Support vector machine

Kookjin Lee

(kookjin.Lee@asu.edu)

Going back to logistic regression

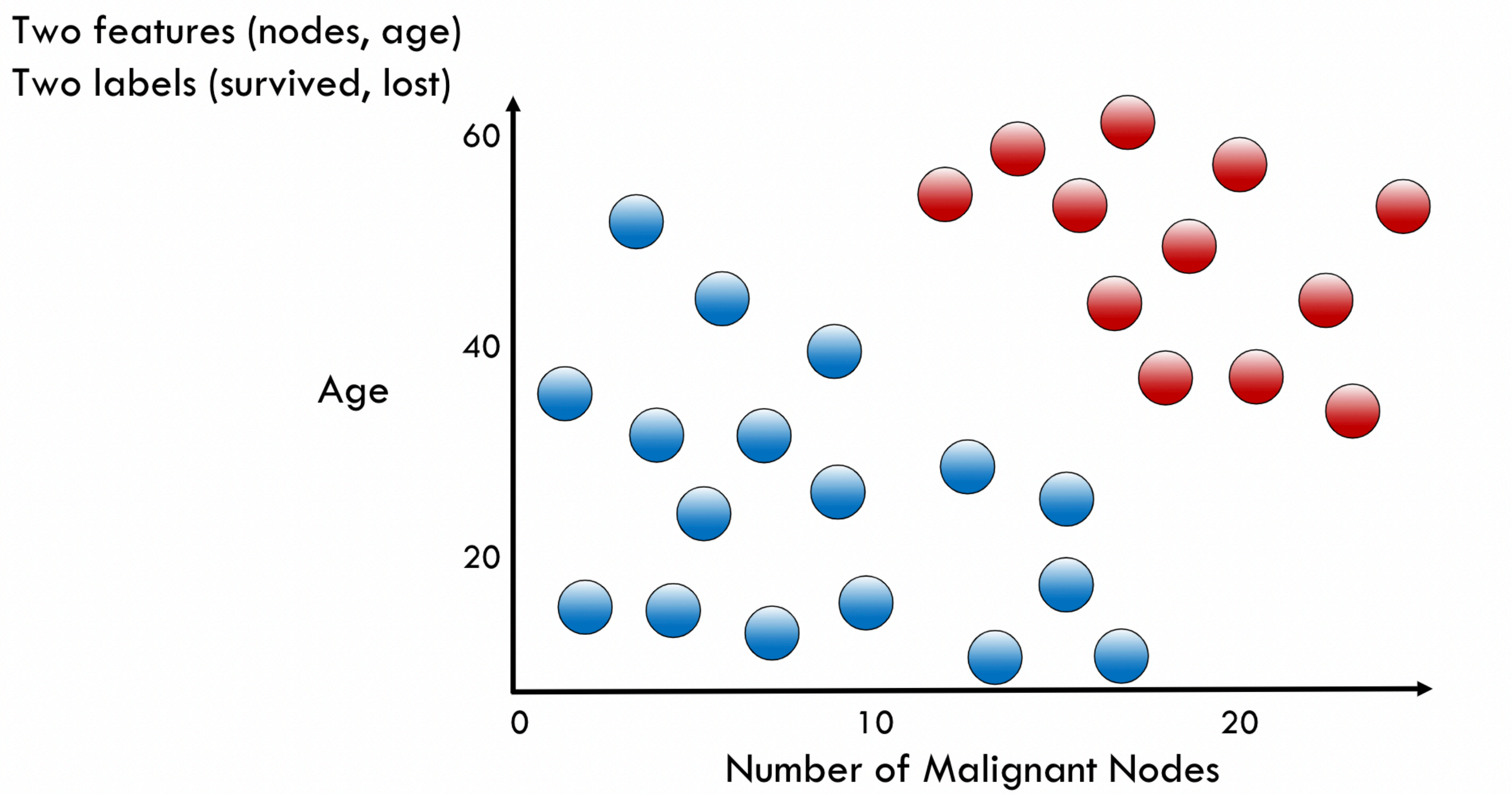
- Logistic regression

$$h_{\boldsymbol{\theta}}(x) = \frac{1}{1 + e^{-\boldsymbol{\theta}^\intercal \boldsymbol{x}}}$$

  - Threshold: .5
    - Predict 1 if $\boldsymbol{\theta}^\intercal \boldsymbol{x} \geq 0$
    - Predict 0 otherwise
  - What we hope for:
    - If "1", $\boldsymbol{\theta}^\intercal \boldsymbol{x} \gg 0$ and if "0", $\boldsymbol{\theta}^\intercal \boldsymbol{x} \ll 0$

