Lecture 16: Learning 4

http://cs.nju.edu.cn/yuy/course_ai15.ashx

Previously...

Learning

Decision tree learning
Neural networks

Why we can learn

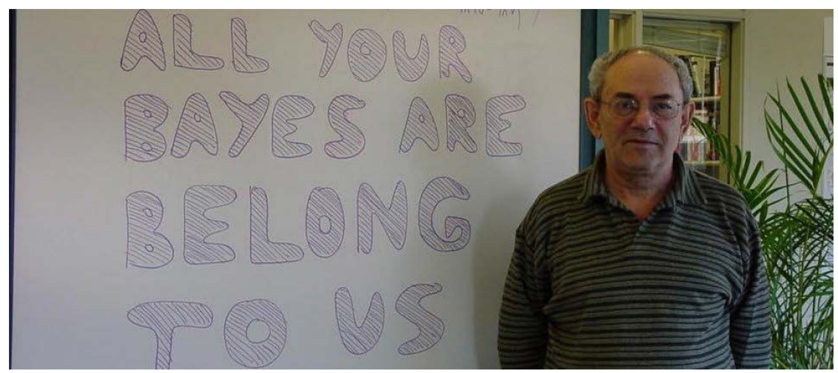
Linear model

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

$$\boldsymbol{w} = \ w_1, w_2, \ldots, w_n \quad b$$

$$w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n + b$$

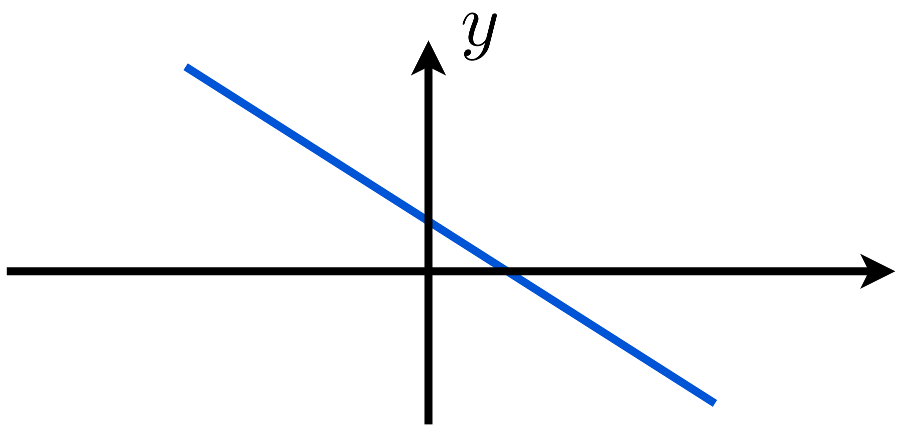
$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$



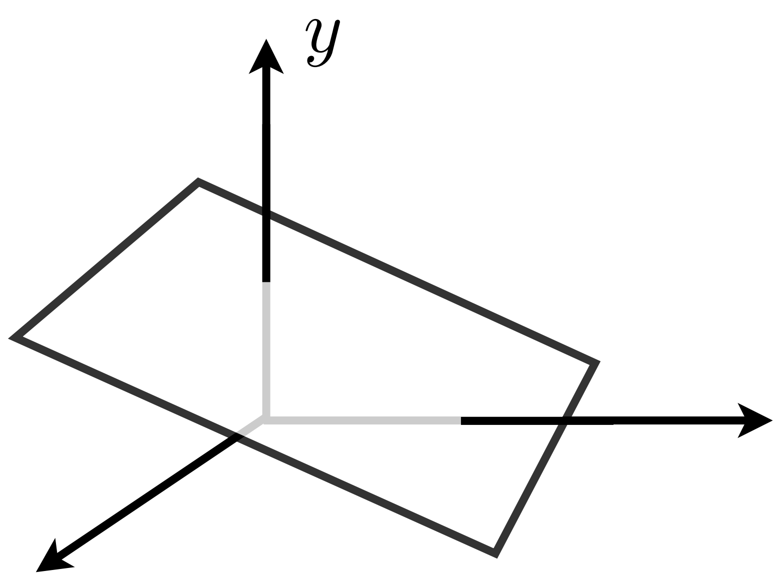
Vladimir Vapnik

Linear model

$$y = ax + b$$

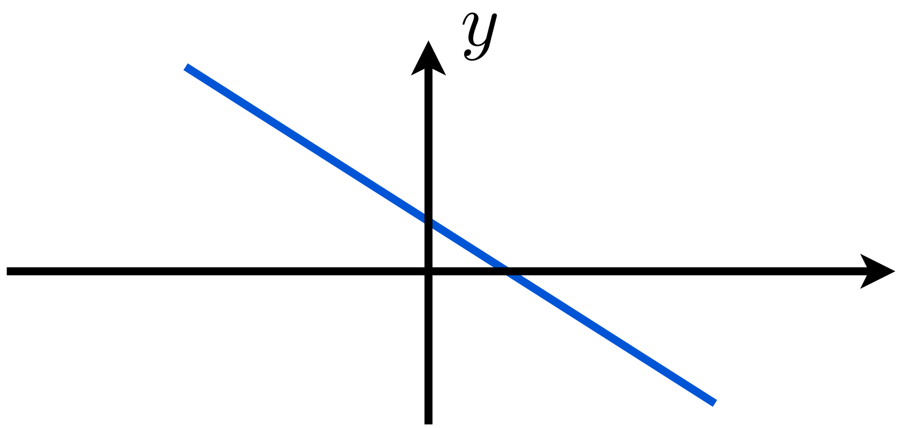


$$y = w_1 \cdot x_1 + w_2 \cdot x_2 + b$$

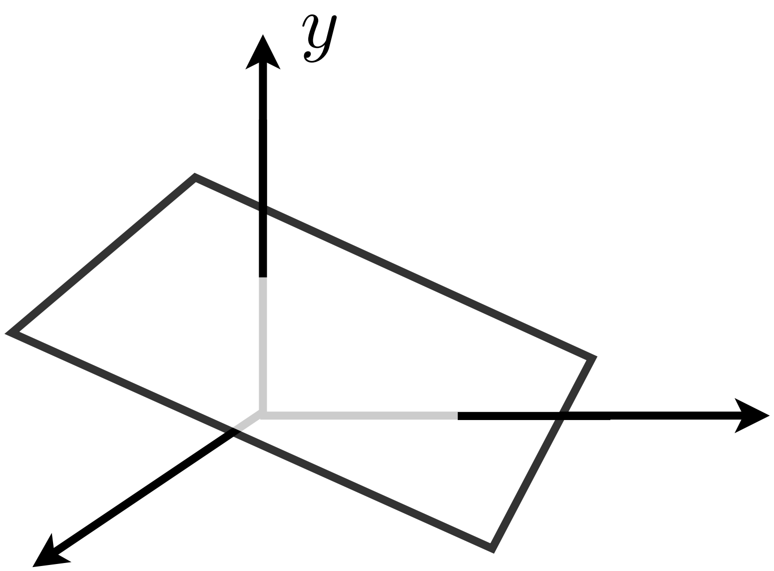


Linear model

$$y = ax + b$$



$$y = w_1 \cdot x_1 + w_2 \cdot x_2 + b$$



is the following a linear model?

$$y = w_1 \cdot x + w_2 \cdot x^2 + b$$

Least square regression

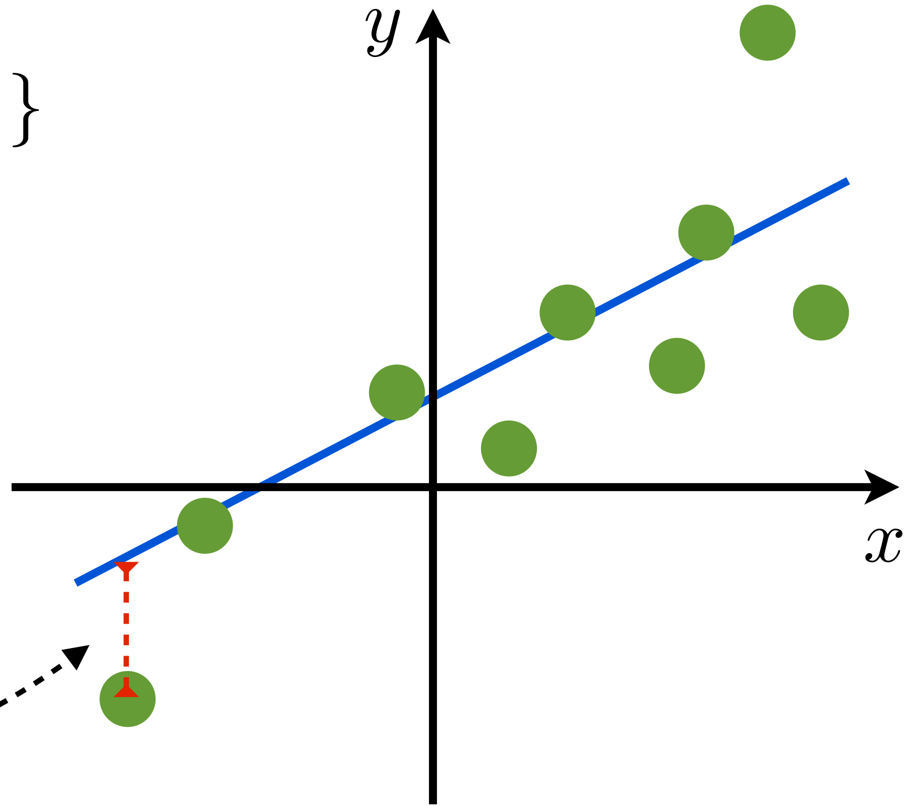
Regression: $y \in \mathbb{R}$

Training data:

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_m, y_m)\}$$

Least square loss:

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$



Least square regression

$$L(\boldsymbol{w}, b) = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} 2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i) = 0$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = \frac{1}{m}\sum_{i=1}^{m} 2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)\boldsymbol{x}_i^\top = 0$$

Least square regression

$$L(\boldsymbol{w}, b) = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} 2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i) = 0$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = \frac{1}{m}\sum_{i=1}^{m} 2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)\boldsymbol{x}_i^\top = 0$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) = \bar{y} - \boldsymbol{w}^\top \bar{\boldsymbol{x}}$$

Least square regression

$$L(\boldsymbol{w}, b) = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} 2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i) = 0$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = \frac{1}{m}\sum_{i=1}^{m} 2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)\boldsymbol{x}_i^\top = 0$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) = \bar{y} - \boldsymbol{w}^\top \bar{\boldsymbol{x}}$$

$$\boldsymbol{w} = \left(\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}(y_i \boldsymbol{x}_i) - \bar{y}\bar{\boldsymbol{x}}\right)$$

$$= var(\boldsymbol{x})^{-1} cov(\boldsymbol{x}, y) = (X^\top X)^{-1} X^\top Y$$

Least square regression

$$L(\boldsymbol{w}, b) = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i) = 0$$

$$\frac{\partial L(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = \frac{1}{m}\sum_{i=1}^{m}2(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)\boldsymbol{x}_i^\top = 0$$

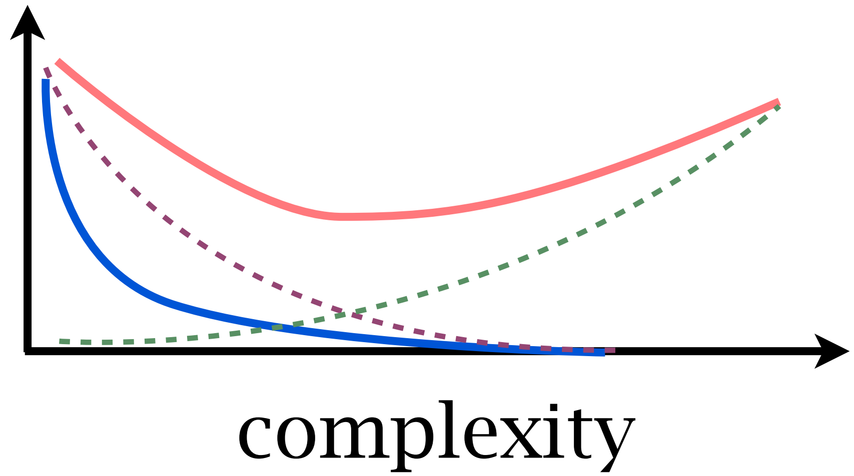
$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) = \bar{y} - \boldsymbol{w}^\top \bar{\boldsymbol{x}}$$

$$\boldsymbol{w} = \left(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}(y_i\boldsymbol{x}_i) - \bar{y}\bar{\boldsymbol{x}}\right)$$

