



# Machine Learning 1

Lecture 11.3 - Kernel Methods  
Support Vector Machines - Maximum Margin  
Classifier

*Erik Bekkers*

*(Bishop 7.1.0)*



# Support vector machines

- ▶ Kernel method with sparse solutions:
  - ▶ prediction for new inputs depend only on kernel function evaluated at a **subset** of the training points

- ▶ Applications:

- ▶ Classification

- ▶ Regression

- ▶ novelty detection/anomaly detection

- ▶ Convex optimization problem, any local solution is at global optimum!

- ▶ No good probabilistic interpretation

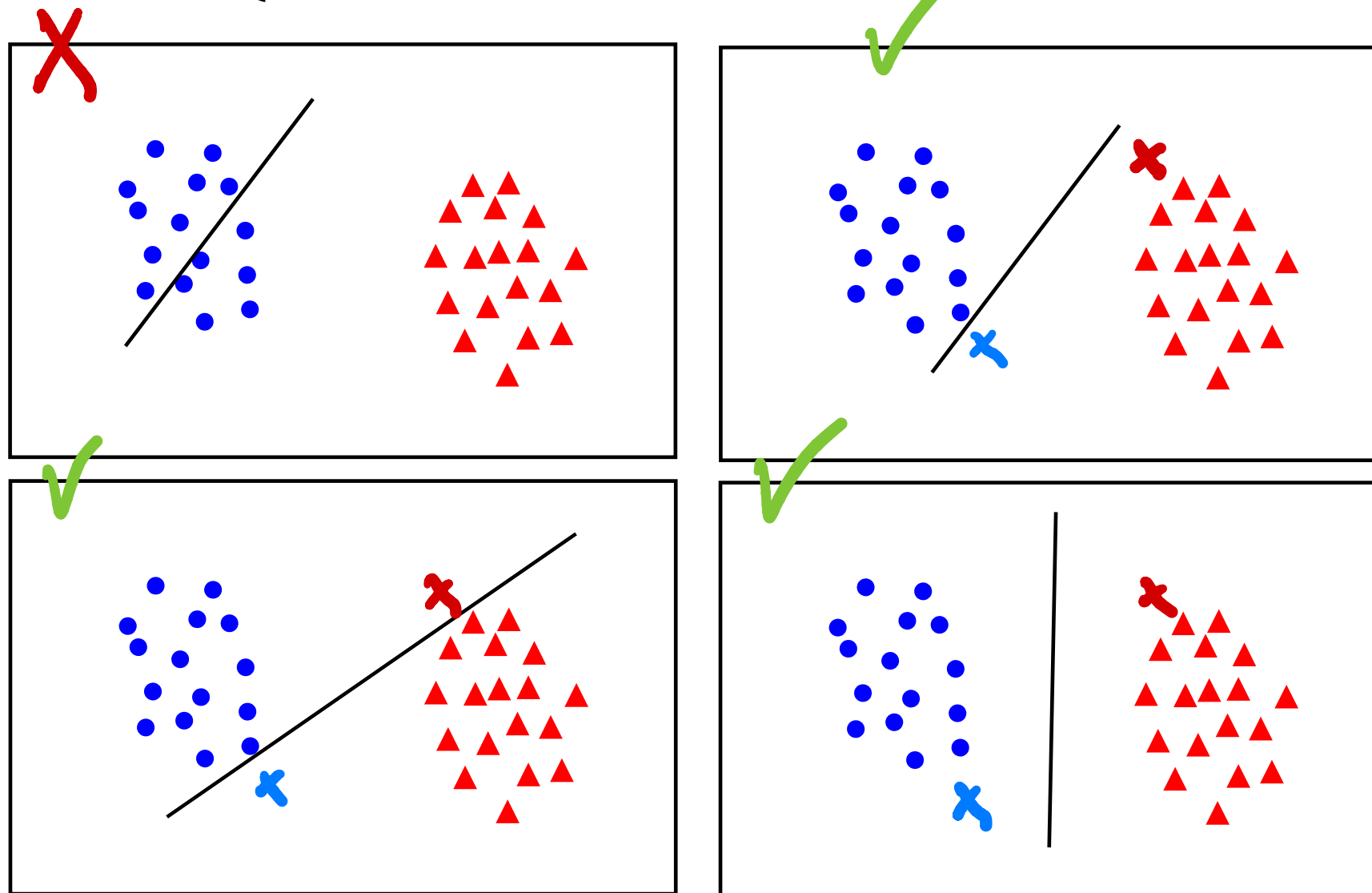
- ▶ Today: SVM for binary classification -> maximum margin classifier!