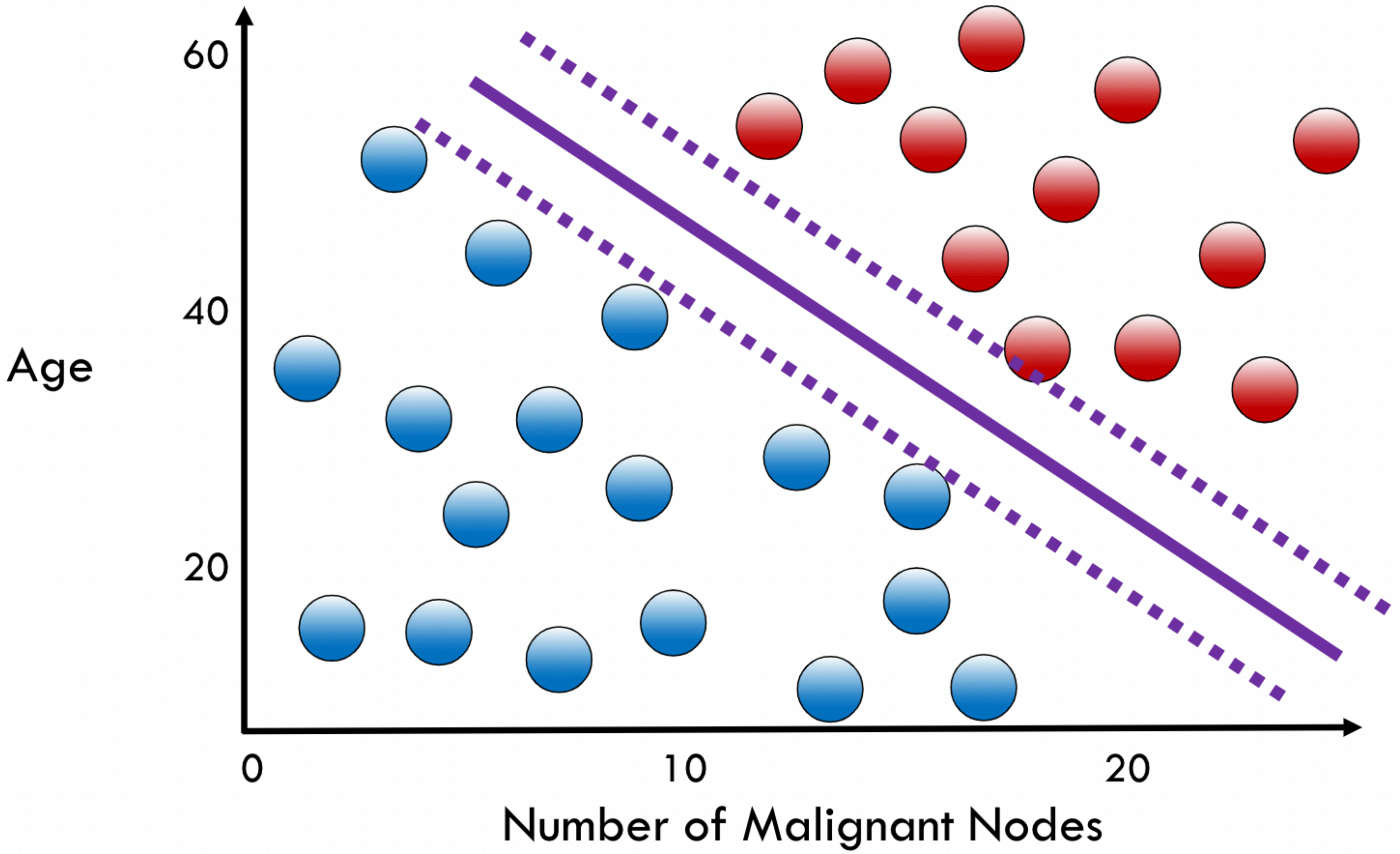
Margins of separation

- Given a data, we could have different lines:



Maximal separation margin

- With SVM, we would like to find



SVM

- Previously in logistic regression:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^\top \boldsymbol{x}}}$$

  - Class labels are 0 or 1

- In SVM,

$$h_{\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}) = g(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

  - Class labels are -1 or 1

SVM

- Maximizing the (functional) margin
  - Margin:

$$\delta^{(i)} = y^{(i)}(\boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b)$$