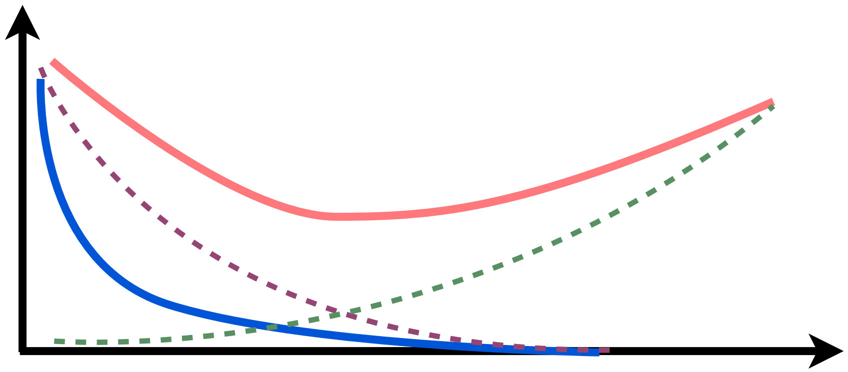
$$= var(\boldsymbol{x})^{-1}cov(\boldsymbol{x}, y) = (X^\top X)^{-1}X^\top Y$$

closed form solution

Complexity of linear models



Complexity of linear models



$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$$

possibility of $\boldsymbol{w}$

Regularization

make hypothesis space small

$\rightarrow$ better generalization ability

make numerical analysis stable

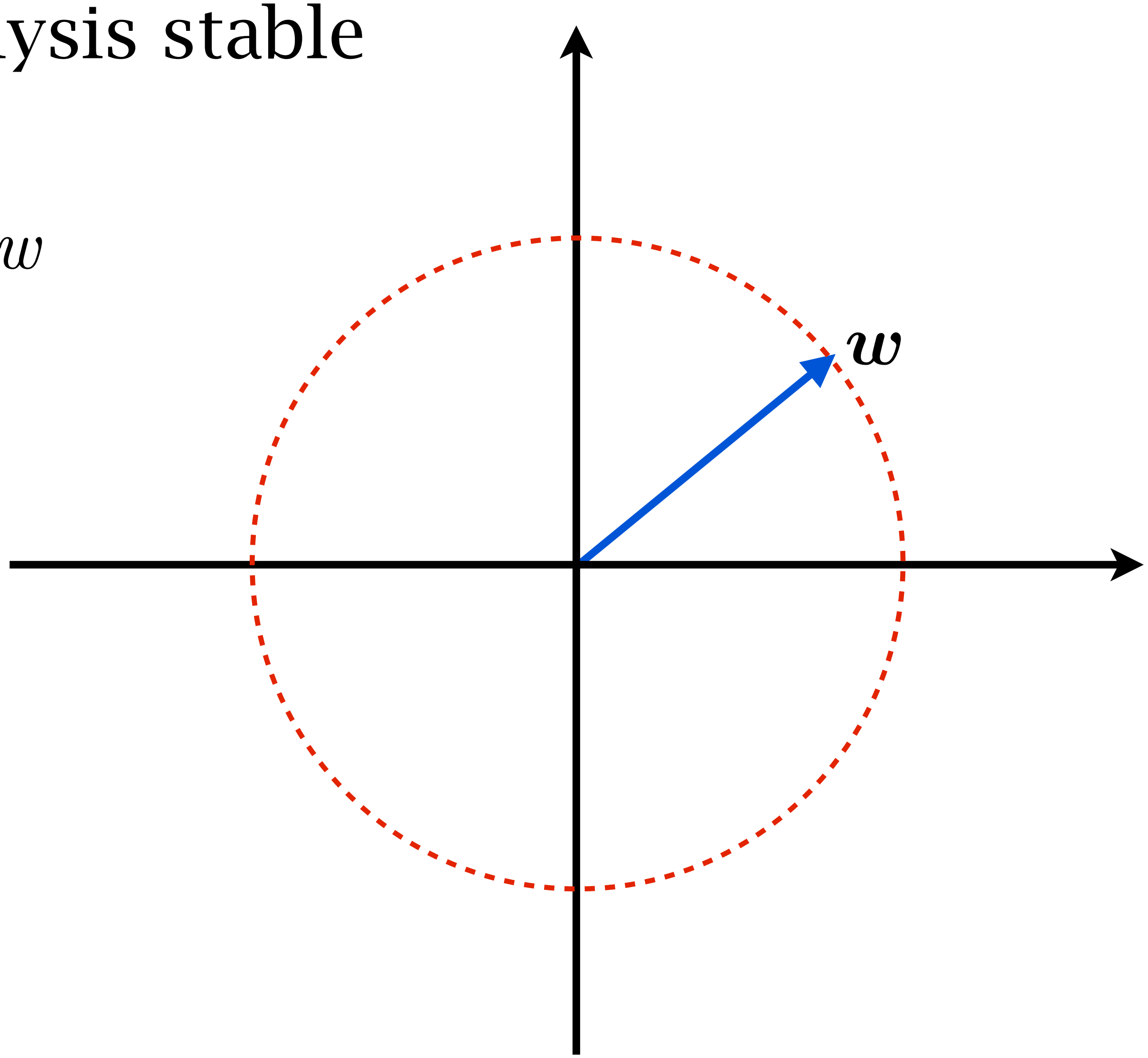
restrict the norm of w

$$\|\boldsymbol{w}\|_p = \left(\sum_{i=1}^{n} |w_i|^p \right)^{1/p}$$

$$\|\boldsymbol{w}\|_2 = \sqrt{\sum_{i=1}^{n} w_i^2}$$

$$\|\boldsymbol{w}\|_1 = \sum_{i=1}^{n} |w_i|$$

$$\|\boldsymbol{w}\|_\infty = \max_{i=1,\ldots,n} |w_i|$$



Ridge regression

Regression: $y \in \mathbb{R}$

Training data:

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_m, y_m)\}$$

objective:

$$\arg\min_{\boldsymbol{w}, b} \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$

$$s.t. \qquad \|\boldsymbol{w}\|_2 \leq \theta$$

or:

$$\arg\min_{\boldsymbol{w}, b} \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2 + \lambda \|\boldsymbol{w}\|_2$$

Ridge regression

centered data, no bias:

$$\arg\min_{\boldsymbol{w}} \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2 + \lambda\|\boldsymbol{w}\|_2$$

closed form solution:

$$\boldsymbol{w} = \Big(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \underbrace{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top}_{0} + \lambda\boldsymbol{I}\Big)^{-1}\Big(\frac{1}{m}\sum_{i=1}^{m}(y_i\boldsymbol{x}_i) - \underbrace{\bar{y}\bar{\boldsymbol{x}}}_{0}\Big)$$

$$= (var(\boldsymbol{x}) + \lambda\boldsymbol{I})^{-1}cov(\boldsymbol{x}, y)$$

$$= (X^\top X + \lambda I)^{-1}X^\top Y$$

$\boldsymbol{I}$ is the identity matrix

Least square v.s. ridge regression

$$\boldsymbol{w} = \Big(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\Big)^{-1}\Big(\frac{1}{m}\sum_{i=1}^{m}(y_i\boldsymbol{x}_i) - \bar{y}\bar{\boldsymbol{x}}\Big)$$

$$= var(\boldsymbol{x})^{-1}cov(\boldsymbol{x}, y) = (X^\top X)^{-1}X^\top Y$$

$$\boldsymbol{w} = \Big(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top + \lambda\boldsymbol{I}\Big)^{-1}\Big(\frac{1}{m}\sum_{i=1}^{m}(y_i\boldsymbol{x}_i) - \bar{y}\bar{\boldsymbol{x}}\Big)$$

$$= (var(\boldsymbol{x}) + \lambda\boldsymbol{I})^{-1}cov(\boldsymbol{x}, y)$$

$$= (X^\top X + \lambda I)^{-1}X^\top Y$$

stable solution

Least absolute shrinkage and selection operator (LASSO)

Regression: $y \in \mathbb{R}$

Training data:

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_m, y_m)\}$$

objective:

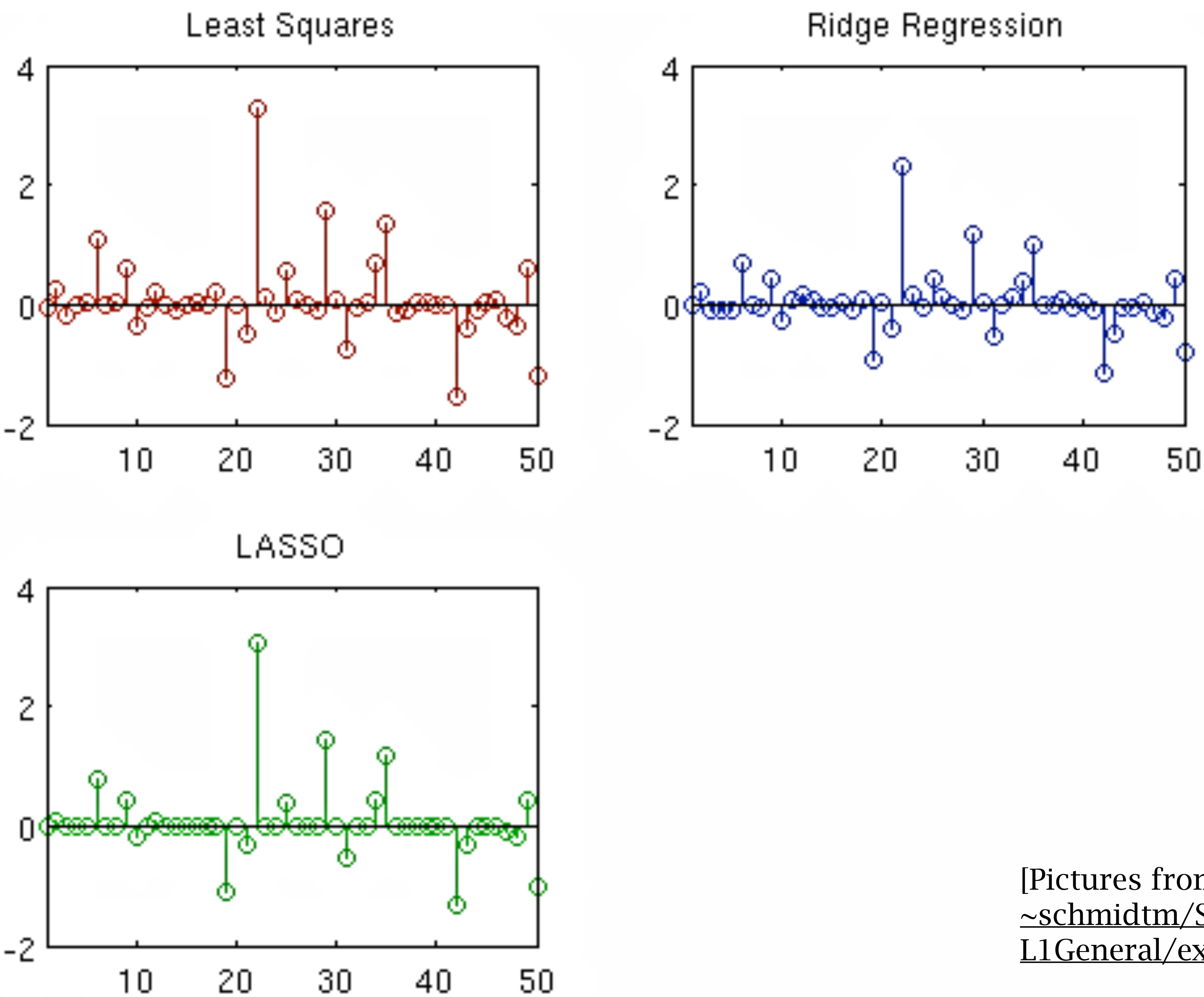
$$\arg\min_{\boldsymbol{w}, b} \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$

$$s.t. \qquad \|\boldsymbol{w}\|_1 \leq \theta$$

or:

$$\arg\min_{\boldsymbol{w}, b} \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2 + \lambda \|\boldsymbol{w}\|_1$$

Comparing different regressions



[Pictures from www.cs.ubc.ca/~schmidtm/Software/L1General/examples.html]

