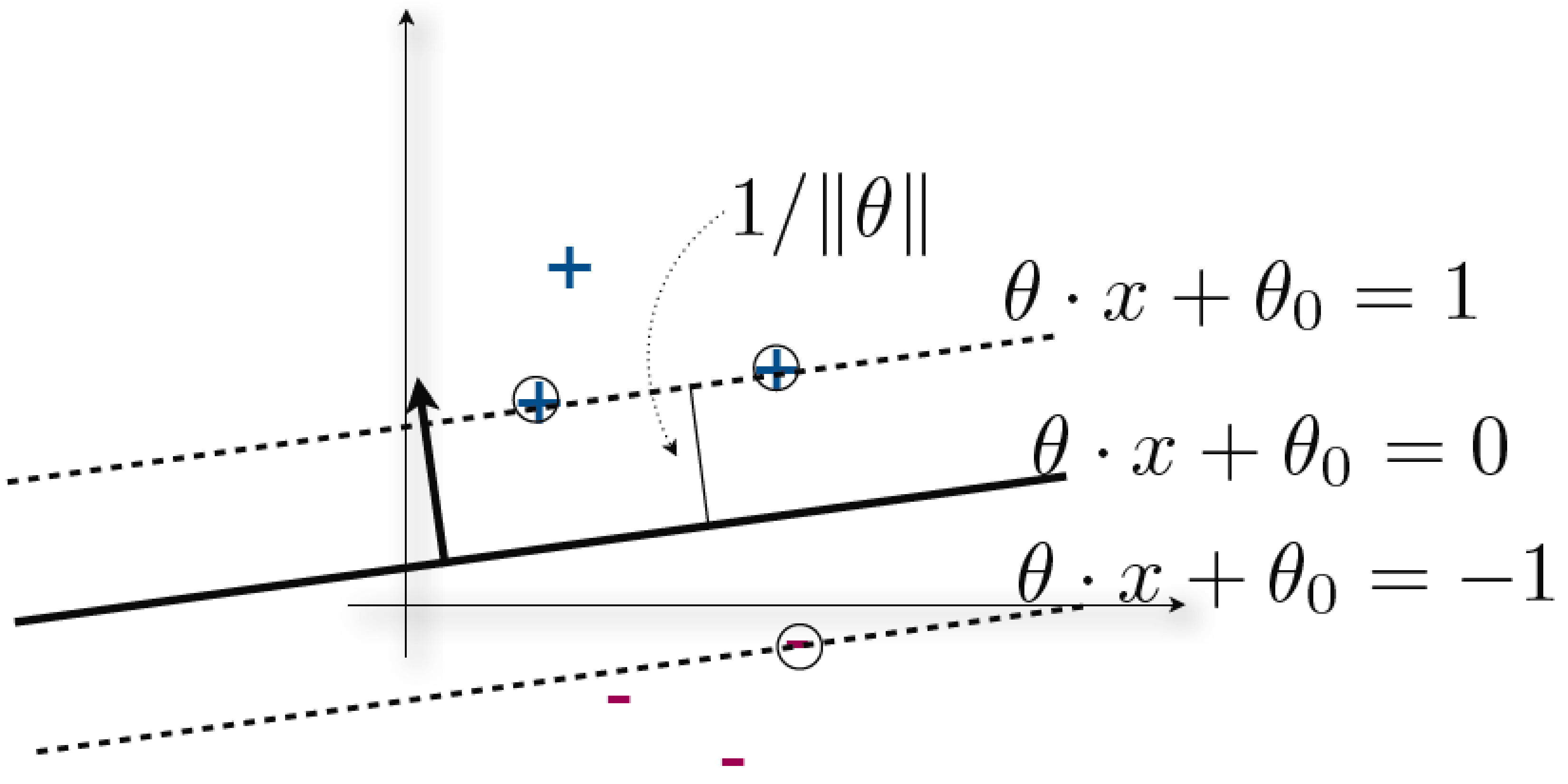
Complementary Slackness

SVM: Maximum Margins

Computing the margin



Computing the margin



Maximum Margin

Our goal is to

maximize $1/\|\theta\|$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Or equivalently,

minimize $\frac{1}{2}\|\theta\|^2$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Lagrangian

Primal. minimize $\frac{1}{2}\|\theta\|^2$

subject to $y(\theta^\top x) \geq 1$ for all data (x, y)