  - We hope
    - If "1",

$$\boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b \gg 0$$

    - If "-1",

$$\boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b \ll 0$$

    - All together,

$$\delta^{(i)} \gg 0$$

SVM – continued

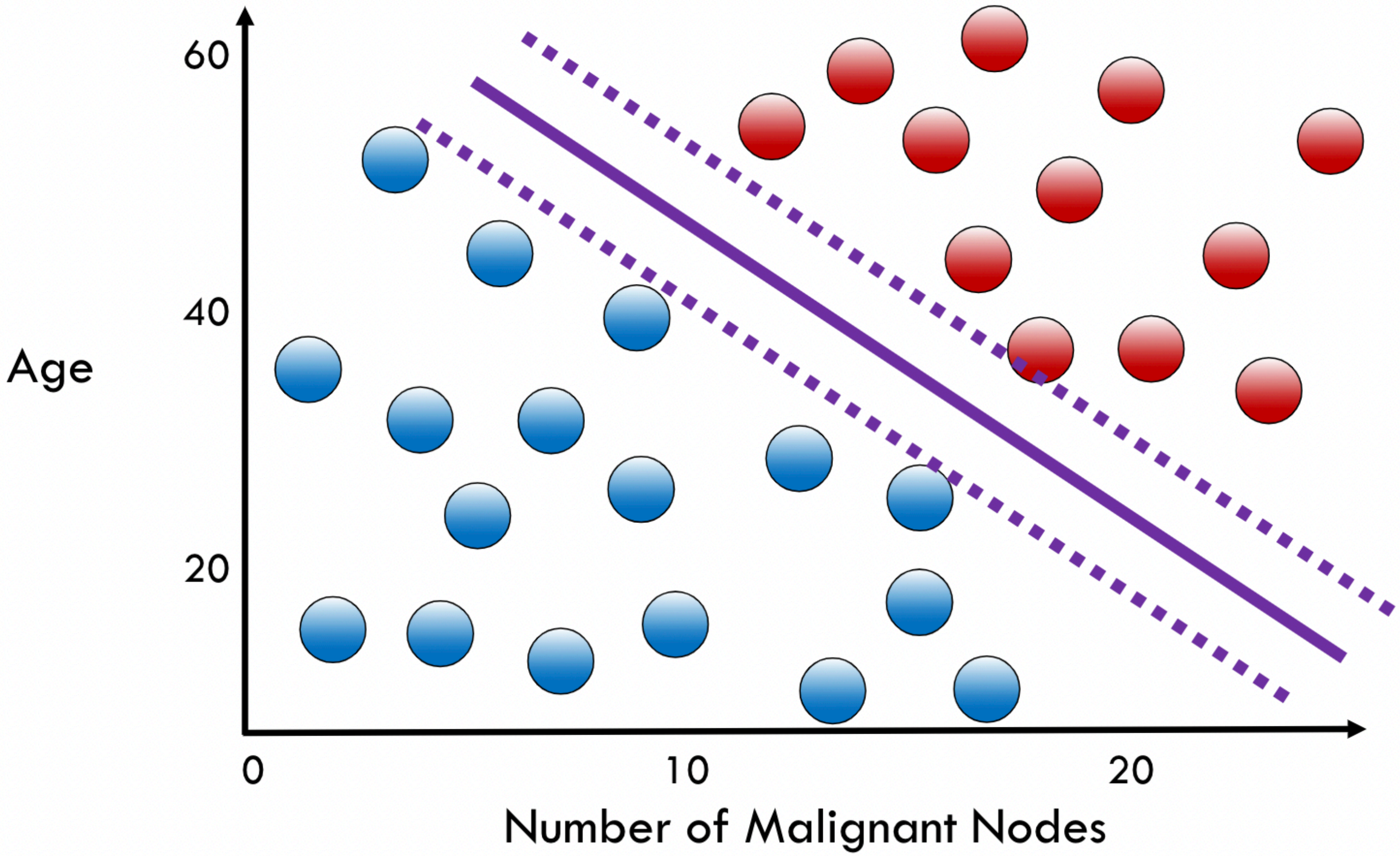
- Functional margin w.r.t. training set

$$\min \delta^{(i)}$$

- Normalization:

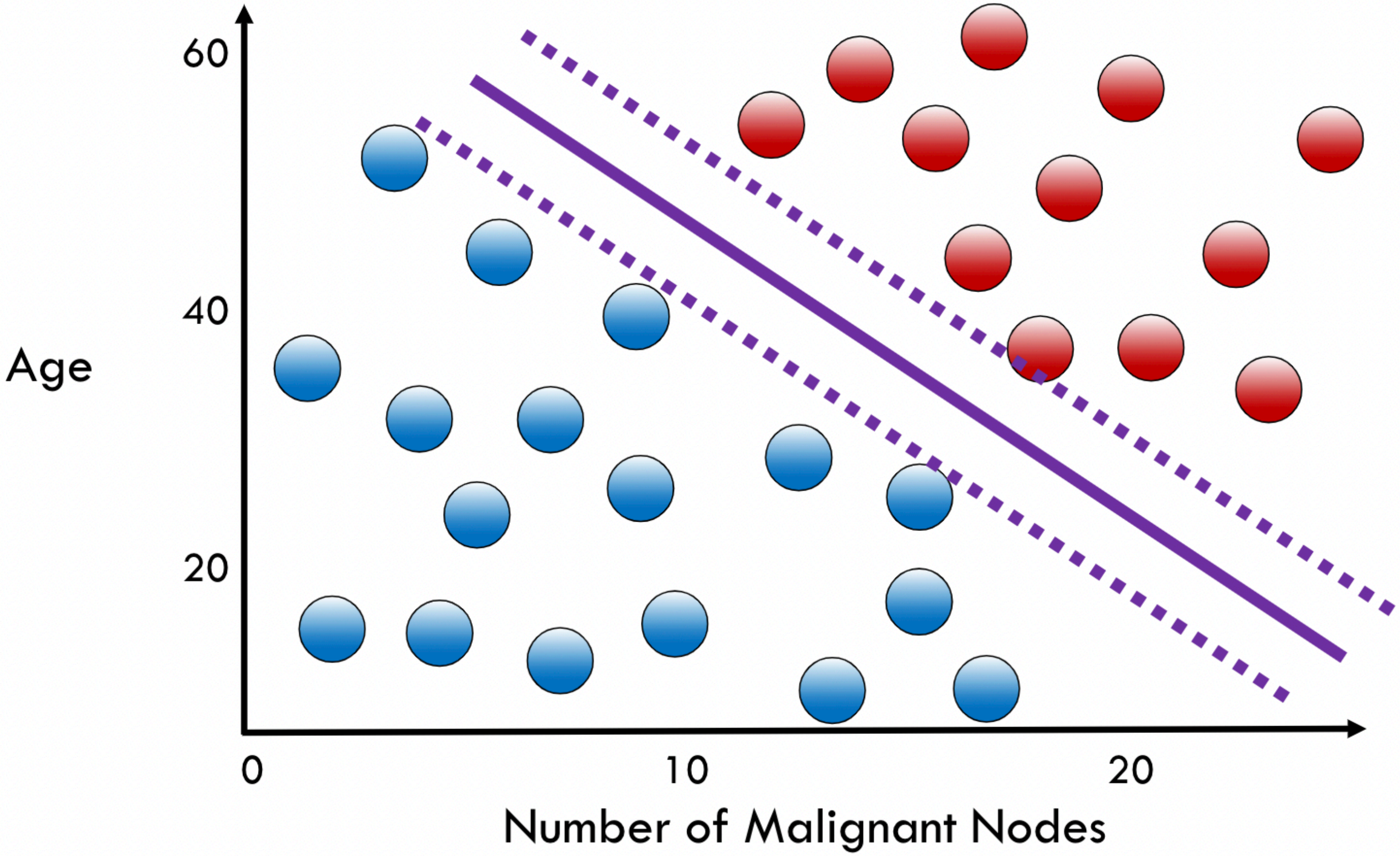
$$\frac{w}{\|w\|_2} \qquad \frac{b}{\|w\|_2}$$

SVM – geometric margin



SVM – geometric margin

- Euclidean distance from a data instance



$$\delta^{(i)} = \frac{\boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b}{\|\boldsymbol{w}\|_2}$$

$$\boldsymbol{w}^\intercal \left(\boldsymbol{x}^{(i)} - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \delta^{(i)} \right) + b = 0$$

$$\delta^{(i)} = \frac{y^{(i)} \left(\boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b \right)}{\|\boldsymbol{w}\|_2}$$

SVM – maximizing the margin

- Optimization problem:

$$\max_{\boldsymbol{w},b} \delta$$

$$\text{such that} \qquad \frac{y^{(i)}\left(\boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b\right)}{\|\boldsymbol{w}\|_2} \geq \delta$$

