

支撑向量机与核方法

张煦尧(xyz@nlpr.ia.ac.cn)

2021年12月15日

History

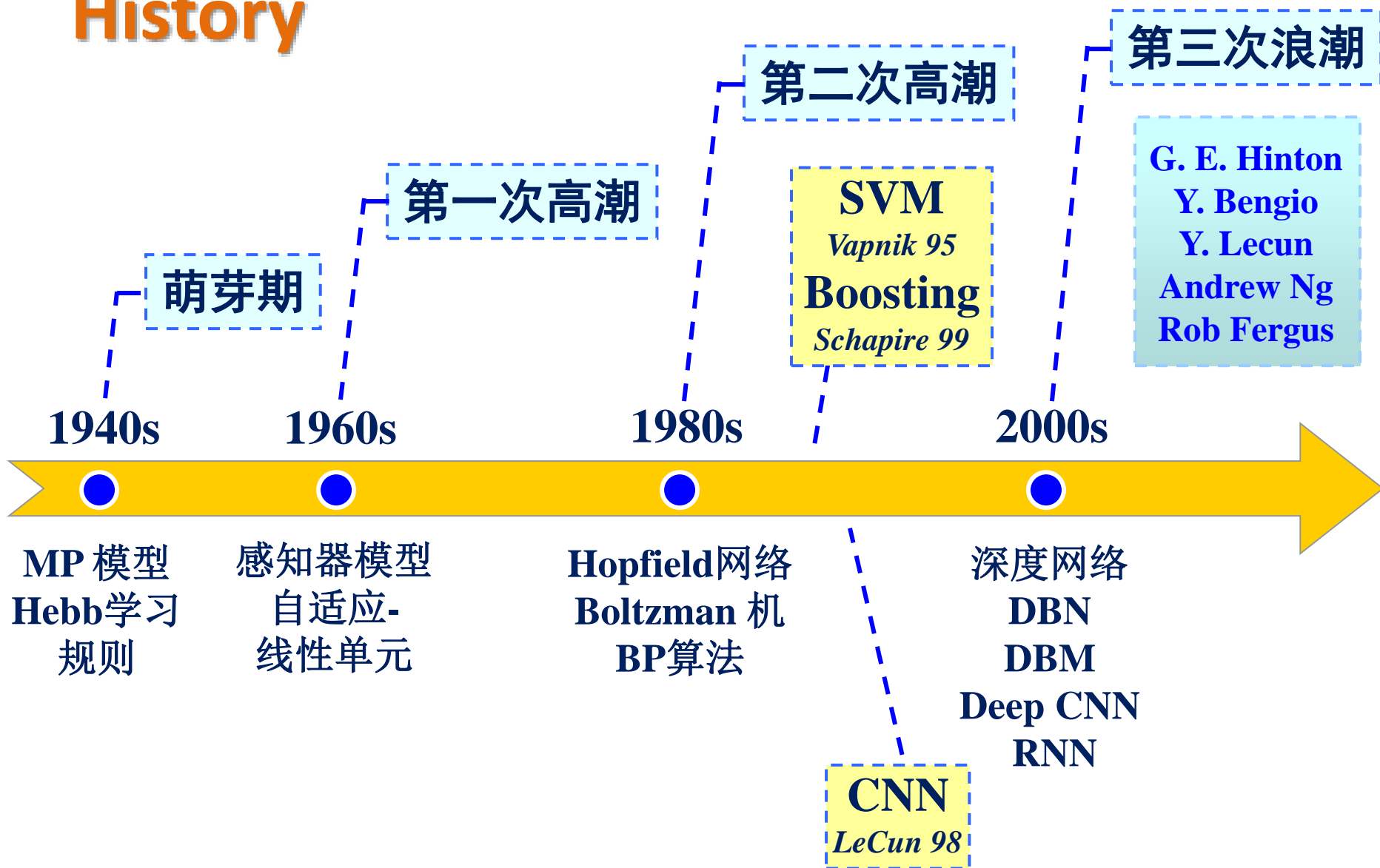
Boser, Guyon, Vapnik, [A training algorithm for optimal margin classifiers](#). COLT 1992.

Cortes, Vapnik, [Support-vector networks](#). Machine Learning, 1995.



- SVM is a classifier derived from statistical learning theory by [Vapnik and Chervonenkis](#)
- SVMs introduced by Boser, Guyon, Vapnik in [COLT-92](#)
- Initially popularized in the [NIPS community](#), now an important and active field of all machine learning research.
- Special issues of [Machine Learning Journal](#), and [Journal of Machine Learning Research](#).

History



History (1995-2005, 2006-now)

正方: Jackel: 1995-2000 五年时间从理论上弄懂神经网络

1. Jackel bets (one fancy dinner) that by March 14, 2000, people will understand quantitatively why big neural nets working on large databases are not so bad. (Understanding means that there will be clear conditions and bounds)

Vapnik bets (one fancy dinner) that Jackel is wrong.

But .. If Vapnik figures out bet.

赌约: one fancy dinner

2. Vapnik bets (one fancy dinner) that by March 14, 2005, no one in his right mind will use neural nets that are essentially like those used in 1995.