Lagrangian. $L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{(x,y)} \alpha_{x,y} \left(1 - y(\theta^\top x)\right)$

To find $\ell(\alpha) = \min_{\theta} L(\theta, \alpha)$, we solve

$$0 = \nabla_\theta L(\theta, \alpha) = \theta - \sum_{(x,y)} \alpha_{x,y}\, yx$$

to get $\theta = \sum_{(x,y)} \alpha_{x,y}\, yx$. Substituting into $L(\theta, \alpha)$ gives

$$\ell(\alpha) = \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2}\sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, yy'(x^\top x').$$

Primal-Dual

It can be shown that the primal and dual problems are equivalent (*strong duality*).

Primal.

minimize $\frac{1}{2}||\theta||^2$

subject to $y(\theta^\top x) \geq 1$ for all data (x, y)

Dual.

maximize $\sum_{(x,y)} \alpha_{x,y} - \frac{1}{2}\sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'} y y' (x^\top x')$

subject to $\alpha_{x,y} \geq 0$ for all (x, y)

After solving the dual to get the optimal $\alpha_{x,y}$'s,
we obtain the optimal θ using $\theta = \sum_{(x,y)} \alpha_{x,y}\, yx$.

Support Vectors

Complementary Slackness.

$$\hat{\alpha}_{x,y} > 0: \quad y\left(\hat{\theta}^\top x\right) = 1$$

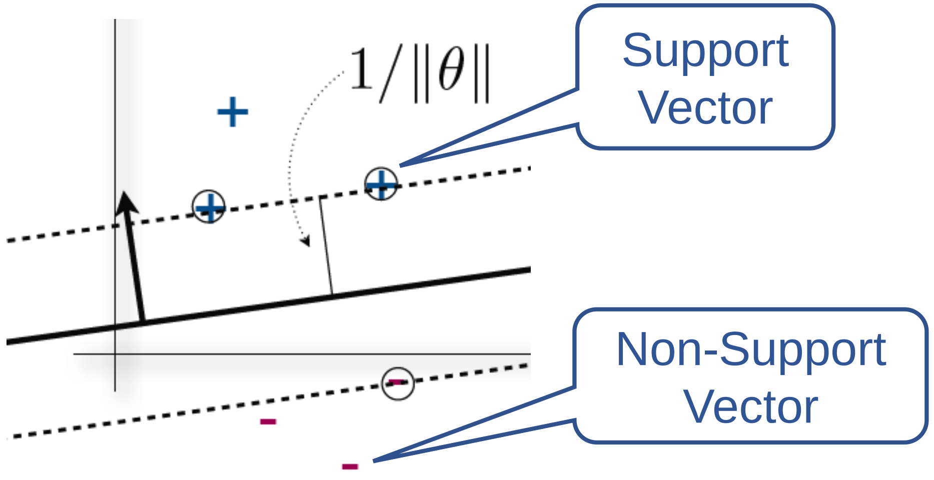
$$\hat{\alpha}_{x,y} = 0: \quad y\left(\hat{\theta}^\top x\right) > 1$$

Support Vectors

Non-Support Vectors

Sparsity

Since very few data points are support vectors, most of the $\hat{\alpha}_{x,y}$ will be zero.



Extensions

SVM with offset

Primal.

minimize $\frac{1}{2}||\theta||^2$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Dual.

maximize $\sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, y y' (x^\top x')$

subject to $\alpha_{x,y} \geq 0$ for all (x, y)

$\sum_{(x,y)} \alpha_{x,y}\, y = 0$

SVM with offset

Dual.

maximize $\quad \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'} y y' (x^\top x')$

subject to $\quad \alpha_{x,y} \geq 0$ for all (x, y)

$\sum_{(x,y)} \alpha_{x,y}\, y = 0$

Parameters. $\quad \hat{\theta} = \sum_{(x,y)} \alpha_{x,y}\, y x$

$\hat{\theta}_0 = y - \hat{\theta}^\top x \qquad$ where (x, y) is a support vector