SVM – (informal) derivation

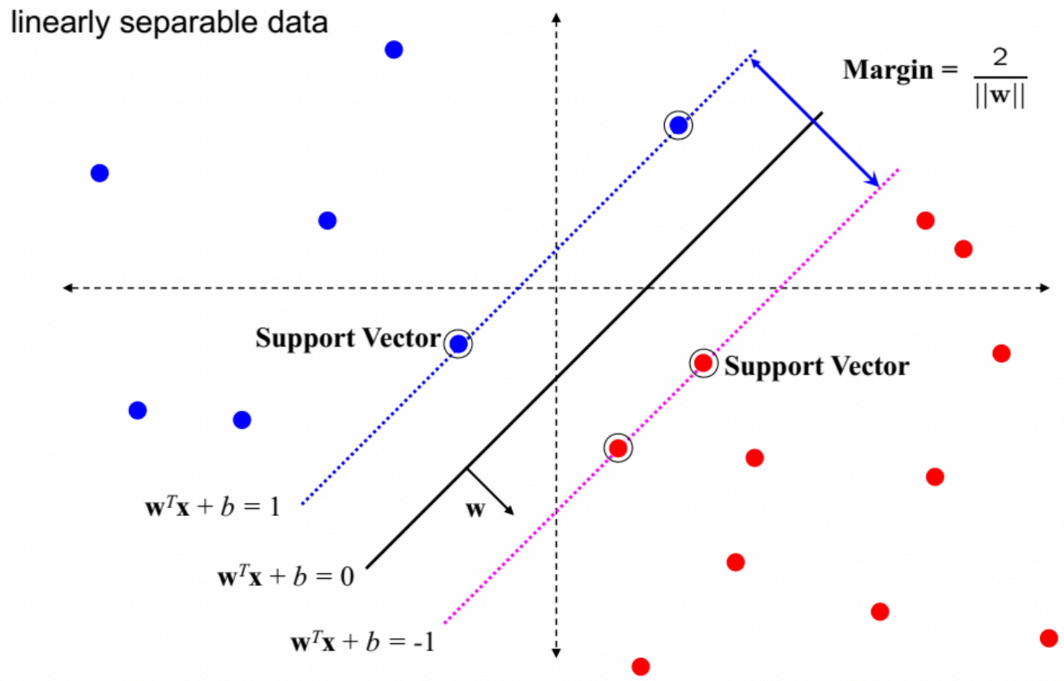
- Choose normalization such that
  - for positive class, $\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b = 1$
  - for negative class, $\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b = -1$

- Optimization objective:

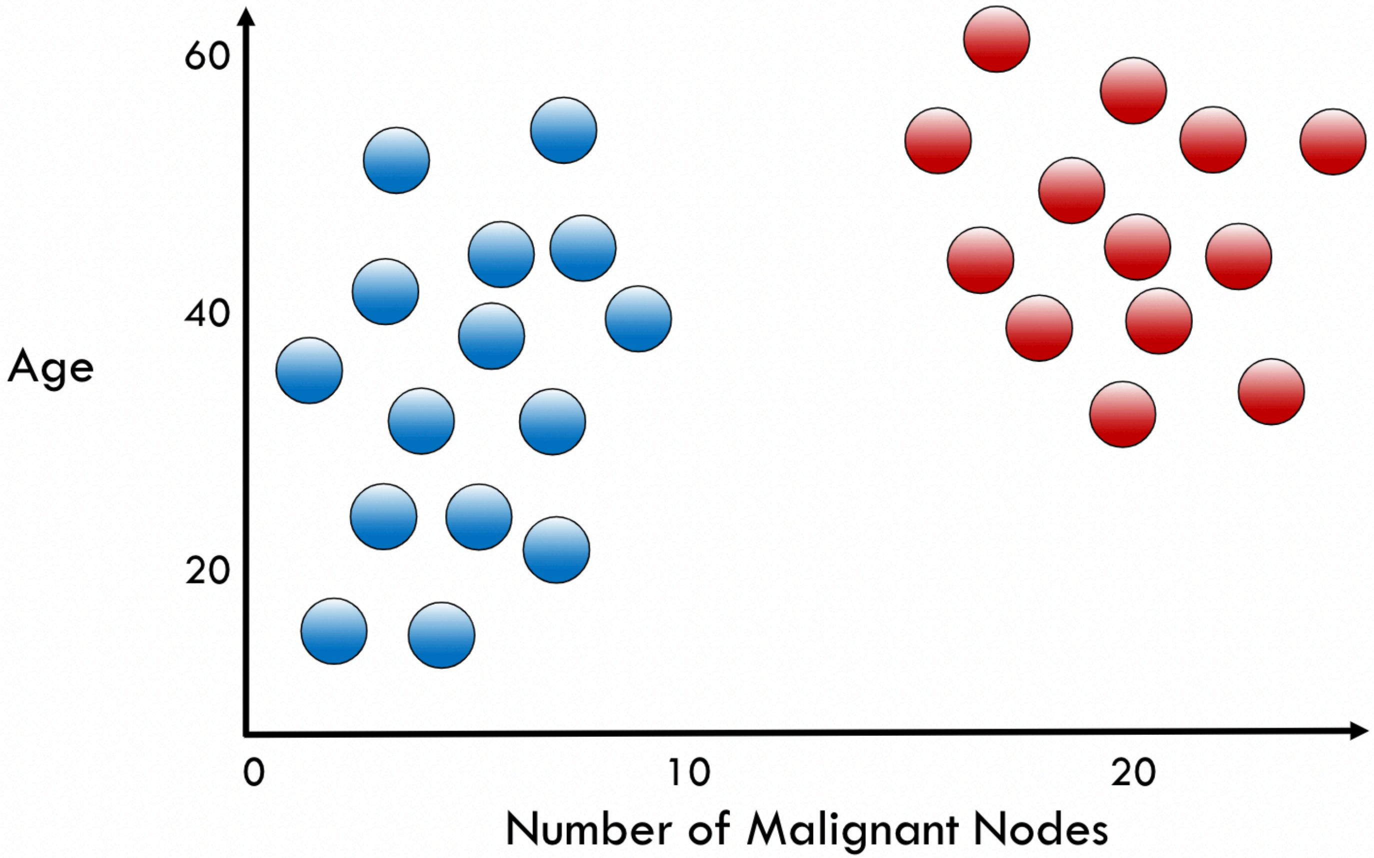
$$\max \frac{2}{\|\boldsymbol{w}\|_2} \quad \text{s.t.} \quad \boldsymbol{w}^\top x^{(i)} + b \geq 1 \text{ if } y^{(i)} = 1$$