Jackel bets (one fancy dinner)

反方: Vapnik: 1995-2005 十年时间也无法弄清楚

V. Vapnik

3/14/95

L. Jackel

3/14/95

Witnessed by Y. LeCun

3/14/95

见证人: Y. LeCun

1995年的赌约

2017 NIPS 风波

Ali Rahimi (阿里·拉希米)

NIPS 2017 "Test of Time" Award

《Random Features for Large-Scale Kernel Machines》



History (1995-2005, 2006-now)

Science
2006

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such “autoencoder” networks, but this works well only if

Nature
2015

Deep learning

Yann LeCun^{1,2}, Yoshua Bengio³ & Geoffrey Hinton^{4,5}

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine

Nature
2016

Mastering the game of Go with deep neural networks and tree search

David Silver^{1*}, Aja Huang^{1*}, Chris J. Maddison¹, Arthur Guez¹, Laurent Sifre¹, George van den Driessche¹, Julian Schrittwieser¹, Ioannis Antonoglou¹, Veda Panneershelvam¹, Marc Lanctot¹, Sander Dieleman¹, Dominik Grewe¹, John Nham², Nal Kalchbrenner¹, Ilya Sutskever², Timothy Lillicrap¹, Madeleine Leach¹, Koray Kavukcuoglu¹, Thore Graepel¹ & Demis Hassabis¹

SVM Theory

Structural Risk Minimization

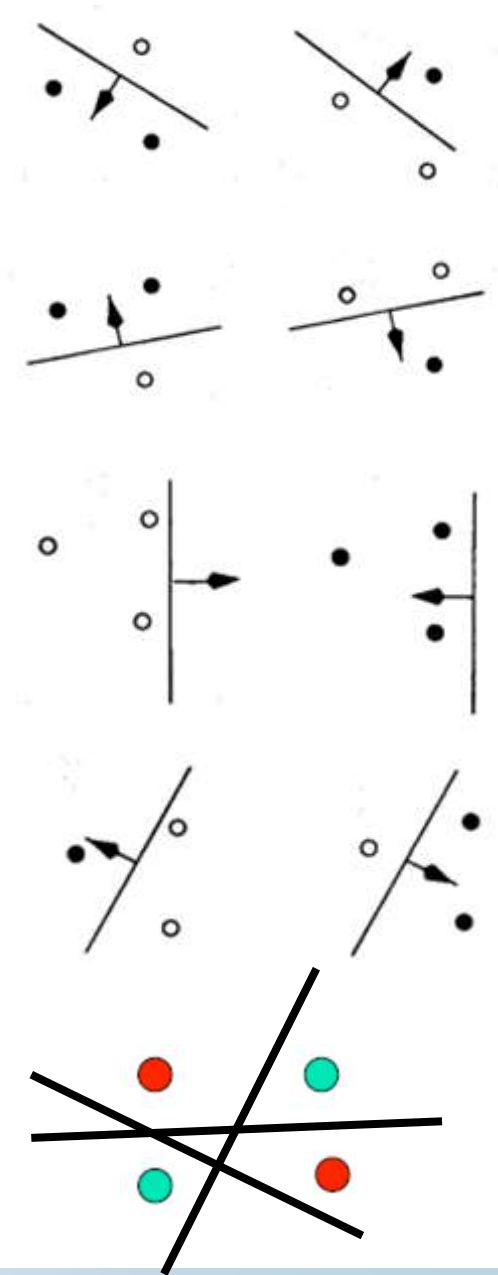
- We want to get a low error rate on unseen data.
 - This is called “structural risk minimization”
 - Training error is “empirical risk minimization”
- It would be really helpful if we could get a guarantee:
$$\text{Test error rate} \leq \text{train error rate} + f(N, h, p)$$

Where N = size of training set,
 h = measure of the model complexity,
 p = the probability that this bound fails

We need p to allow for really unlucky test sets.
- Then we could choose the model complexity that minimizes the bound on the test error rate.

VC Dimension

- Pick n datapoints
- Assign labels of + or – to them at random
- If our model (e.g. a neural net with a certain number of hidden units) is powerful enough to learn **any** association of labels with the data, its too powerful!
- Characterize the power of a model class by asking **how many datapoints** it can “shatter”
- **learn perfectly for all possible assignments of labels.**
- This number of datapoints is called the **Vapnik-Chervonenkis dimension**.
- VC dimension for a 2D plane?
 - In 2-D, we can find a plane (i.e. a line) to deal with any labeling of three points. A 2-D hyperplane shatters 3 points



VC Dimension

- The VC dimension of a hyperplane in 2-D is 3.
 - In k dimensions it is $k+1$.
- It is just a **coincidence** that the VC dimension of a hyperplane is almost identical to the number of parameters it takes to define a hyperplane.
- A **sine wave has infinite VC dimension and only 2 parameters!**
 - By choosing the phase and period carefully we can shatter any random collection of one-dimensional datapoints
 - Let $x_i = 10^i$ where i ranges from 1 to n . The classifier $y = \text{sign}(\sin(\alpha x))$ can classify all x_i correctly for all possible combination of class labels on x_i

$$f(x) = b \sin(ax)$$



VC Dimension

- The VC-dimension of the nearest neighbor classifier is infinity, because no matter how many points you have, you get perfect classification on training data
 - 1-NN \rightarrow K-NN: reduce VC dimension
- The higher the VC-dimension, the more flexible a classifier is
- VC-dimension, however, is a theoretical concept
- VC-dimension of most classifiers, in practice, is difficult to be computed exactly
- Qualitatively, if we think a classifier is flexible, it probably has a high VC-dimension