Primal viewpoint  
Dual viewpoint

$$y(\underline{x}, \underline{w}) = \underline{w}^T \underline{\phi}(\underline{x})$$
$$y(\underline{x}, \underline{a}) = \sum_n a_n k(\underline{x}, \underline{x}_n)$$

# Linearly separable dataset

- ▶ Linear classifier:  $y(\mathbf{x}_n) = \mathbf{w}^t \mathbf{x}_n + b$
- ▶ Classification: 
$$\begin{cases} t_n = +1 & \text{if } y(\mathbf{x}_n) \geq 0 \\ t_n = -1 & \text{if } y(\mathbf{x}_n) < 0 \end{cases}$$



- ▶ Maximum Margin: most stable under perturbations of the input

# Linearly Separable Dataset

- ▶ If  $\mathbf{x}'$  lies on decision boundary:  $y(\mathbf{x}') = \mathbf{w}^T \mathbf{x}' + b = 0$
- ▶ Recall: distance from  $\mathbf{x}$  to decision boundary is

$$r = \frac{|y(\mathbf{x}_n)|}{\|\mathbf{w}\|} = \frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|}$$

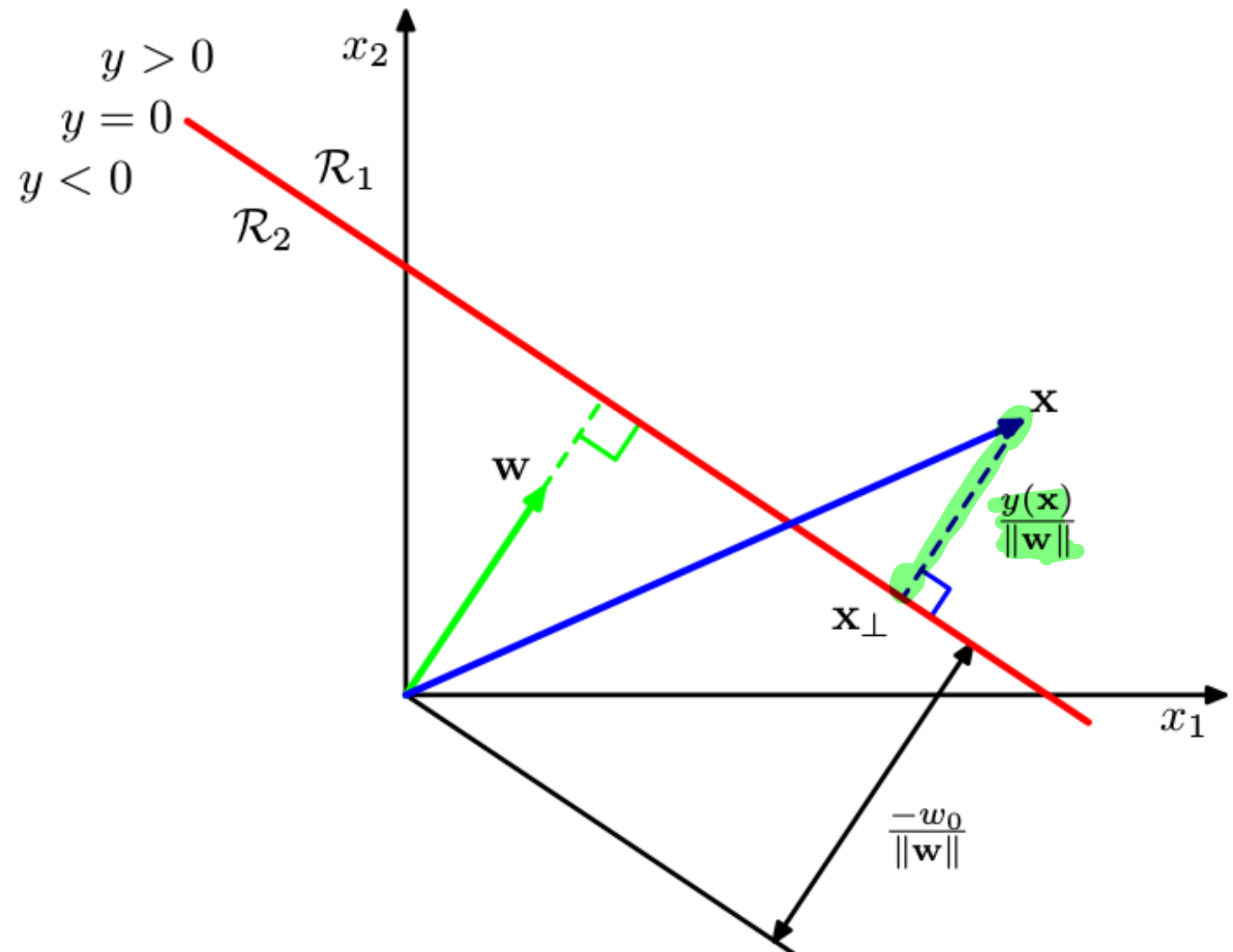
- ▶ For correct classification:

$$y(\mathbf{x}_n) \geq 0 \text{ if } t_n = +1$$

$$y(\mathbf{x}_n) < 0 \text{ if } t_n = -1$$

- ▶ So for all  $n = 1, \dots, N$

$$t_n y(\mathbf{x}_n) \geq 0$$



# Maximum Margin Classifier

- Margin: perpendicular distance from decision boundary to closest point  $\mathbf{x}_n$