$$\boldsymbol{w}^\top x^{(i)} + b \leq -1 \text{ if } y^{(i)} = -1$$

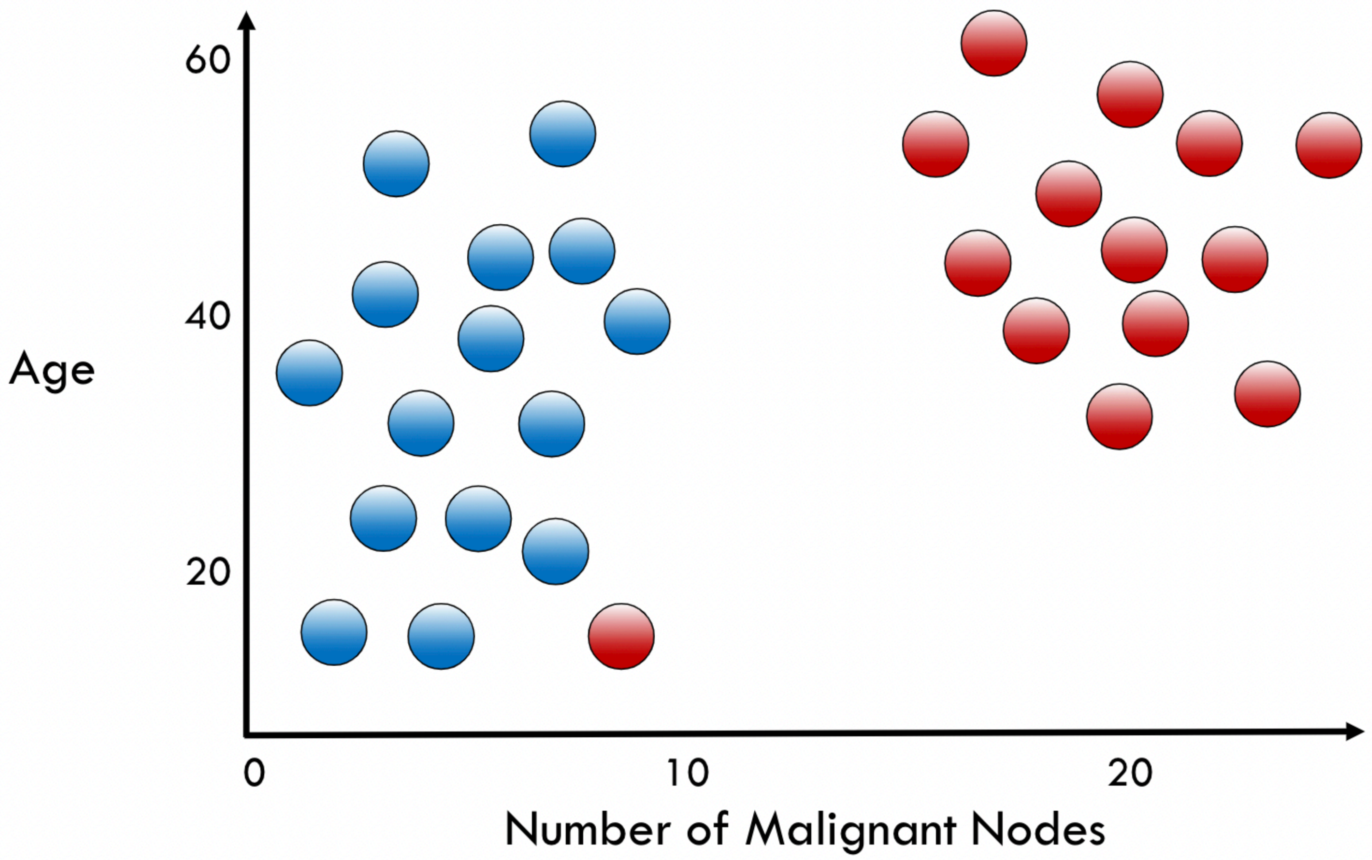
$$\min \|\boldsymbol{w}\| \quad \text{s.t.} \quad y^{(i)} \left(\boldsymbol{w}^\top x^{(i)} + b\right) \geq 1$$