The Probabilistic Guarantee

$$E_{test} \leq E_{train} + \left(\frac{h + h \log(2N / h) - \log(p / 4)}{N} \right)^{\frac{1}{2}}$$

where N = size of training set

h = VC dimension of the model class

p = upper bound on probability that this bound fails

Good generalization \rightarrow large N and small h

So if we train models with different complexity, we should pick the one that minimizes this bound

Actually, this is only sensible if we think the bound is fairly tight, which it usually isn't. The theory provides insight, but in practice we still need some witchcraft. --- G. Hinton

Large Margin and VC Dimension

- If we use a large set of non-adaptive features, we can often make the two classes linearly separable.
 - But if we just fit any separating plane, it will not generalize well to new cases.
- If we fit the separating plane that **maximizes the margin** (the minimum distance to any of the data points), we will get much better generalization.
 - Intuitively, by maximizing the margin we are **squeezing out all the surplus capacity** that came from using a high-dimensional feature space.
- This can be justified by a whole lot of clever mathematics which shows that
 - large margin separators have lower VC dimension.
 - models with lower VC dimension have a smaller gap between the training and test error rates.

Some Philosophy



- **Occam's Razor** (奥卡姆剃刀原理)
 - Entities should not be multiplied unnecessarily
 - 如无必要，勿增实体
 - All other things being equal, the simplest solution is the best



- **Vapnik's principle**
 - never to solve a problem that is more general than you actually need to solve



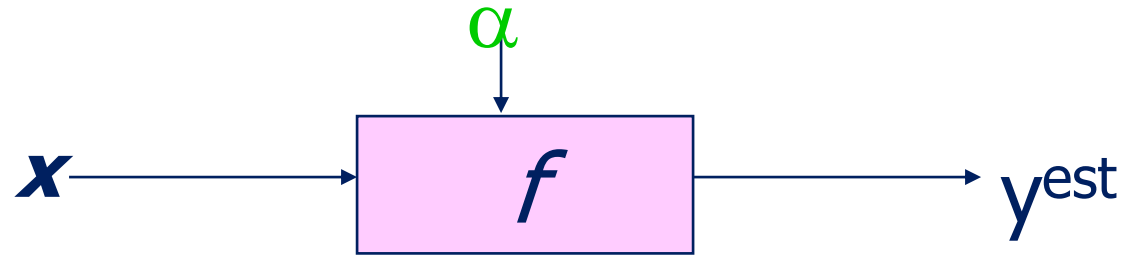
- **牛顿**
 - 如果某一原因既真又足以解释自然事物的特性，则我们不当接受比这更多的原因

Some Philosophy

- **Simplest pattern classifier ? 最简单的模式分类器 ?**
 - Binary two-class problem
 - Linear classifier
- **Which linear classifier?**
 - The large margin one
- **Extension for nonlinear case**
 - Any function can be a linear function if transformed to a high-dimension space
 - **Kernel method**: the inner product in high-dimension space
 - Almost all kernel methods (including SVM) adopt the linear classifier as its initial study objective
- **Extension for multi-class case**
 - 1vsAll, 1vs1 ...