- For all data points distance to decision boundary is

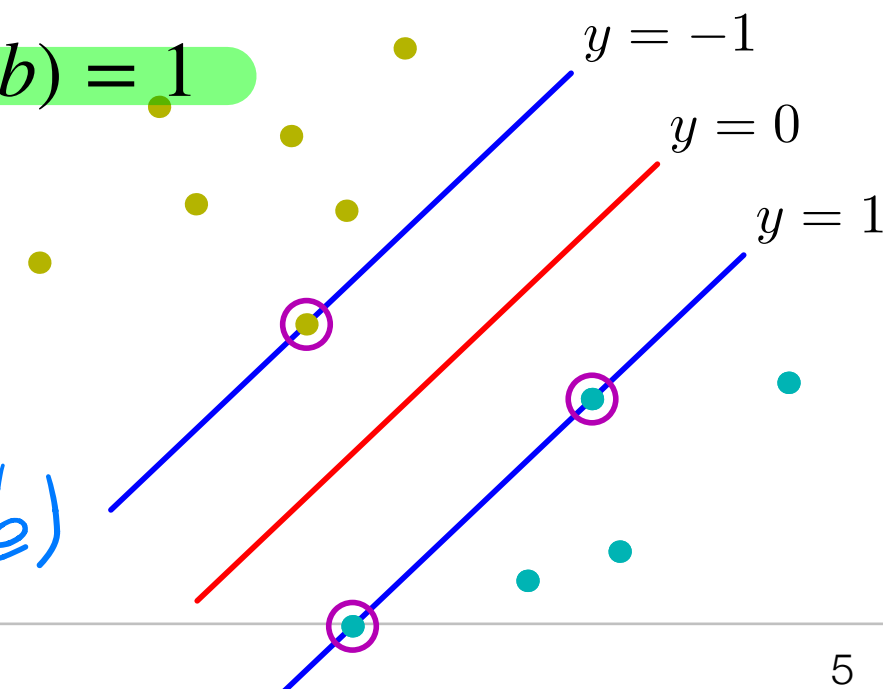
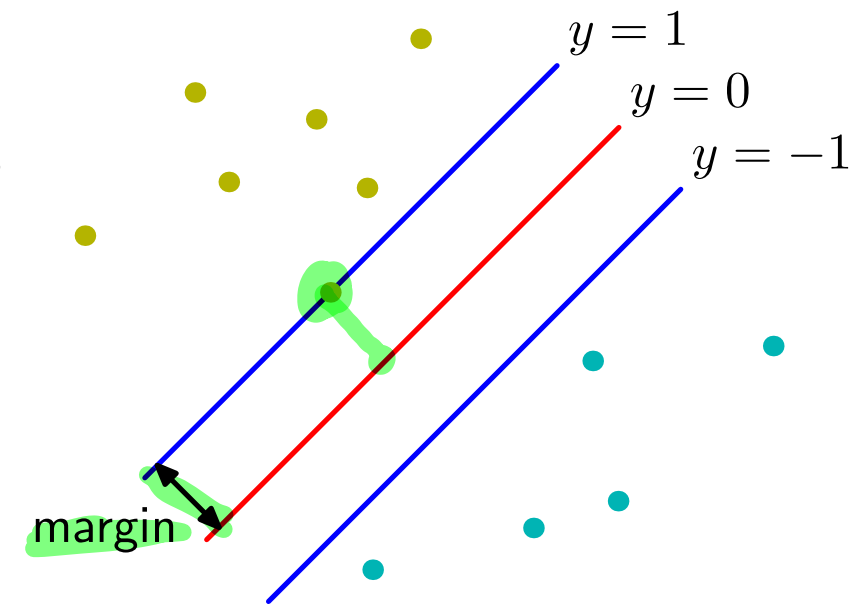
$$r = \frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T \mathbf{x}_n + b)}{\|\mathbf{w}\|}$$

- Margin:  $\min_n \frac{t_n(\mathbf{w}^T \mathbf{x}_n + b)}{\|\mathbf{w}\|} = \min_n \frac{t_n(\kappa \mathbf{w}^T \mathbf{x}_n + \kappa b)}{\|\kappa \mathbf{w}\|}$