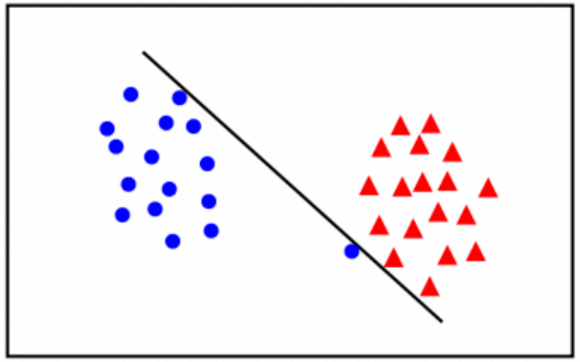
Outlier sensitivity



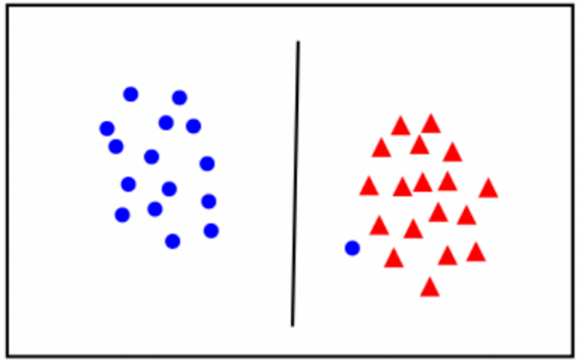
Outlier sensitivity



What is the best model parameter?