A general framework

objective function:

$$\arg\min_{\boldsymbol{w},b} L(\boldsymbol{w},b) + \|\boldsymbol{w}\|_p$$

general optimization: gradient descent

$$(\boldsymbol{w},b) -= \eta \frac{\partial (L(\boldsymbol{w},b) + \|\boldsymbol{w}\|_p)}{\partial(\boldsymbol{w},b)}$$

good for convex objective functions

$$f(\alpha \boldsymbol{w}_1 + (1-\alpha)\boldsymbol{w}_2)) \geq \alpha f(\boldsymbol{w}_1) + (1-\alpha) f(\boldsymbol{w}_2)$$

linear, quadratic

convex + convex → convex

Linear classifier

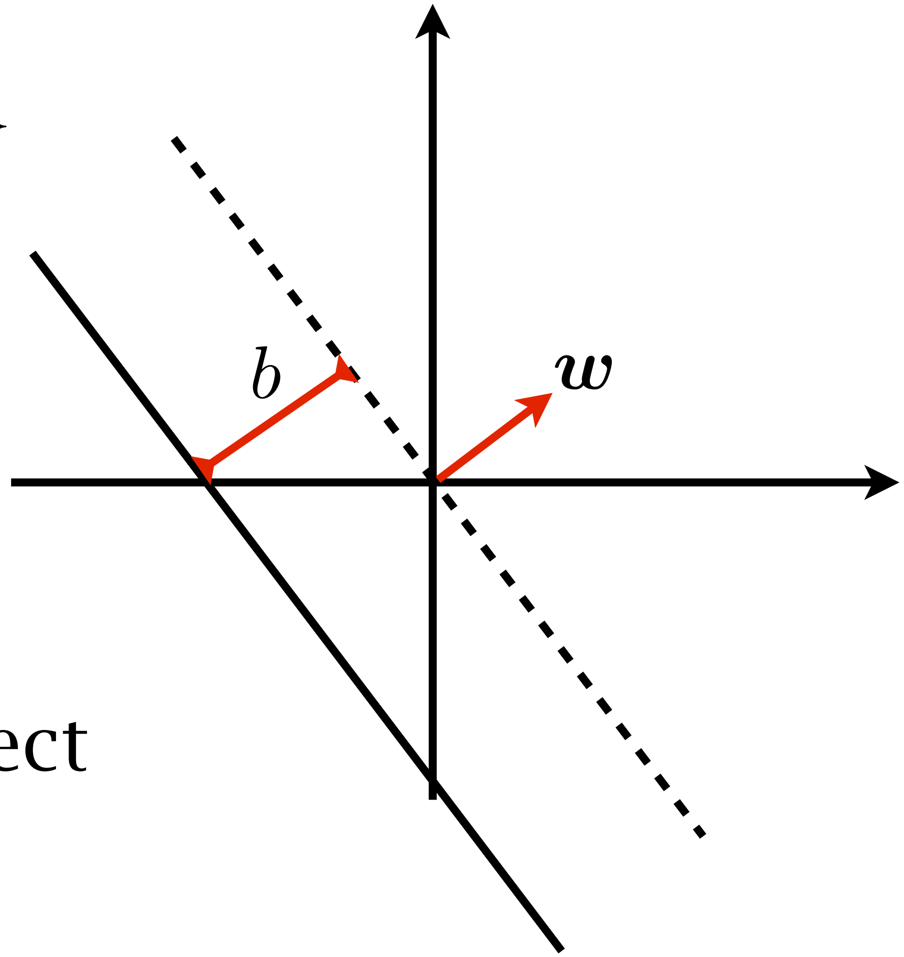
model space: $\mathbb{R}^{n+1}$

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

for classification $y \in \{-1, +1\}$

we predict an instance by

$$\operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b) = \begin{cases} +1, & \boldsymbol{w}^\top \boldsymbol{x} + b > 0 \\ -1, & \boldsymbol{w}^\top \boldsymbol{x} + b < 0 \\ \text{random}, & otherwise \end{cases}$$



for an example $(\boldsymbol{x}, y)$, a correct prediction means

$$y(\boldsymbol{w}^\top \boldsymbol{x} + b) > 0$$

Idea classifier

$$\arg\min_{\boldsymbol{w},b} \sum_i I(y(\boldsymbol{w}^\top \boldsymbol{x} + b) \leq 0)$$

not convex
hard to solve

Prototype

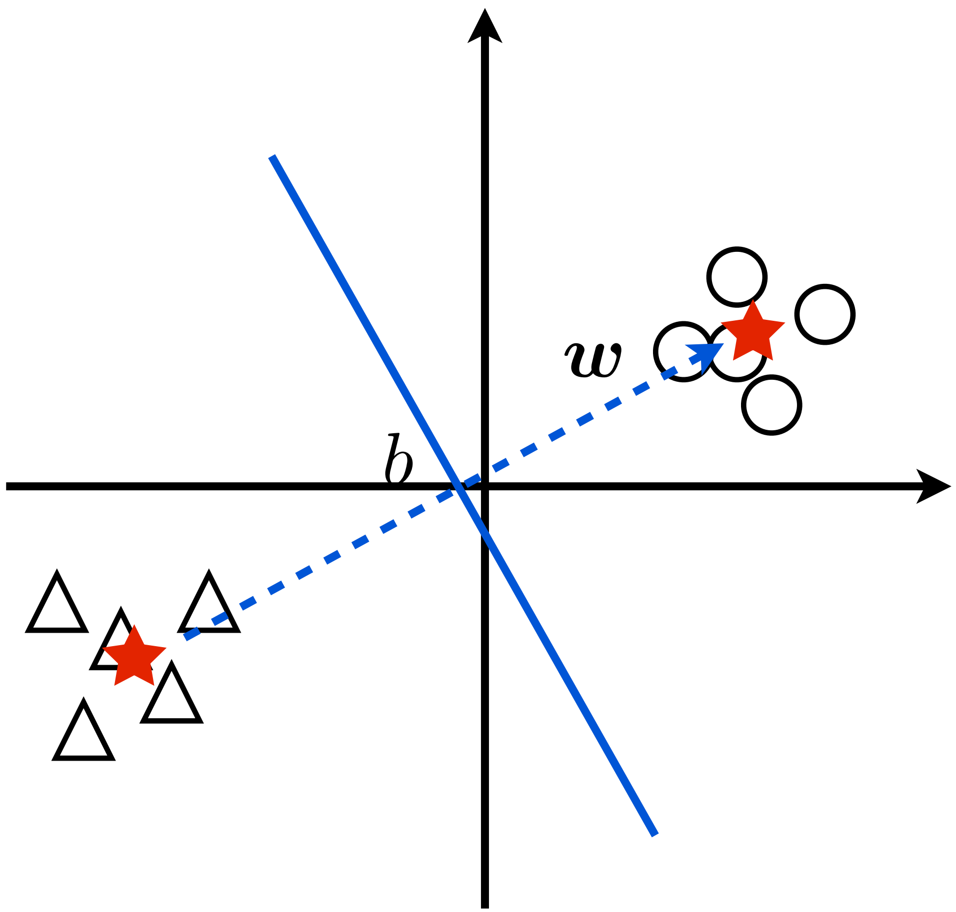
simple, but too restricted

$$\bar{\boldsymbol{x}}^+ = \frac{1}{\sum_{i:y_i=+1} 1} \sum_{i:y_i=+1} \boldsymbol{x}_i$$

$$\bar{\boldsymbol{x}}^- = \frac{1}{\sum_{i:y_i=-1} 1} \sum_{i:y_i=-1} \boldsymbol{x}_i$$

$$\boldsymbol{w} = \bar{\boldsymbol{x}}^+ - \bar{\boldsymbol{x}}^-$$

$$b = -\boldsymbol{w}^\top \cdot \frac{\bar{\boldsymbol{x}}^+ + \bar{\boldsymbol{x}}^-}{2}$$



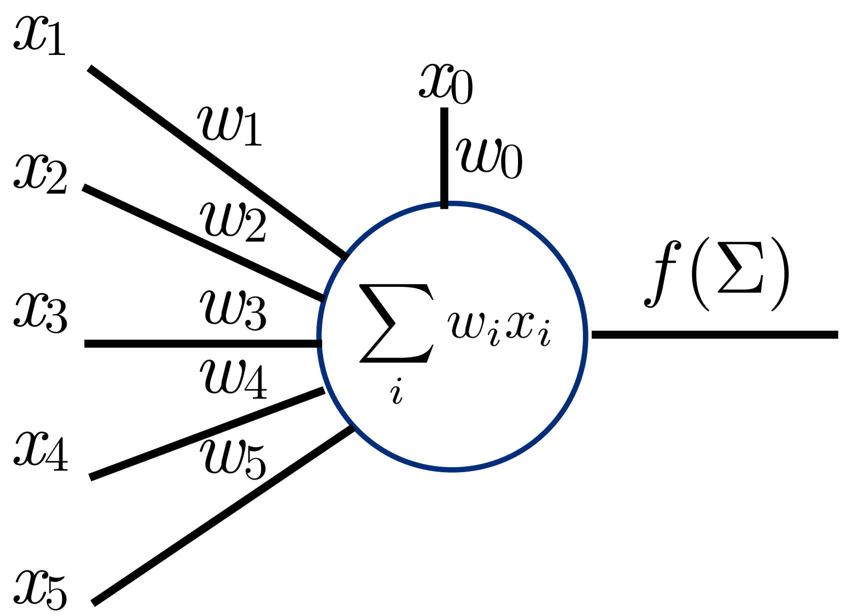
Perceptron

feed training examples one by one

1. $\boldsymbol{w} = 0$

2. for each example $(\boldsymbol{x}, y)$
   if $\mathrm{sign}(y\boldsymbol{w}^\top \boldsymbol{x}) < 0$

$$\boldsymbol{w} = \boldsymbol{w} + y\boldsymbol{x}$$



$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

Perceptron

feed training examples one by one

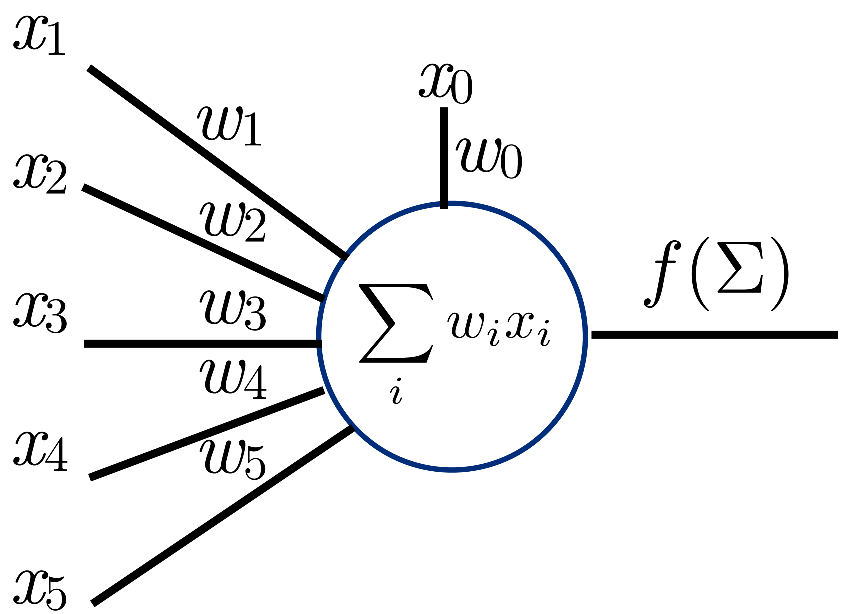
1. $\boldsymbol{w} = 0$

2. for each example $(\boldsymbol{x}, y)$
   if $\mathrm{sign}(y\boldsymbol{w}^\top \boldsymbol{x}) < 0$

$$\boldsymbol{w} = \boldsymbol{w} + y\boldsymbol{x}$$

gradient ascent

$$\frac{\partial y\boldsymbol{w}^\top \boldsymbol{x}}{\partial \boldsymbol{w}} = y\boldsymbol{x}$$



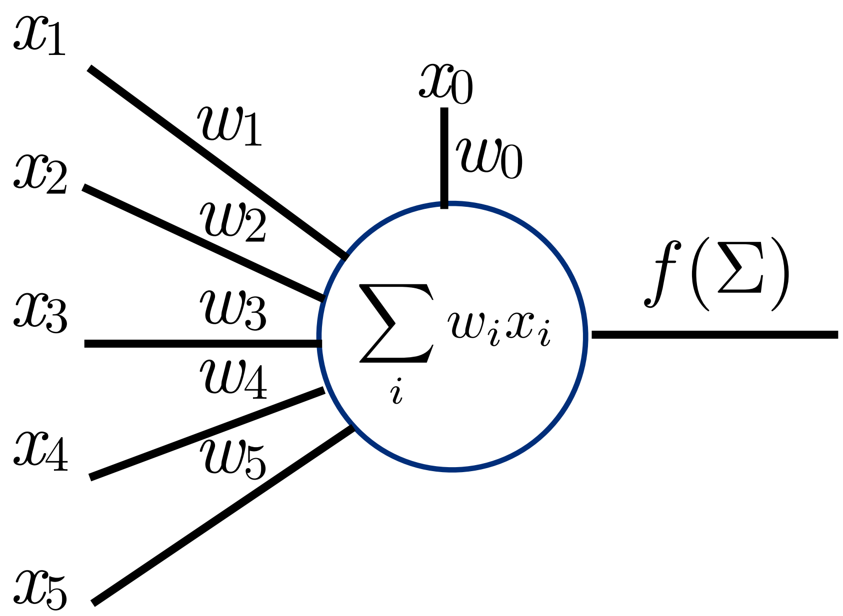
$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

Perceptron

feed training examples one by one

1. $\boldsymbol{w} = 0$
2. for each example $(\boldsymbol{x}, y)$
   if $\text{sign}(y\boldsymbol{w}^\top \boldsymbol{x}) < 0$

$$\boldsymbol{w} = \boldsymbol{w} + y\boldsymbol{x}$$



gradient ascent

$$\frac{\partial y\boldsymbol{w}^\top \boldsymbol{x}}{\partial \boldsymbol{w}} = y\boldsymbol{x}$$