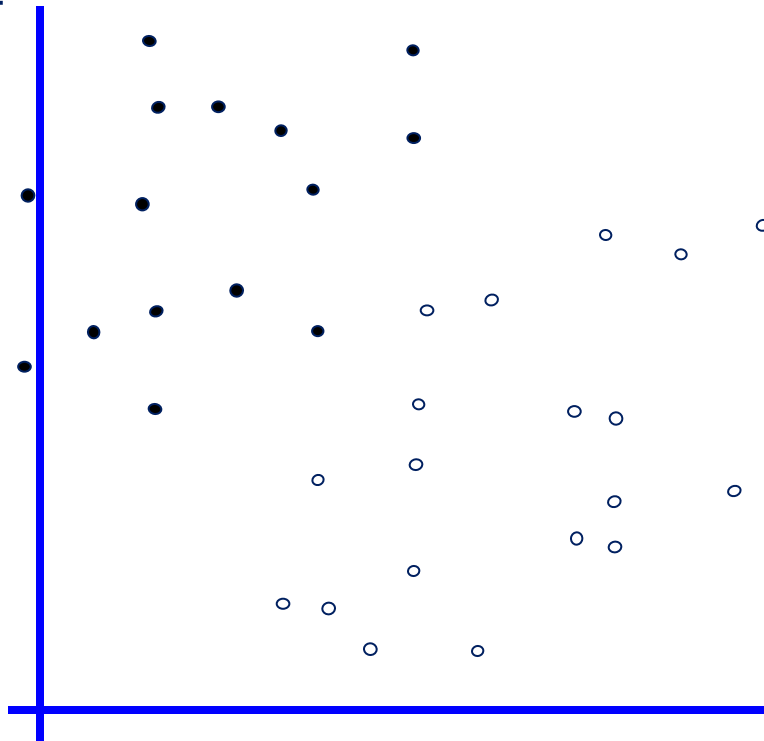
Hard-Margin SVM

Linear Classifiers



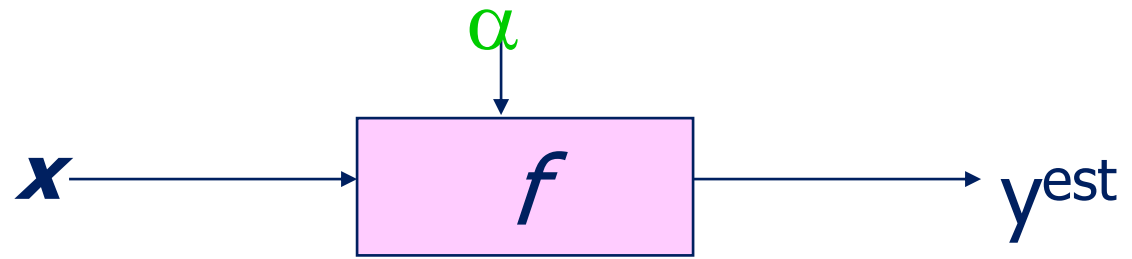
$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

- denotes +1
- denotes -1

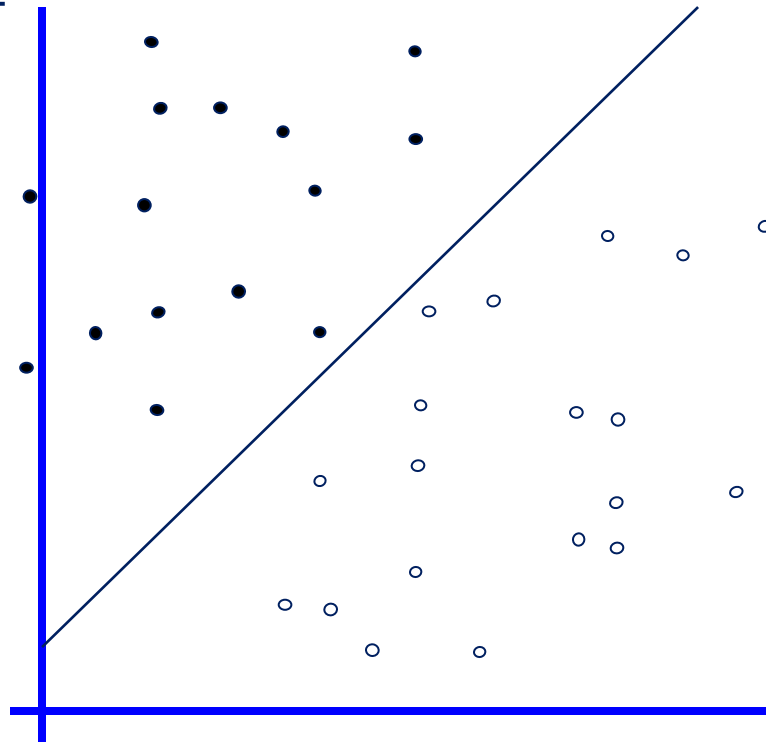


How would you classify this data?