• the points can be linearly separated but there is a very narrow margin

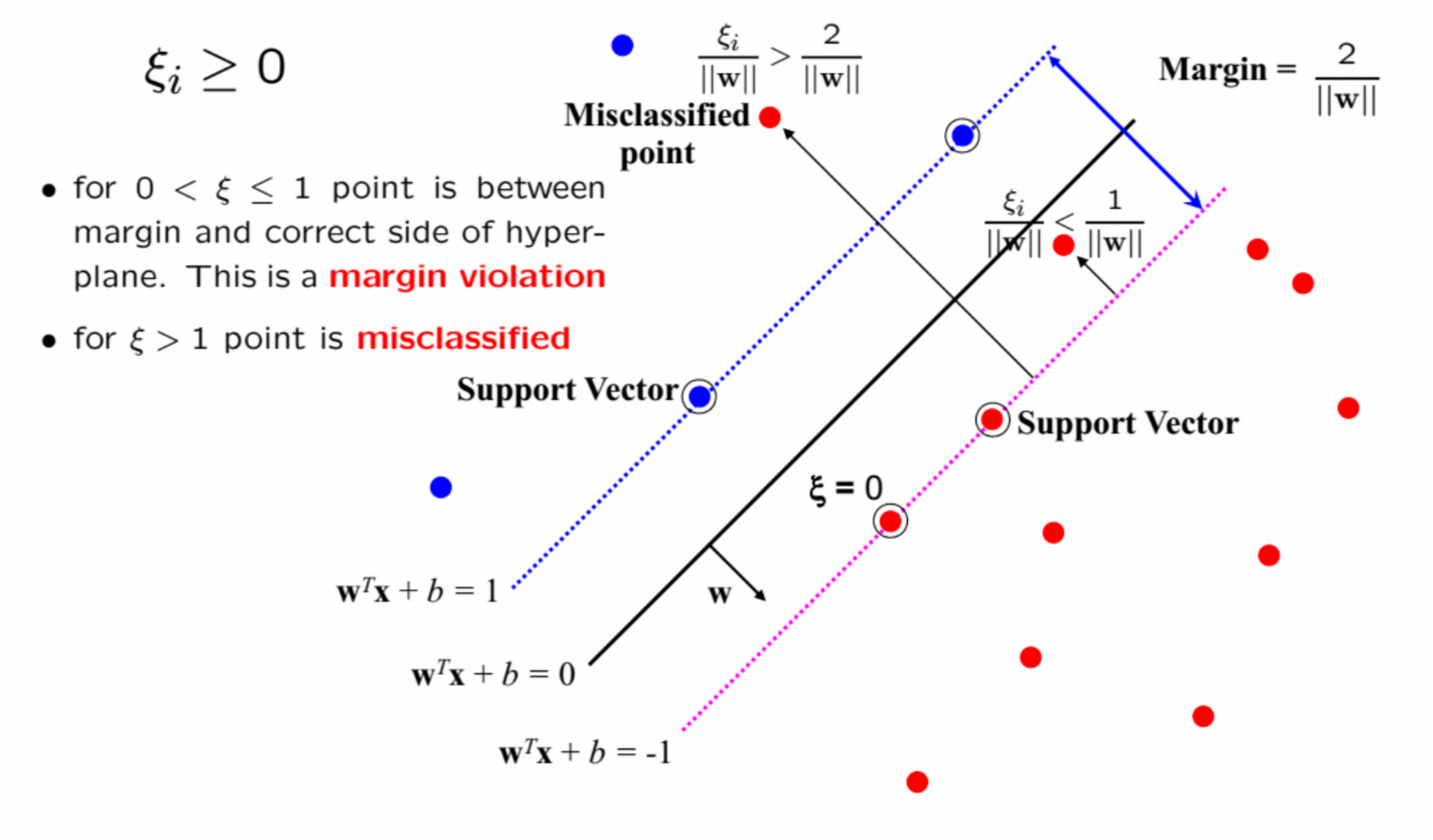


• but possibly the large margin solution is better, even though one constraint is violated

In general there is a trade off between the margin and the number of mistakes on the training data

Regularization

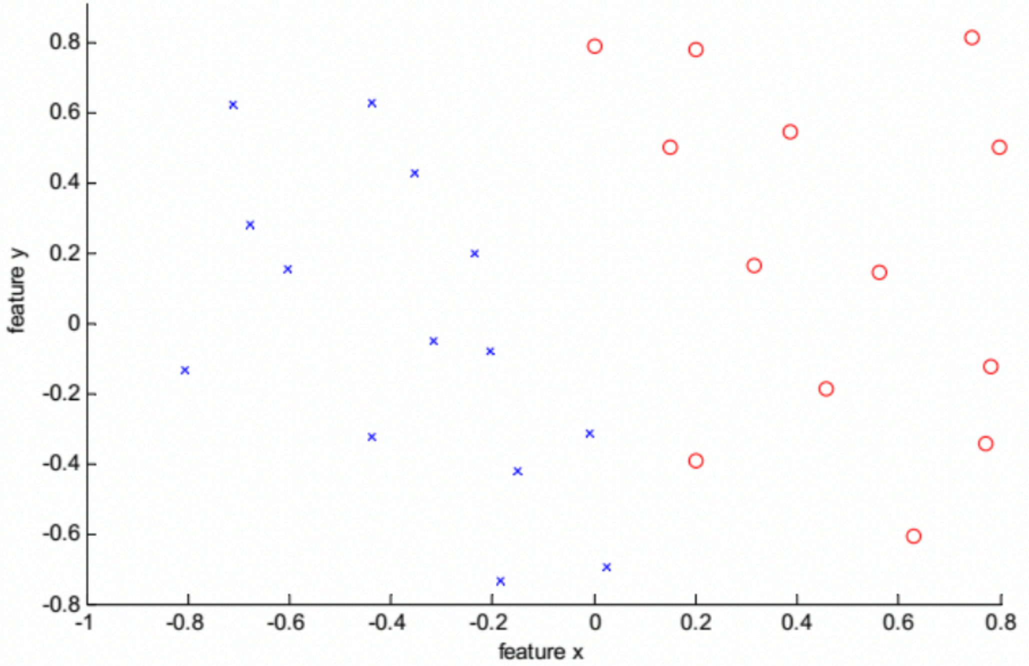
- Introducing slack variables



"Soft" margin SVM

- A new optimization objective:

$$\min \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{N} \xi_i \quad \text{s.t.} \quad y^{(i)} \left(\boldsymbol{w}^\intercal x^{(i)} + b \right) \geq 1 - \xi_i$$



Hard vs soft SVM

- With different values of C