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

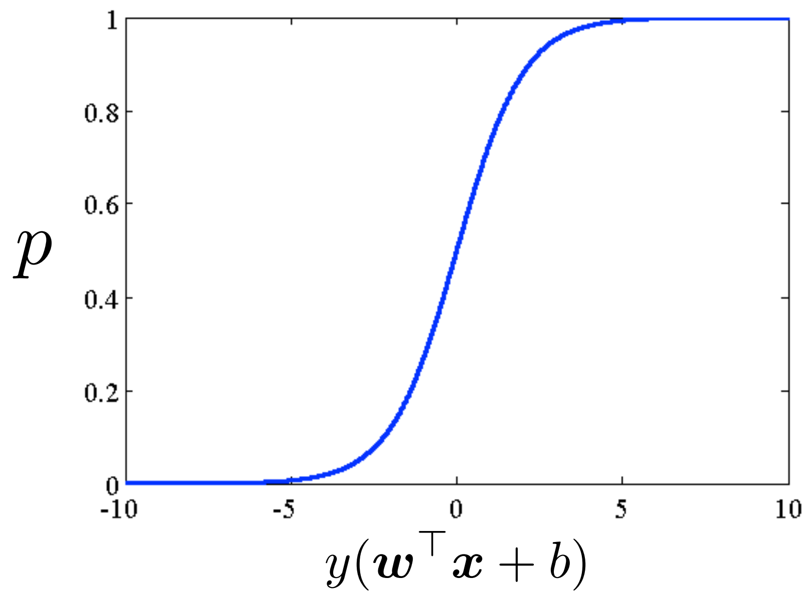
when all examples are with length 1 and are linearly separable by w^*, perceptron algorithm makes at most $\left(1/\min_{\boldsymbol{x}} \frac{|\boldsymbol{w}^{*\top}\boldsymbol{x}|}{\|\boldsymbol{x}\|_2}\right)^2$ mistakes

Logistic regression

assume logit model: for a positive example

$$\boldsymbol{w}^\top \boldsymbol{x} = \log \frac{p(+1 \mid \boldsymbol{x})}{1 - p(+1 \mid \boldsymbol{x})}$$

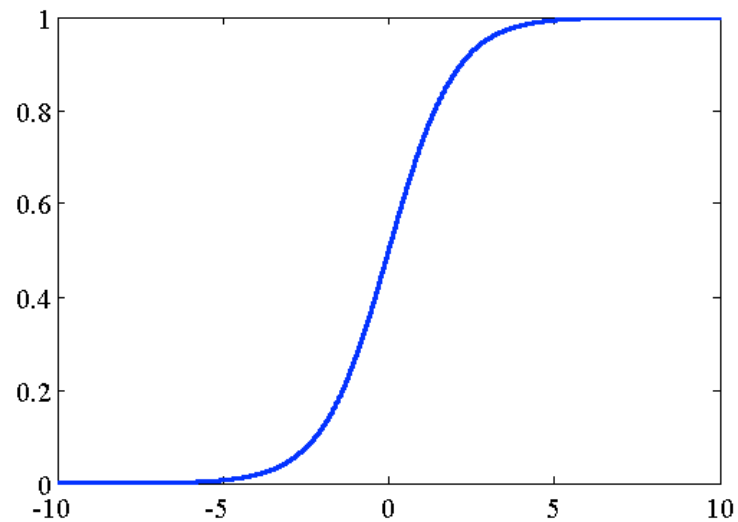
so that $p(y \mid \boldsymbol{x}, \boldsymbol{w}) = \dfrac{1}{1 + e^{-y(\boldsymbol{w}^\top \boldsymbol{x})}}$



Logistic regression

assume logit model: for a positive example

$$\boldsymbol{w}^\top \boldsymbol{x} = \log \frac{p(+1 \mid \boldsymbol{x})}{1 - p(+1 \mid \boldsymbol{x})}$$

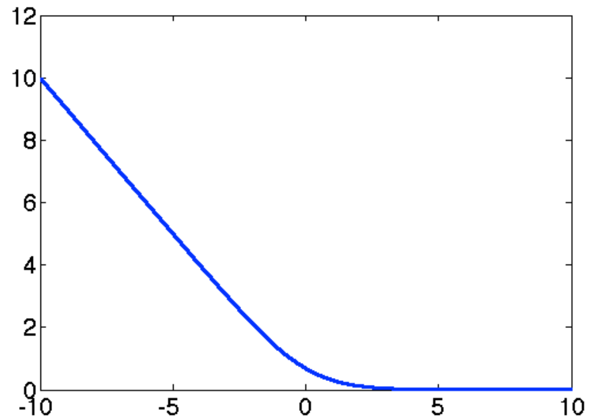


so that $p(y \mid \boldsymbol{x}, \boldsymbol{w}) = \dfrac{1}{1 + e^{-y(\boldsymbol{w}^\top \boldsymbol{x})}}$

minimize negative log-likelihood:

$$\arg\min_{\boldsymbol{w}, b} - \log \prod_{i=1}^{m} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{w}) = -\sum_i \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{w})$$

$$= \sum_i \log \left(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i)}\right)$$



convex

Linear classifier revisit

model space: $\mathbb{R}^{n+1}$

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

for classification $y \in \{-1, +1\}$

Original objective:

$$\arg\min_{\boldsymbol{w},b} \sum_i I(y(\boldsymbol{w}^\top \boldsymbol{x} + b) \leq 0)$$

0-1 loss
hard to optimize

Surrogate objective:

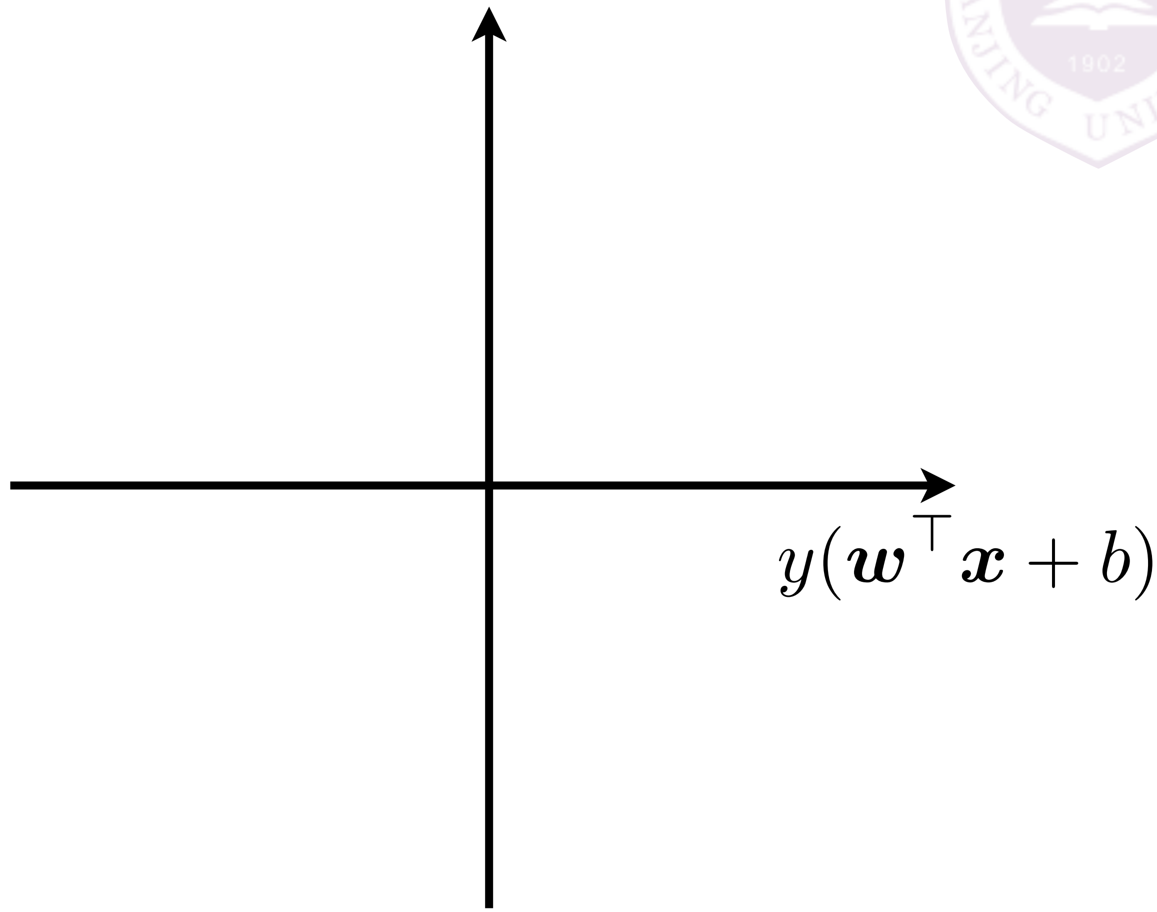
$$\arg\min_{\boldsymbol{w},b} \sum_i \log\left(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}\right)$$

logistic regression

$$\arg\min_{\boldsymbol{w},b} \sum_i \max\{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0\}$$