- For point closest to decision boundary  $t_n(\mathbf{w}^T \mathbf{x}_n + b) = 1$

- For all data points:  $t_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$

- Maximum margin classifier:  $\max_{\underline{w}, \underline{b}} \text{margin}(\underline{w}, \underline{b})$



# Maximum Margin Classifier

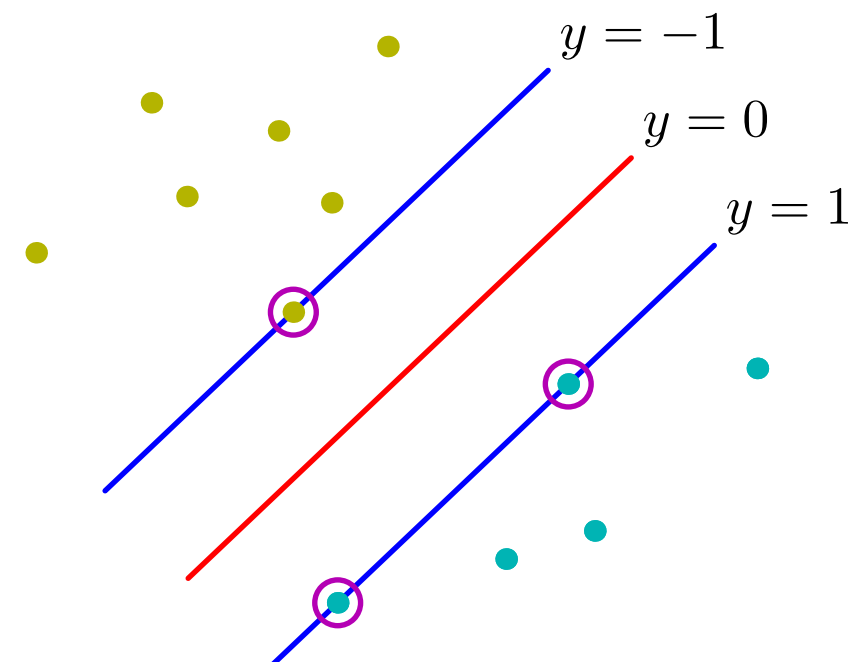
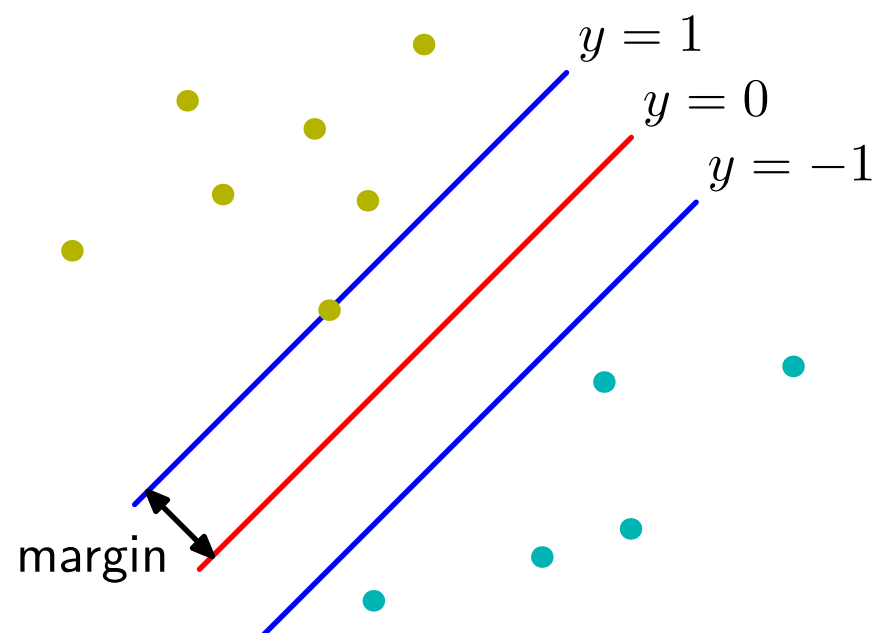
- For all data points distance to decision boundary is

$$r = \frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T \mathbf{x}_n + b)}{\|\mathbf{w}\|}$$

- For point closest to decision boundary  $t_n(\mathbf{w}^T \mathbf{x}_n + b) = 1$

- For all data points:  $t_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$

- Size of the margin:  $\frac{1}{\|\mathbf{w}\|}$  ← maximize



# Maximum Margin Classifier

- ▶ Size of the margin:  $\frac{1}{\|\mathbf{w}\|}$
- ▶ For all data points:  $t_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$
- ▶ **Maximizing the margin:**

$$\arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \text{ subject to } N \text{ constraints } t_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$$

- ▶ Quadratic programming problem!