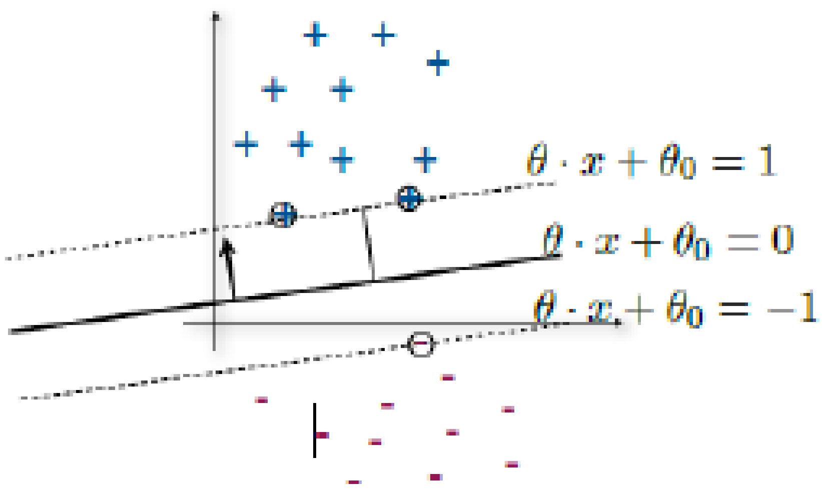
Derivation for $\hat{\theta}_0$

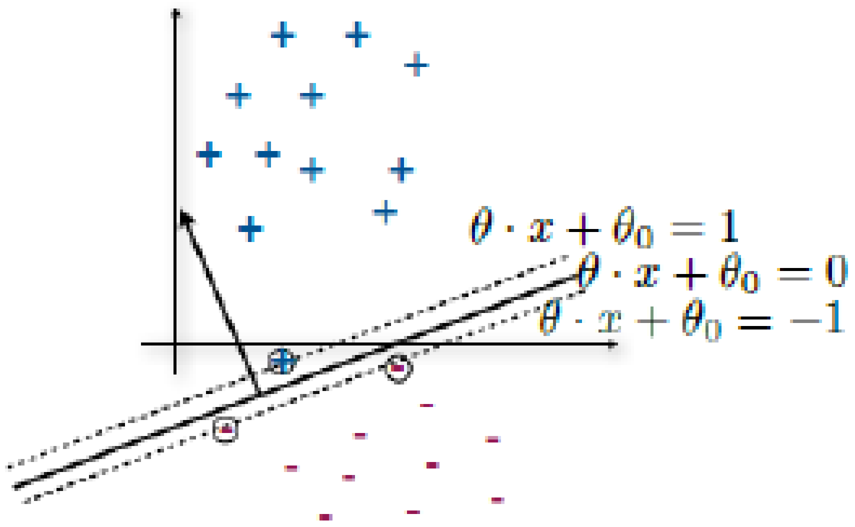
$$y^{(t)}(\hat{\theta} \cdot x^{(t)} + \hat{\theta}_0) = y^{(t)}(\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')}(x^{(t')} \cdot x^{(t)}) + \theta_0) = 1$$

$$\hat{\theta}_0 = y^{(t)} - (\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')}(x^{(t')} \cdot x^{(t)}))$$