perceptron

Linear classifier revisit



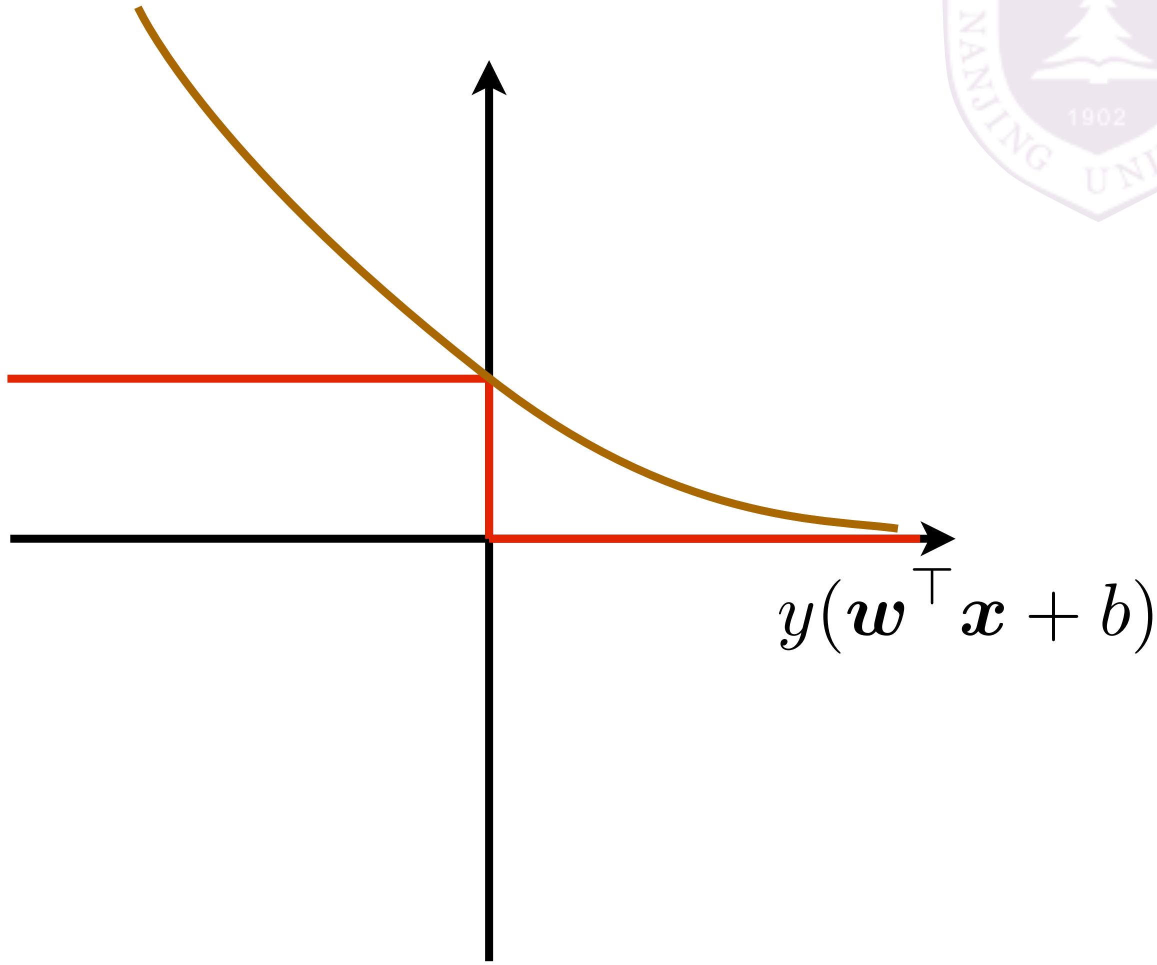
Linear classifier revisit

0-1 loss

$I(y(\boldsymbol{w}^\top \boldsymbol{x} + b) \leq 0)$

logistic regression

$\log_2(1 + e^{-y(\boldsymbol{w}^\top \boldsymbol{x}+b)})$



Linear classifier revisit

0-1 loss

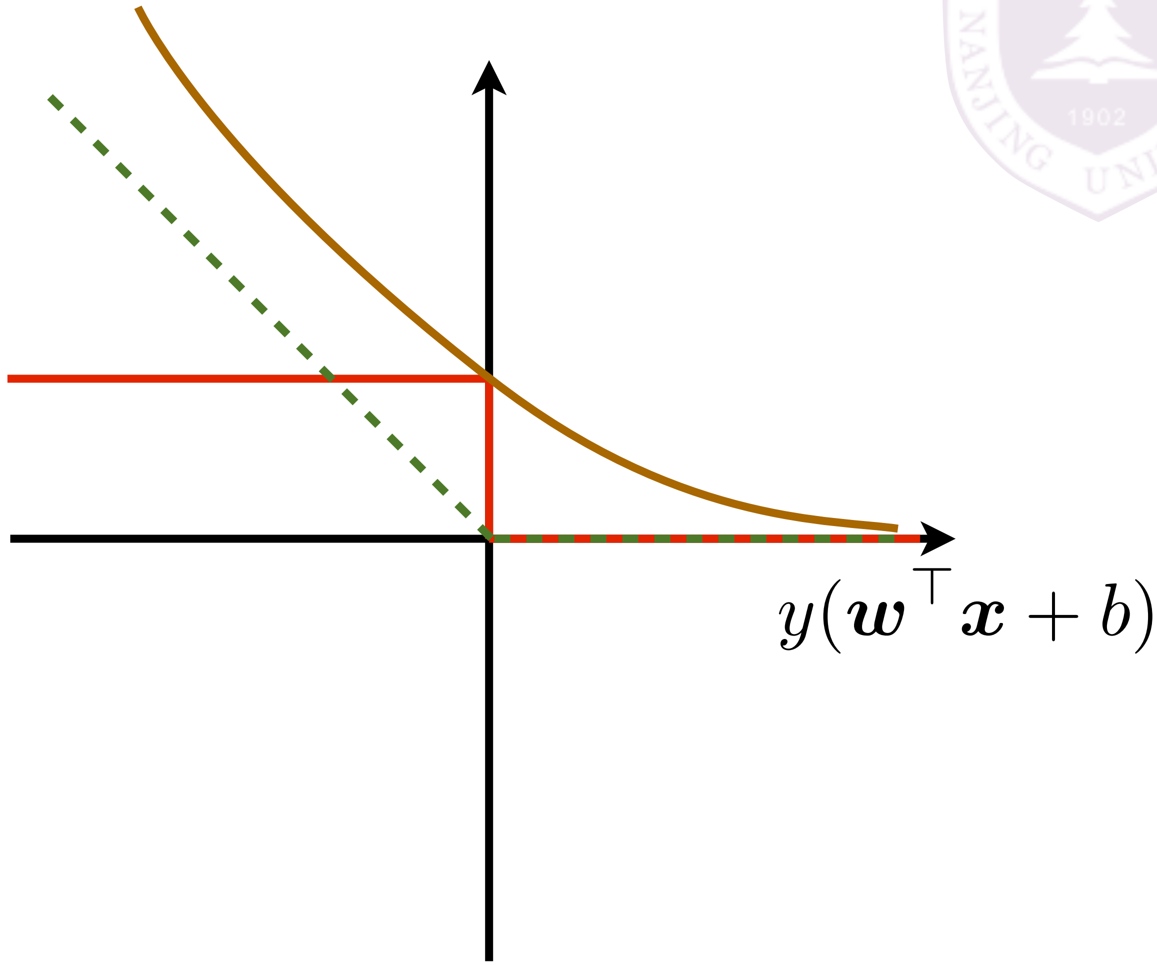
$I(y(\boldsymbol{w}^\top \boldsymbol{x} + b) \leq 0)$

logistic regression

$\log_2(1 + e^{-y(\boldsymbol{w}^\top \boldsymbol{x}+b)})$

perceptron

$\max\{-y(\boldsymbol{w}^\top \boldsymbol{x} + b), 0\}$



Linear classifier revisit

0-1 loss

$I(y(\boldsymbol{w}^\top \boldsymbol{x} + b) \leq 0)$

logistic regression

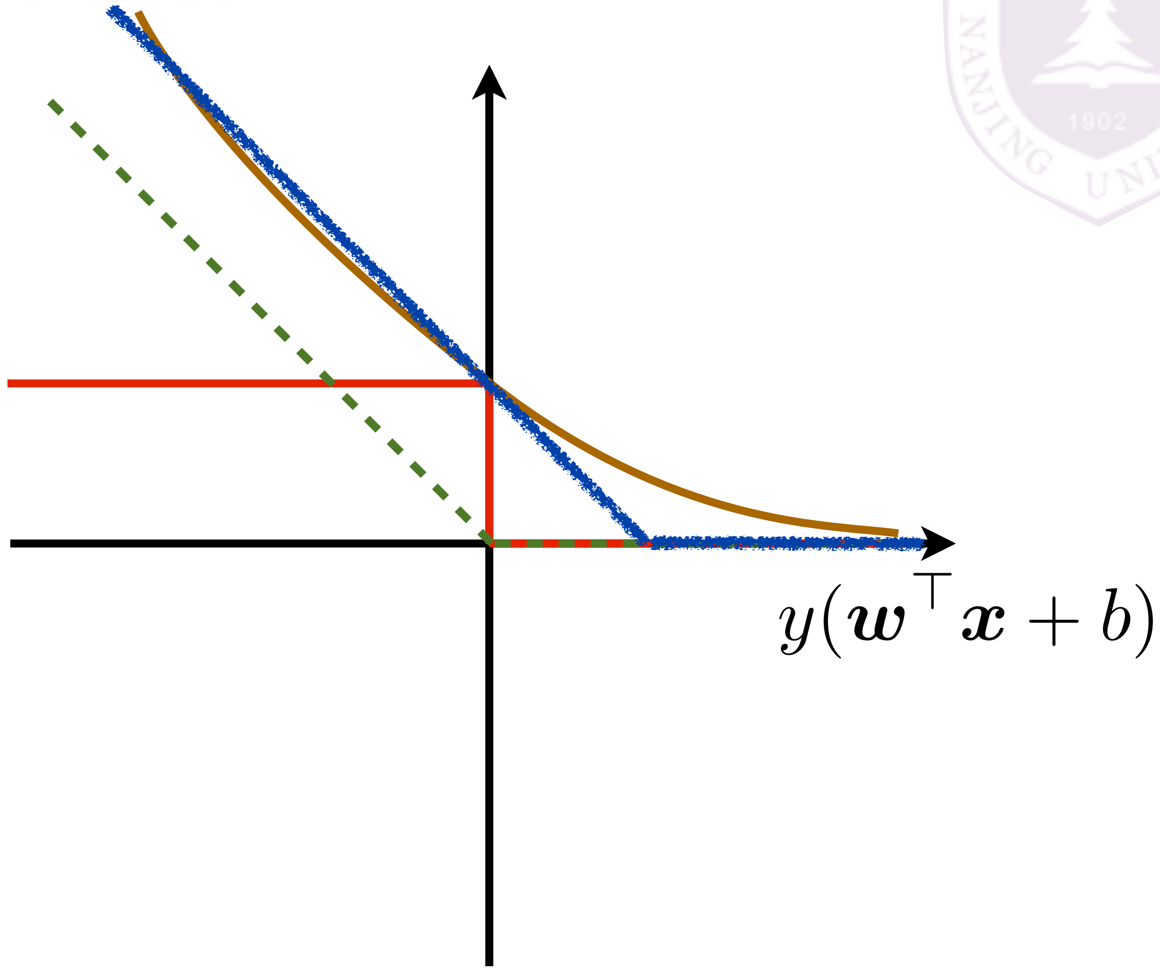
$\log_2(1 + e^{-y(\boldsymbol{w}^\top \boldsymbol{x} + b)})$

perceptron

$\max\{-y(\boldsymbol{w}^\top \boldsymbol{x} + b), 0\}$

hinge loss

$\max\{1 - y(\boldsymbol{w}^\top \boldsymbol{x} + b), 0\}$



Support vector machines (SVM)

hinge loss + L2-norm

$$\arg\min_{\boldsymbol{w},b} \sum_i \max(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0) + \lambda \|\boldsymbol{w}\|_2$$

Support vector machines (SVM)

hinge loss + L2-norm

$$\arg\min_{\boldsymbol{w},b} \sum_i \max(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0) + \lambda \|\boldsymbol{w}\|_2$$

$\max(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0) = \xi_i$
$\xi_i \geq 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)$
$\xi_i \geq 0$

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|_2 + C \sum_i \xi_i$$

$$s.t. \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i$$

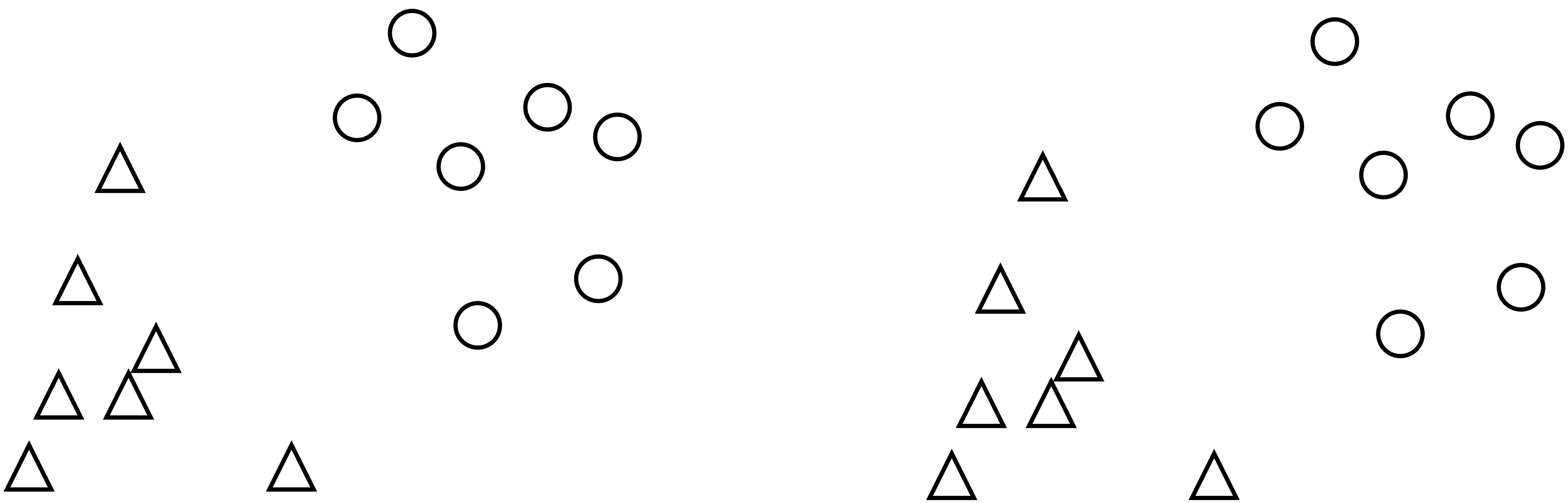
$$\xi_i \geq 0$$

quadratic

Support vector machines (SVM)

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2$$

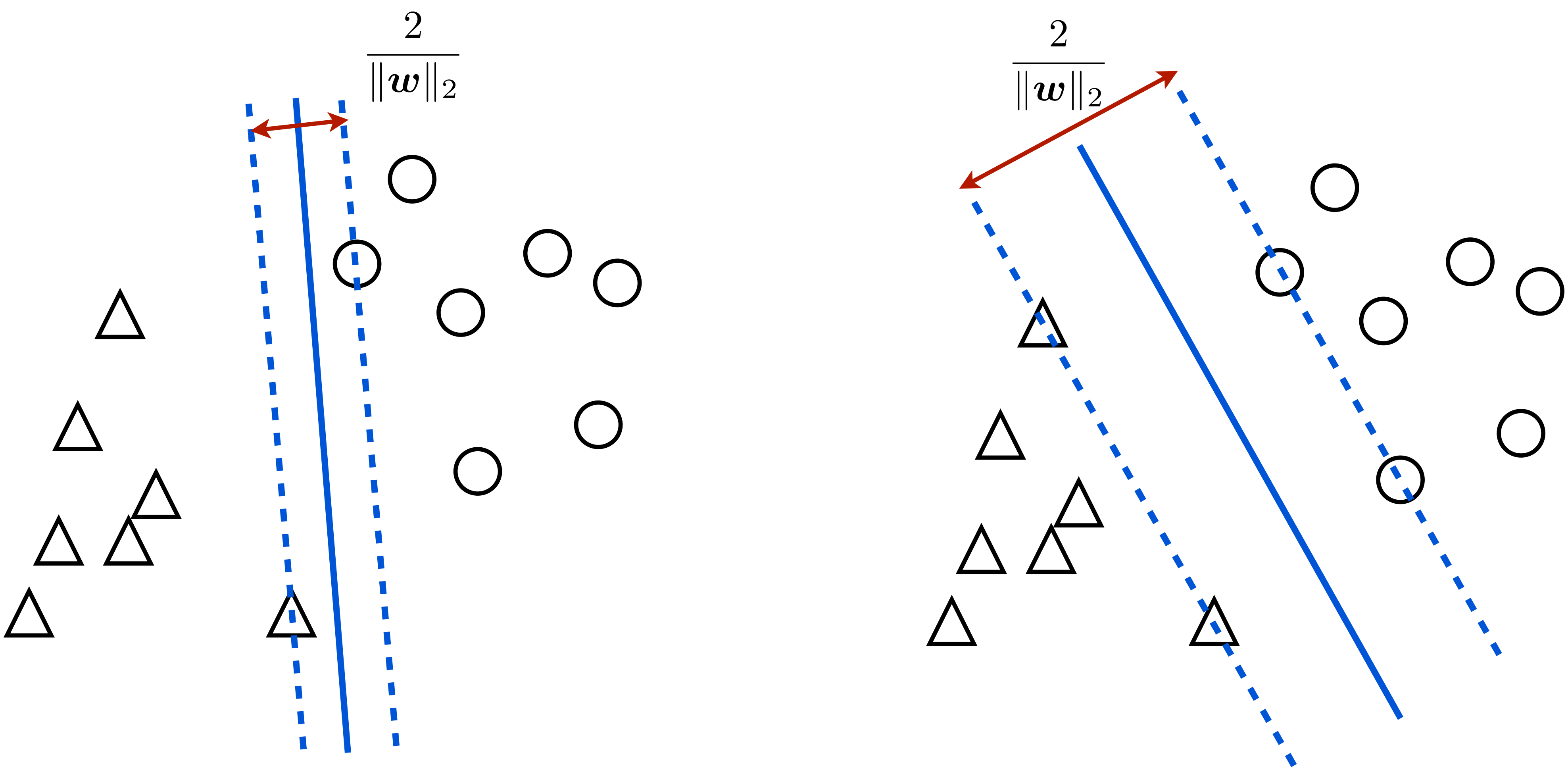
$$s.t. \qquad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$$