Linear Classifiers



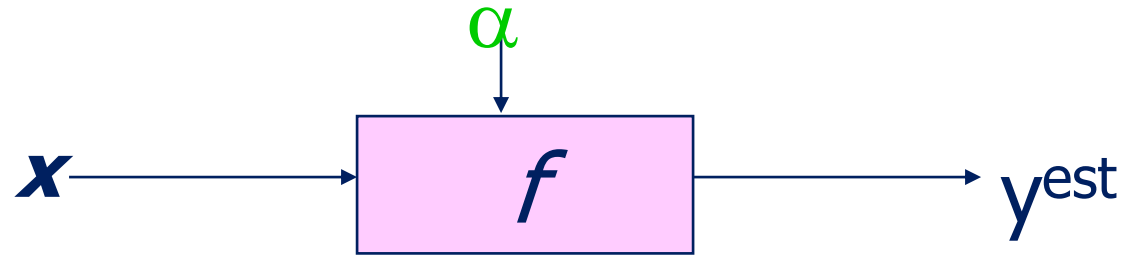
- denotes +1
- denotes -1



$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

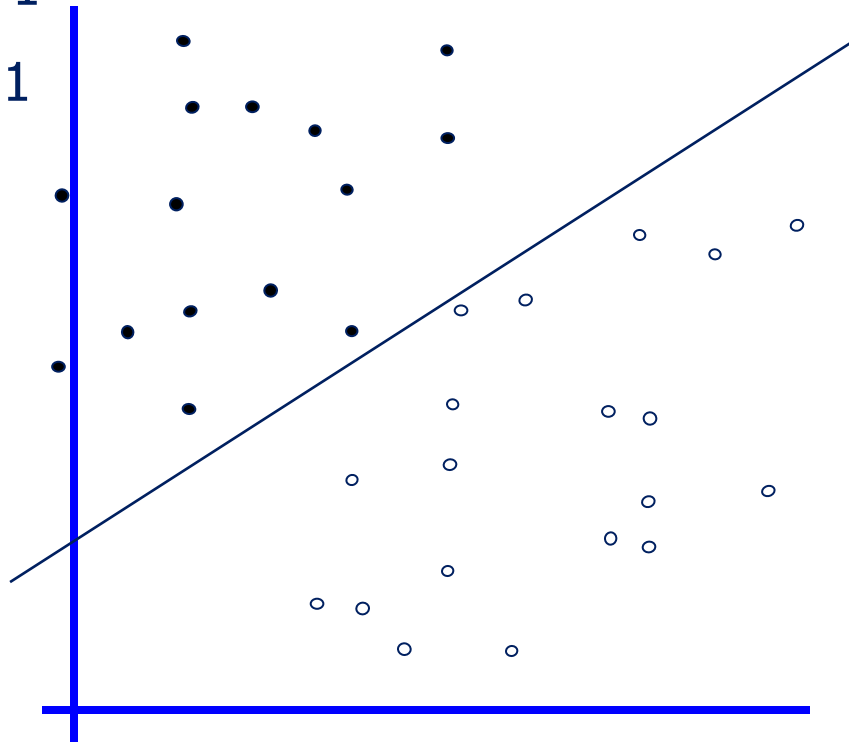
How would you classify this data?

Linear Classifiers



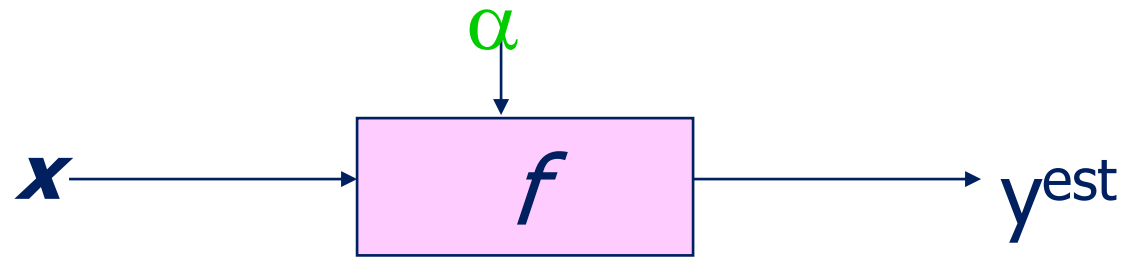
$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

- denotes +1
- denotes -1

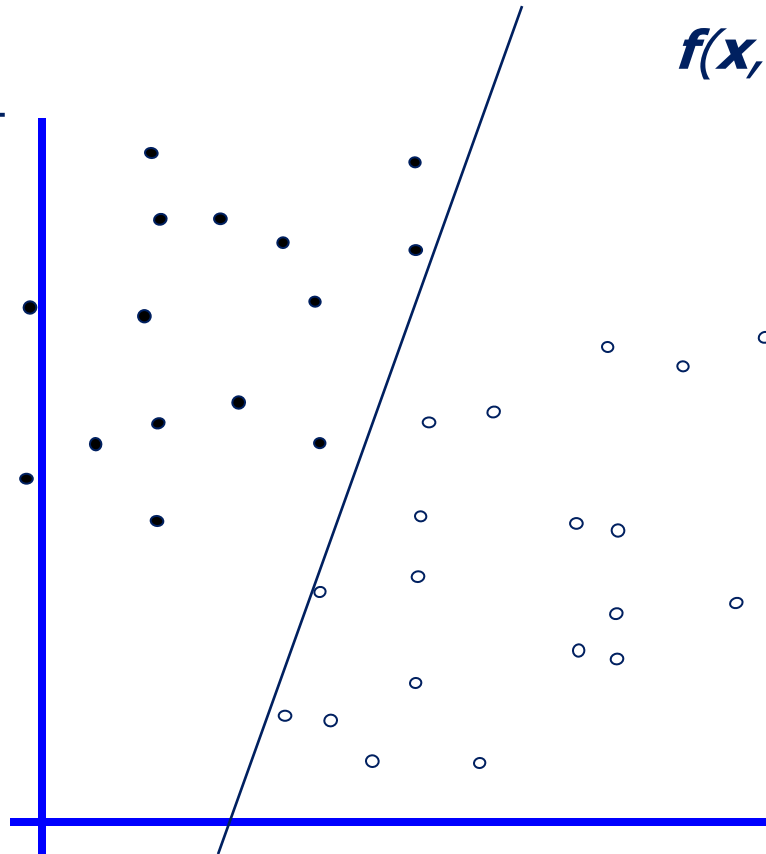


How would you classify this data?