SVM with errors

- **Effect of errors in labelling training examples:**





SVM with errors

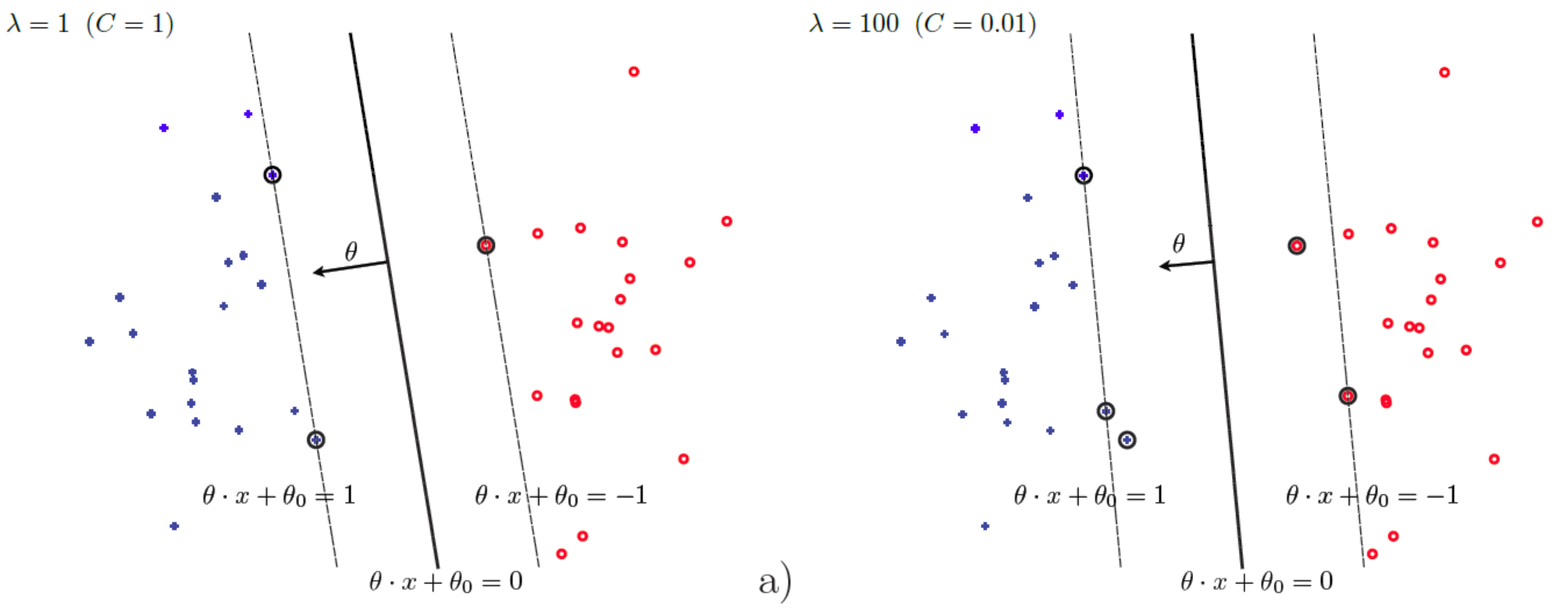
- **Allowing misclassified points, yet maximize the margin**
- Convert **hard constraints to soft constraints**
- Add a **slack variable** ξ and a **regularization parameter** λ
- Slack variables allow constraints to be violated for a cost.
- Regularization parameter balances favouring between increasing the margin and misclassifications.

Primal.

$$\begin{array}{lll} \text{minimize} & \frac{\lambda}{2}||\theta||^2 + \frac{1}{n}\sum_{(x,y)} \xi_{x,y} & \\ \text{subject to} & y(\theta^\top x + \theta_0) \geq 1 - \xi_{x,y} & \text{for all data } (x,y) \\ & \xi_{x,y} \geq 0 & \text{for all data } (x,y) \end{array}$$