Support vector machines (SVM)

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2$$

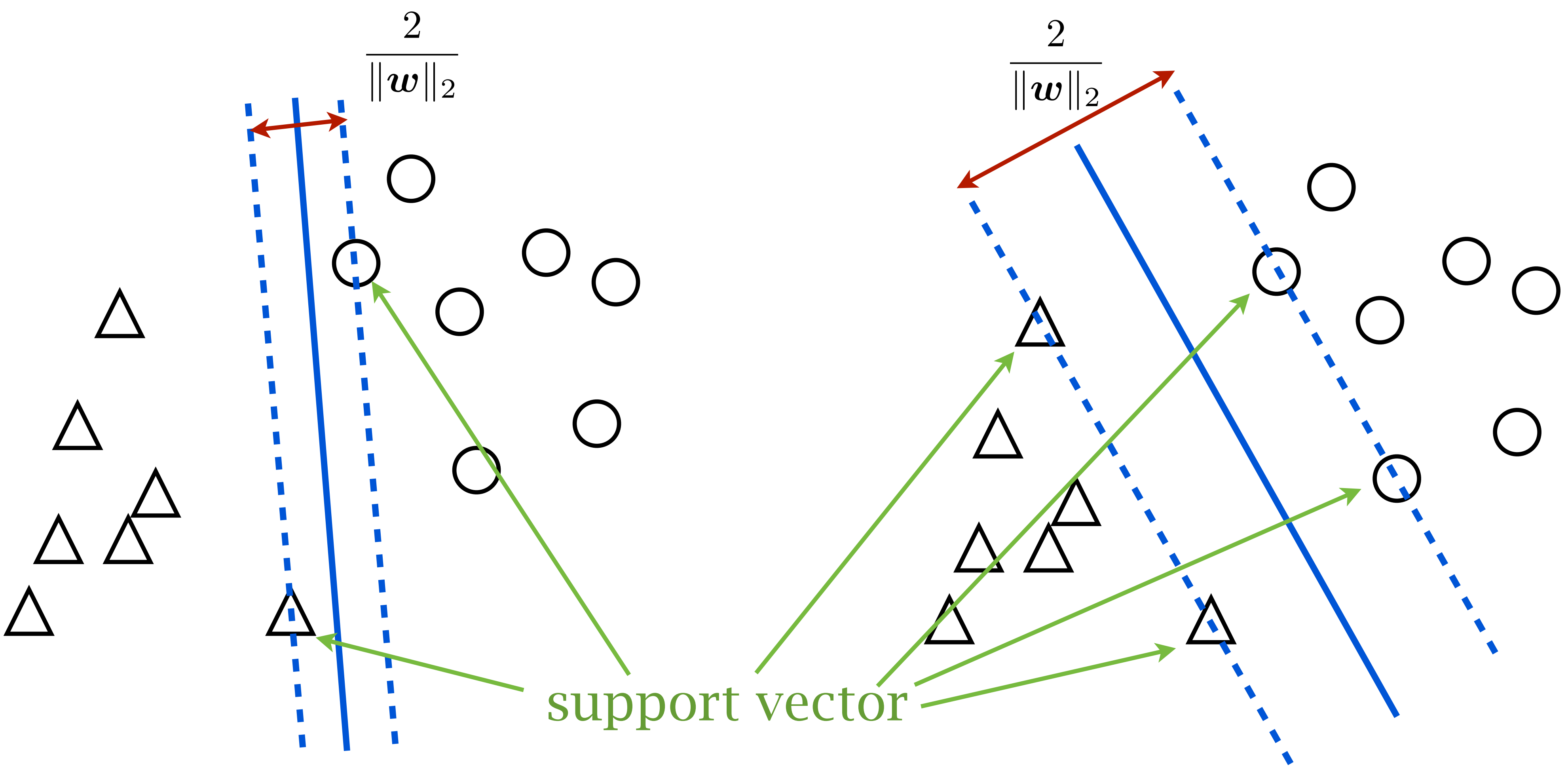
$$s.t. \qquad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$$



Support vector machines (SVM)

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2$$

$$s.t. \qquad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$$



Scoring functions

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2$$ least square regression

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i|$$ LAD regression

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2 + \lambda\|\boldsymbol{w}\|_2$$ ridge regression

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2 + \lambda\|\boldsymbol{w}\|_1$$ LASSO

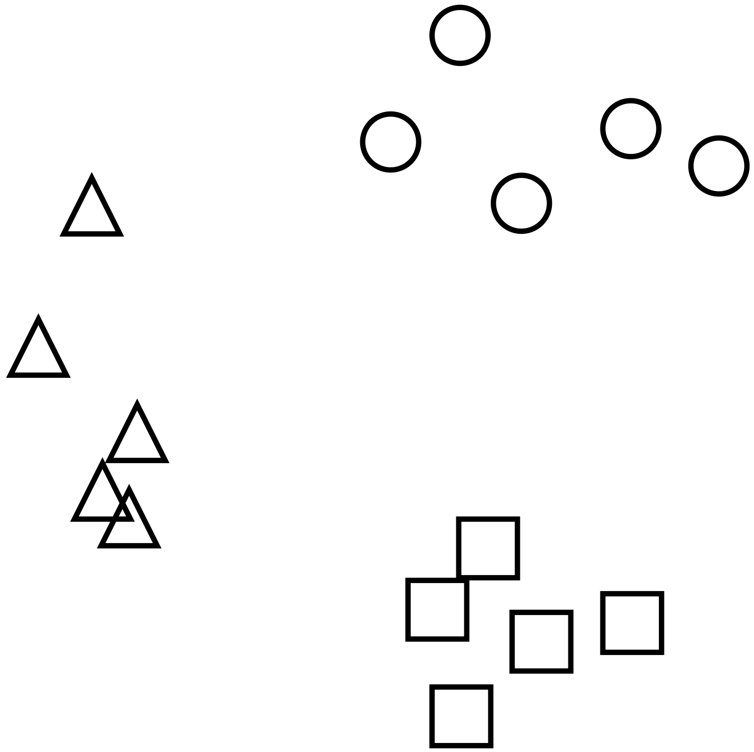
Scoring functions

$\sum_i I(y(\boldsymbol{w}^\top \boldsymbol{x} + b) > 0)$	0-1 loss
$\sum_i \max\{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0\}$	perceptron
$\sum_i \log\left(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}\right)$	logistic regression
$\sum_i \log\left(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}\right) + \lambda \|\boldsymbol{w}\|_2$	regularized LR
$\sum_i \max(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0) + \lambda \|\boldsymbol{w}\|_2$	SVM