Linear Classifiers



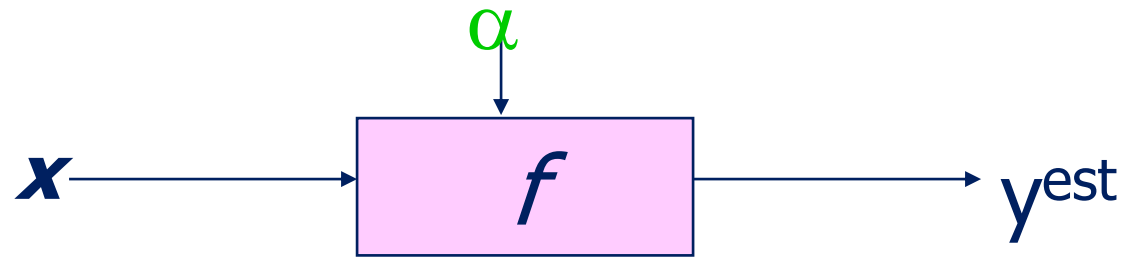
- denotes +1
- denotes -1



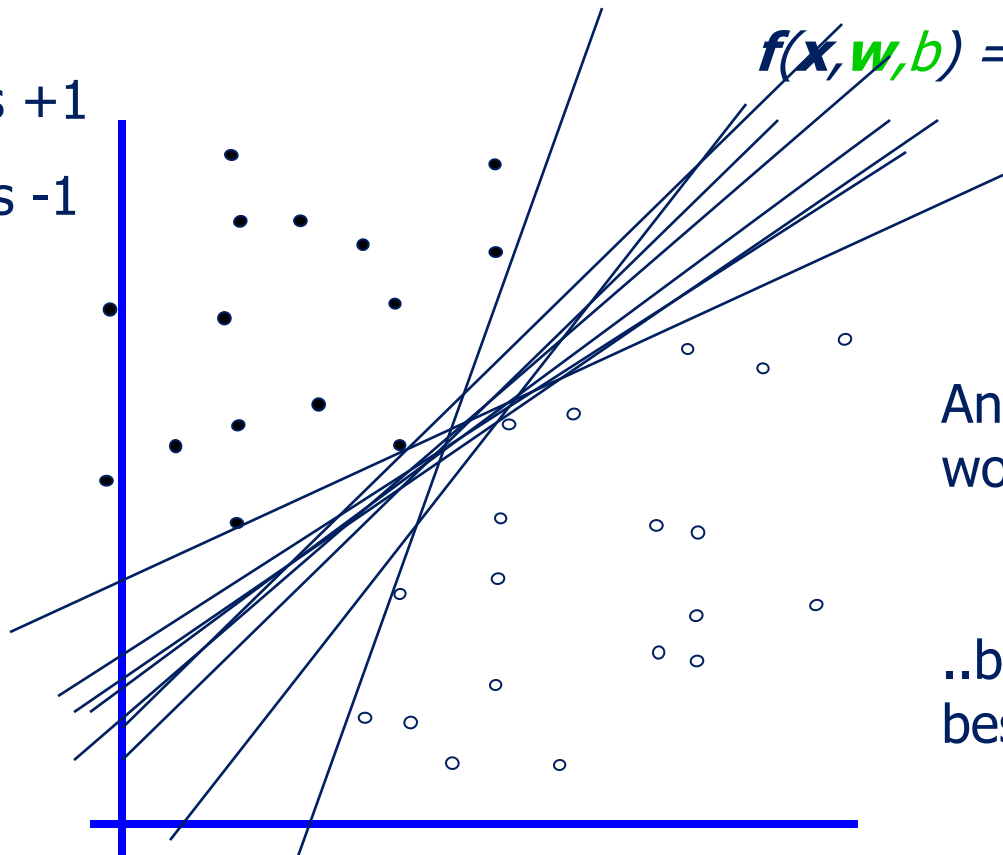
$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

How would you classify this data?

Linear Classifiers



- denotes +1
- denotes -1

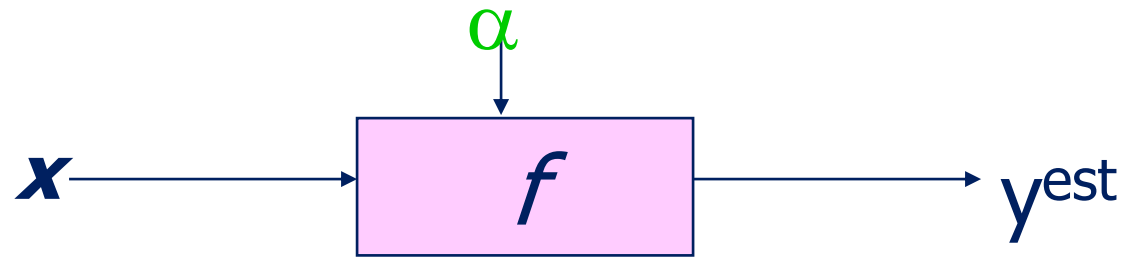


$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

Any of these
would be fine..