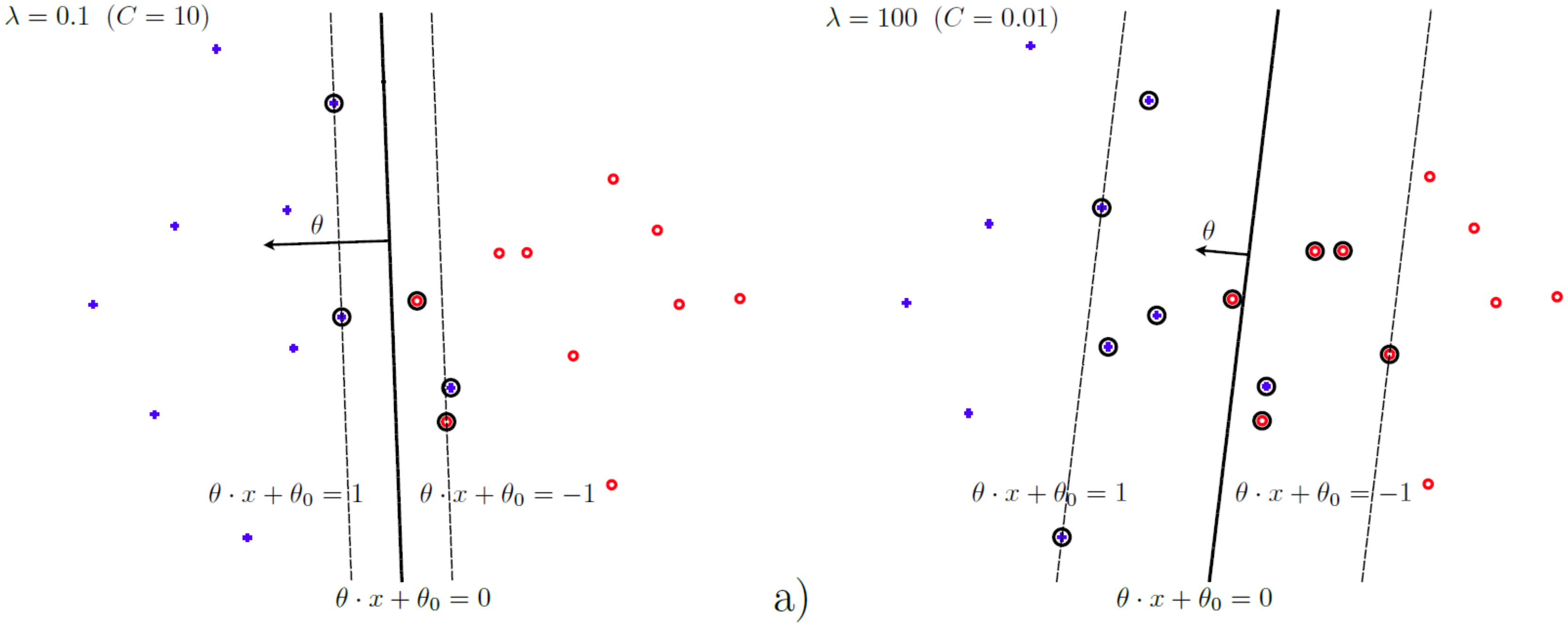
SVM with errors

- **Linearly Separable**



SVM with errors

- **Not Linearly Separable**



SVM with errors

Primal.

minimize $\frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{(x,y)} \xi_{x,y}$

subject to $y(\theta^\top x + \theta_0) \geq 1 - \xi_{x,y}$ for all data (x, y)

$\xi_{x,y} \geq 0$ for all data (x, y)

Equivalent Primal.

minimize $\frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{(x,y)} \mathrm{Loss}_{\mathrm{H}}\big(y(\theta^\top x + \theta_0)\big)$

Hinge-loss classifier with regularization!

SVM with errors

Primal.

minimize $\frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{(x,y)} \xi_{x,y}$

subject to $y(\theta^\top x + \theta_0) \geq 1 - \xi_{x,y}$ for all data (x, y)

$\xi_{x,y} \geq 0$ for all data (x, y)

Dual.

maximize $\sum_{(x,y)} \alpha_{x,y} - \frac{1}{2}\sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, y y' (x^\top x')$

subject to $1/\lambda \geq \alpha_{x,y} \geq 0$ for all (x, y)

$\sum_{(x,y)} \alpha_{x,y}\, y = 0$

There are many efficient solvers for quadratic problems with box constraints.

Putting limits on what the adversary can do.

SVM with errors

- **Complementary slackness**

$$\hat{\alpha}_t = 0 \Rightarrow y^{(t)}(\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')}(x^{(t')} \cdot x^{(t)}) + \hat{\theta}_0) \geq 1 \qquad \text{(non-support vectors)}$$

$$\hat{\alpha}_t \in (0, 1/\lambda) \Rightarrow y^{(t)}(\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')}(x^{(t')} \cdot x^{(t)}) + \hat{\theta}_0) = 1 \qquad \text{(SVs, on the margin)}$$

$$\hat{\alpha}_t = 1/\lambda \Rightarrow y^{(t)}(\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')}(x^{(t')} \cdot x^{(t)}) + \hat{\theta}_0) \leq 1 \qquad \text{(SVs, margin violations)}$$

Summary

- **Lagrange Multipliers**
  - Lagrangian
  - Primal-Dual Problems
  - Inequality Constraints
  - Complementary Slackness
- **Support Vector Machines**
  - Maximum Margins
  - Dual Problem
  - Support Vectors
- **Regularization**
  - Slack Variables
  - Regularized Hinge Loss
  - Bounded Multipliers

Intended Learning Outcomes

Extensions

- Describe the dual problem for the SVM with offset.
- Describe the primal problem for SVM with slack variables. Show that the primal is equivalent to regularized hinge loss. Explain how the regularizing parameter λ affects the margins. Describe the dual problem in terms of box constraints.