minimize loss + regularization

Multi-class classification

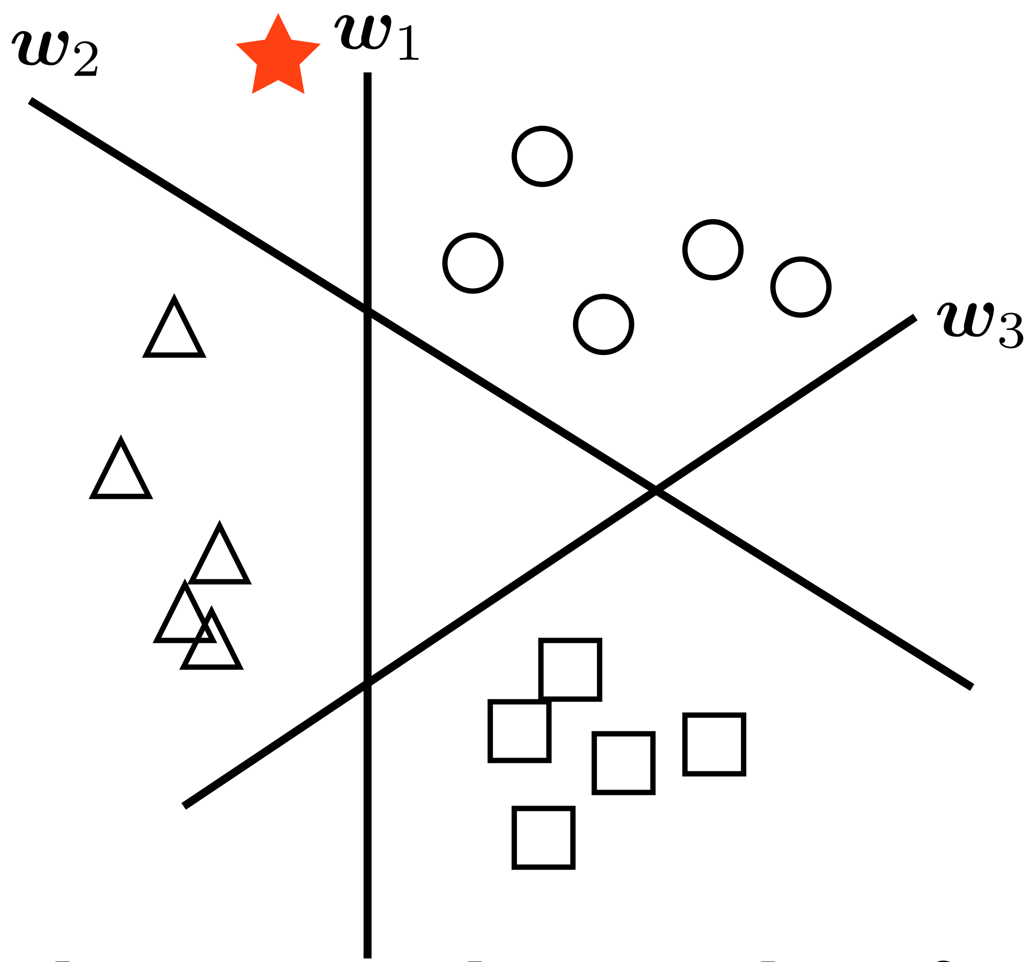
one-vs-rest



for C classes, need to train C binary classifiers

Multi-class classification

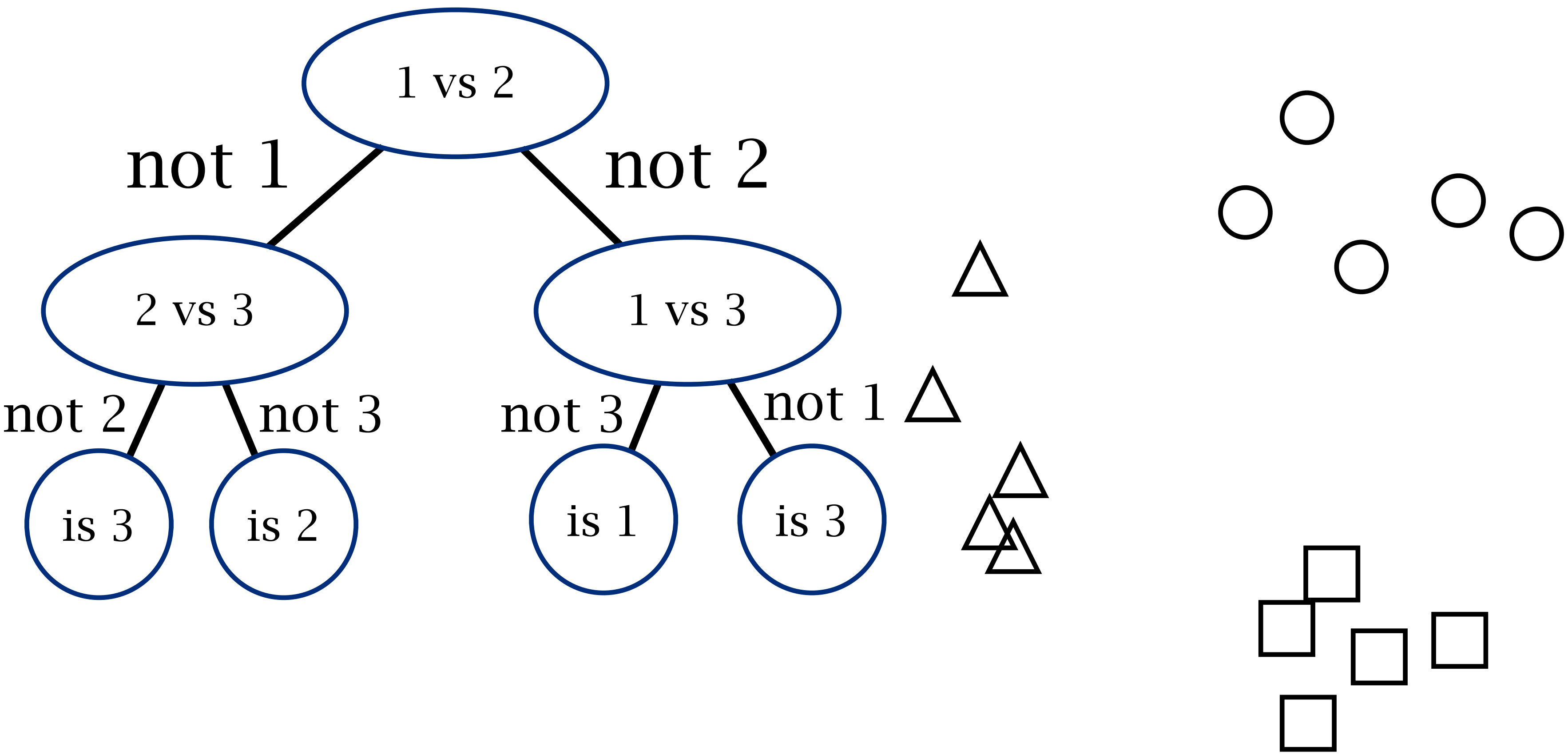
one-vs-rest



for C classes, need to train C binary classifiers

Multi-class classification

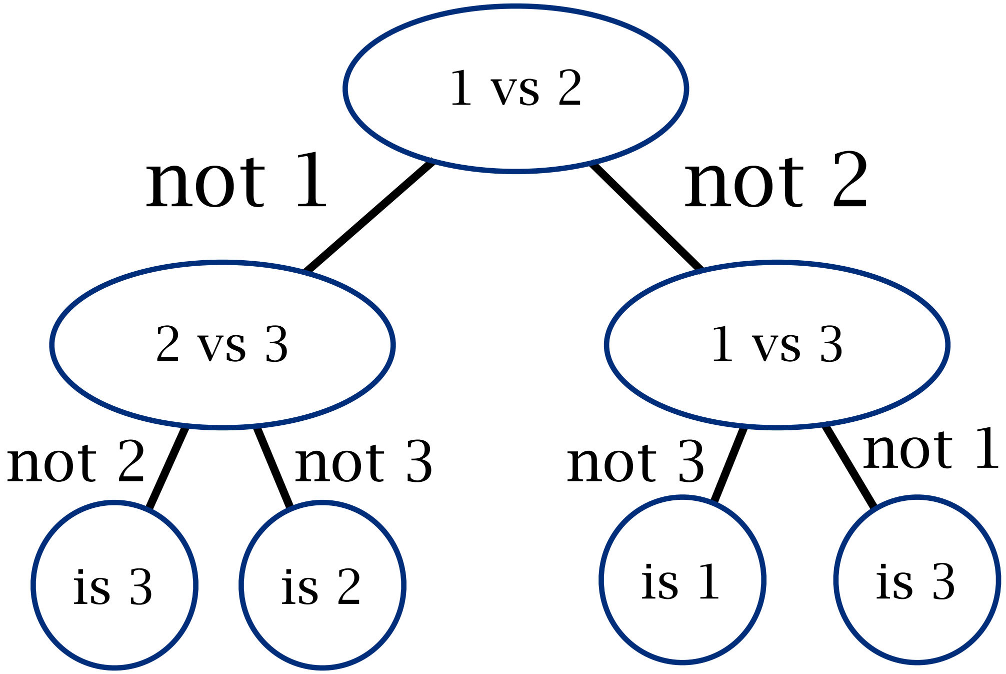
one-vs-one

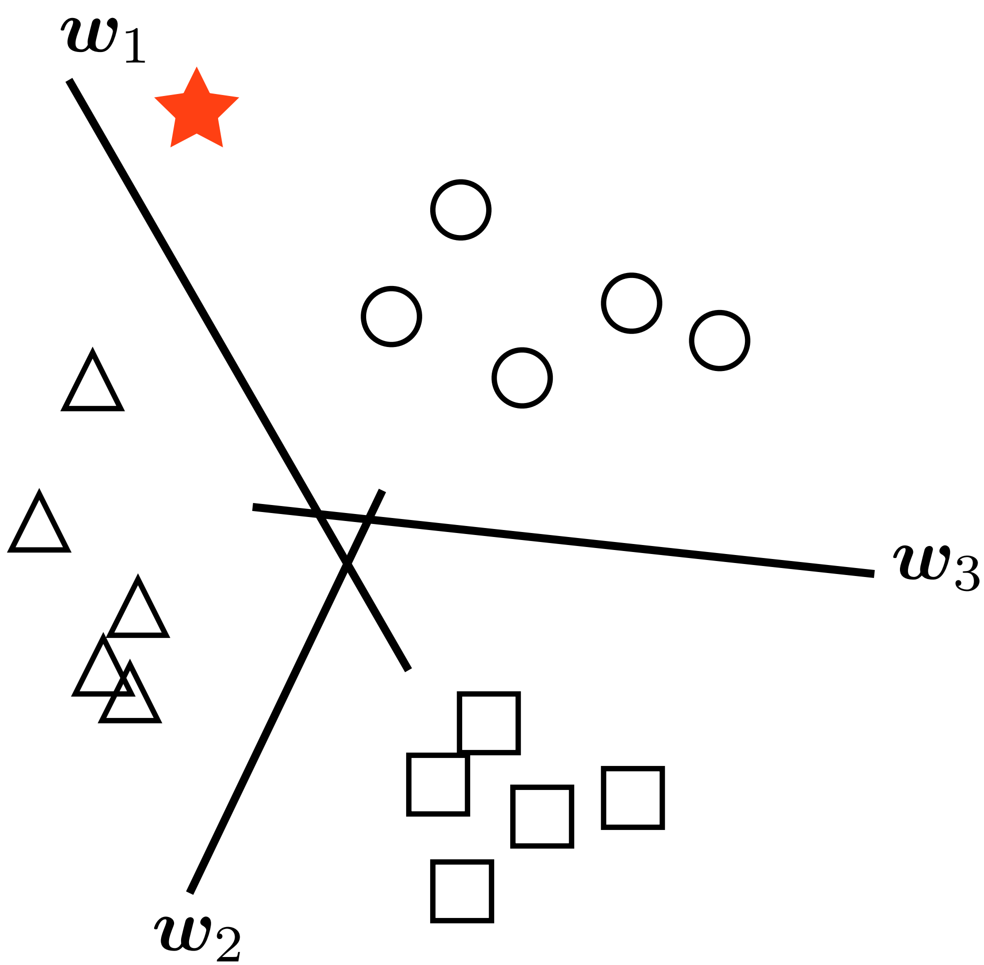


for C classes, need to train $C(C\text{-}1)/2$ binary classifiers

Multi-class classification

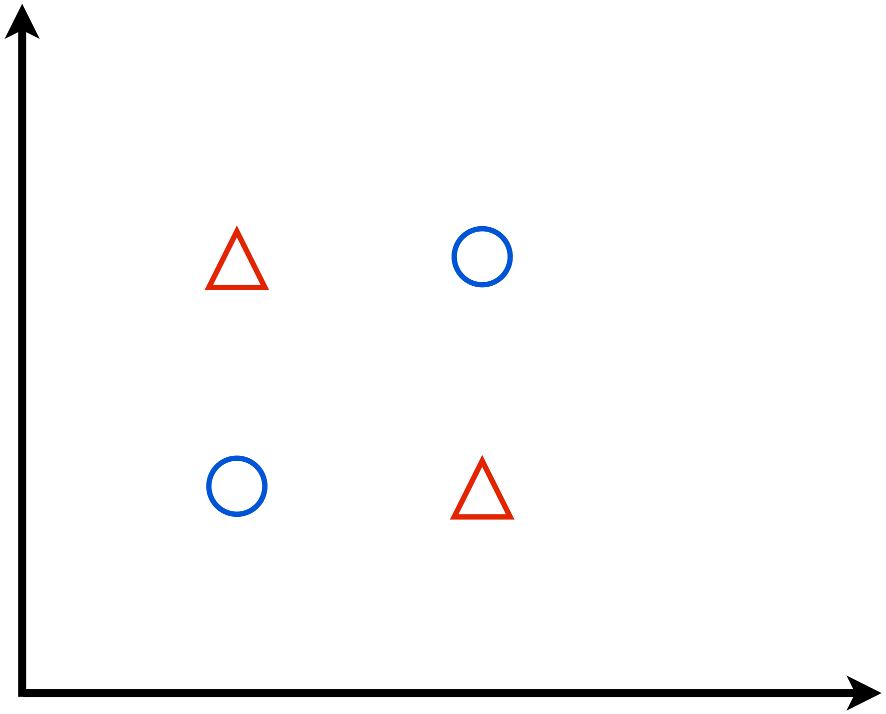
one-vs-one





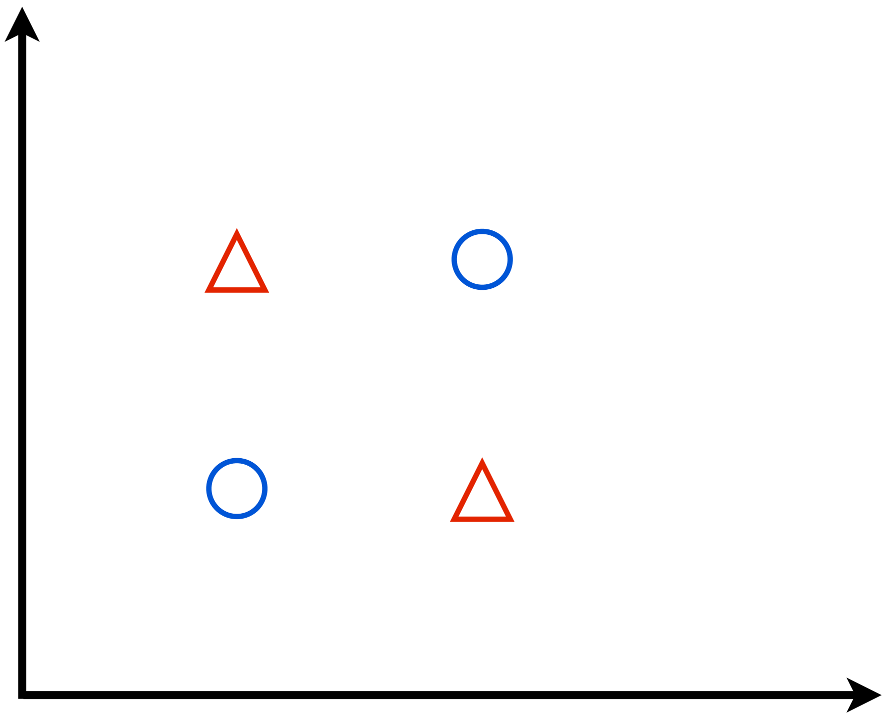
for C classes, need to train $C(C\text{-}1)/2$ binary classifiers