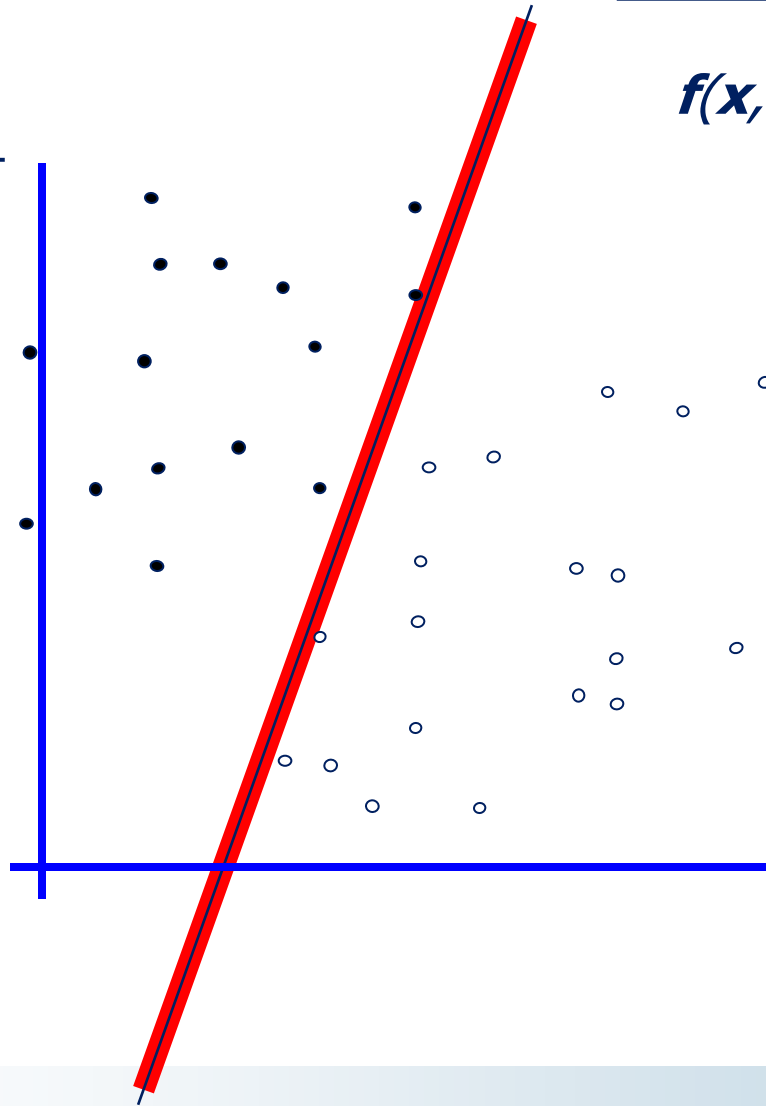
..but which is
best?

Classifier Margin



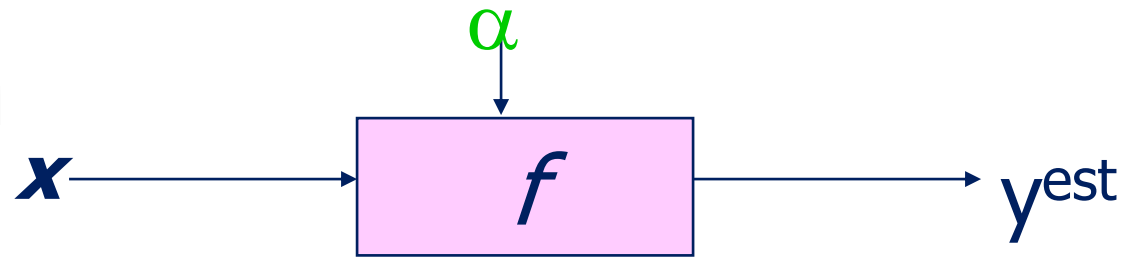
$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