Linearity v.s. dimensionality

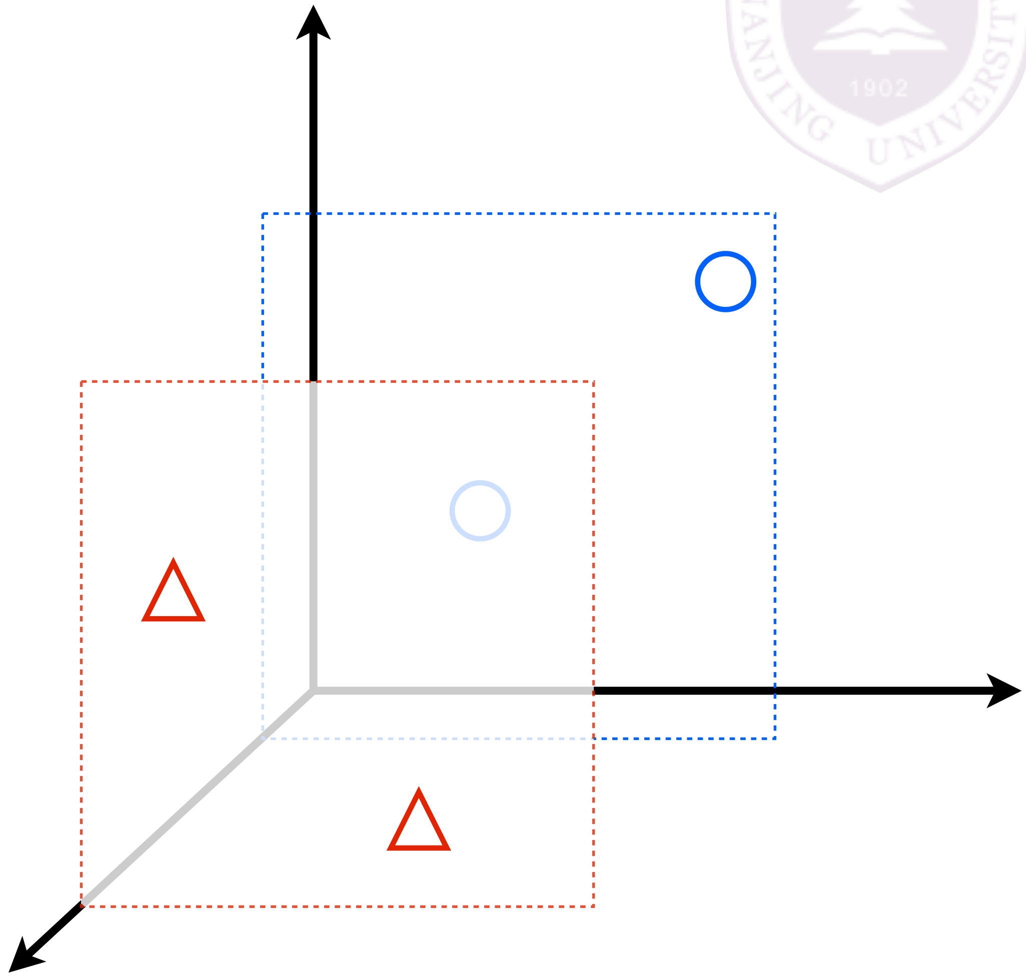


XOR in 2D

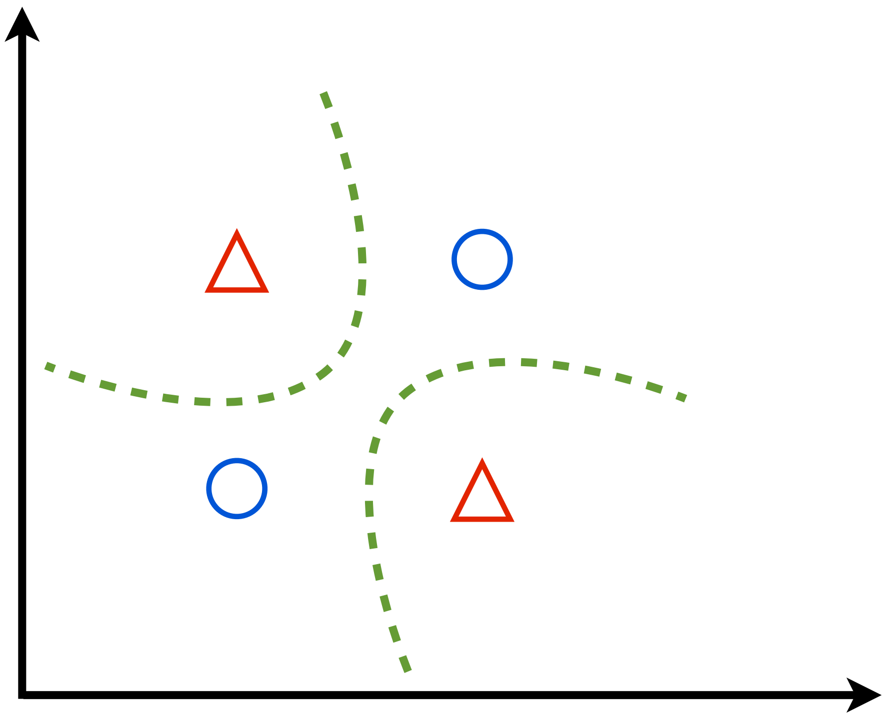
Linearity v.s. dimensionality



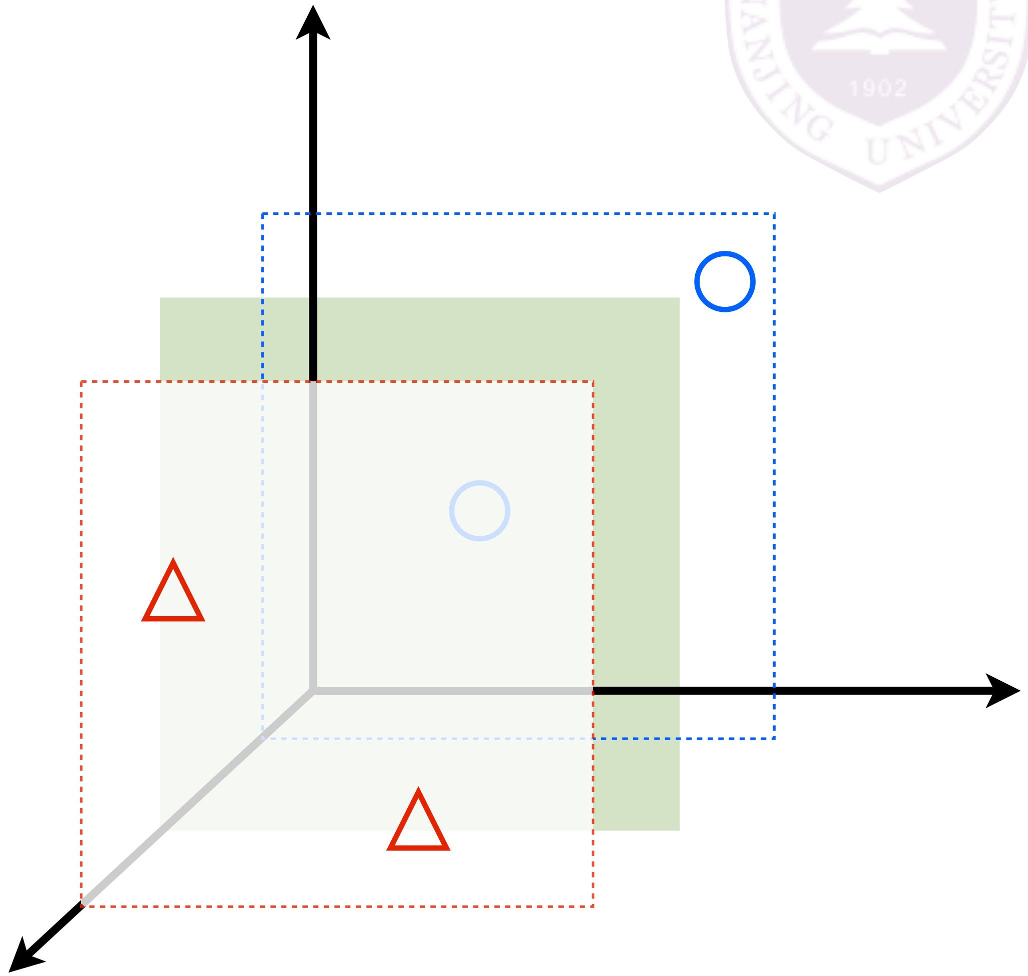
XOR in 2D



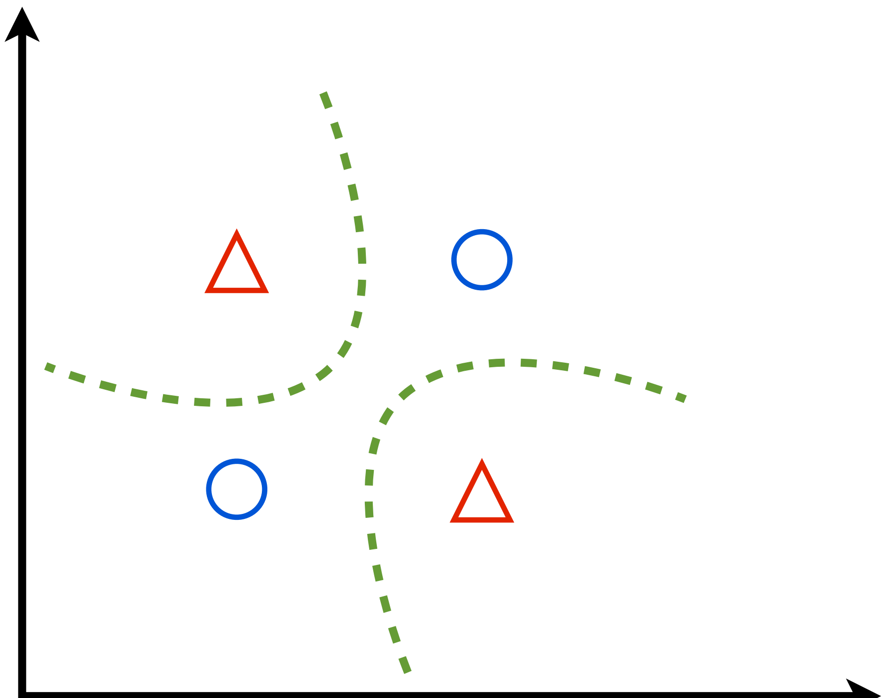
Linearity v.s. dimensionality



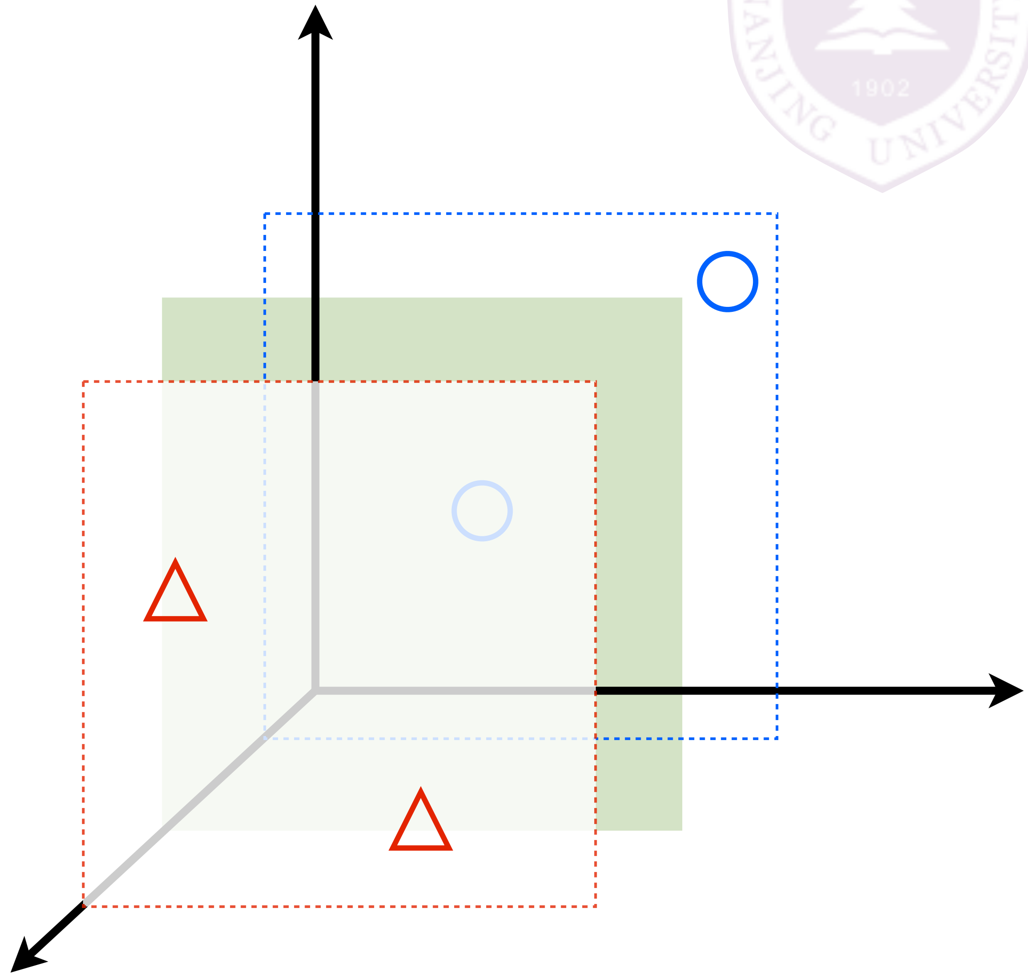
XOR in 2D



Linearity v.s. dimensionality



XOR in 2D



x_1	x_2	y
0	0	+1
0	1	-1
1	0	-1
1	1	+1

x_1	x_2	x_1x_2	y
0	0	0	+1
0	1	0	-1
1	0	0	-1
1	1	1	+1

$$\boldsymbol{w} = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}, b = -0.5$$