- denotes +1
- denotes -1

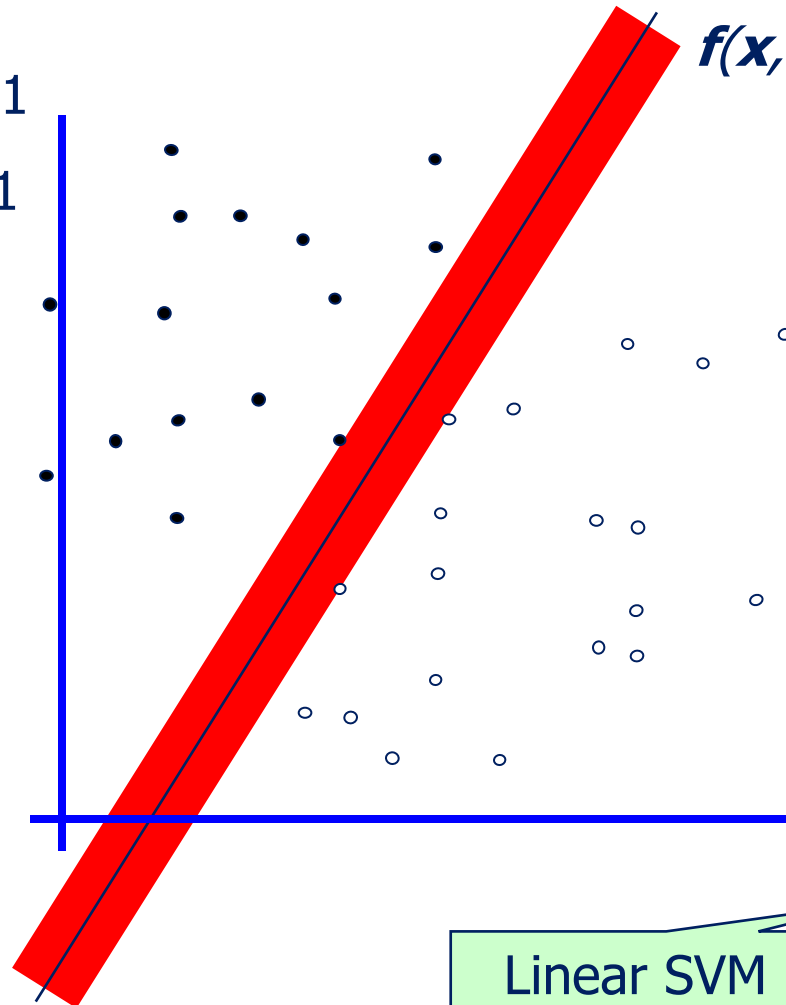


Define the **margin** of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

Maximum Margin



- denotes +1
- denotes -1



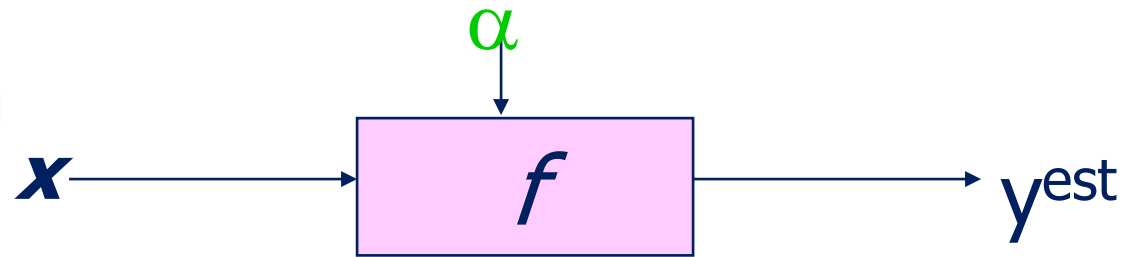
$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

The maximum margin linear classifier is the linear classifier with the maximum margin.

This is the simplest kind of SVM (Called an LSVM)

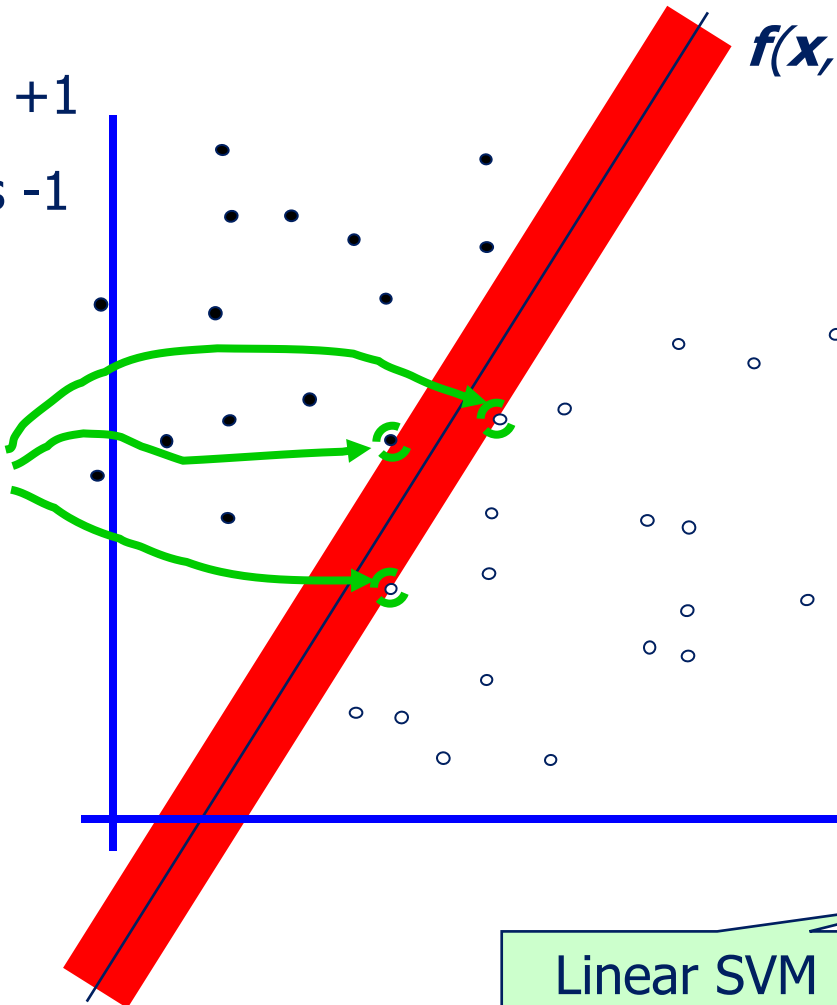
Linear SVM

Maximum Margin



- denotes +1
- denotes -1

Support Vectors
are those
datapoints that
the margin
pushes up
against



$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

The maximum margin linear classifier is the linear classifier with the maximum margin.

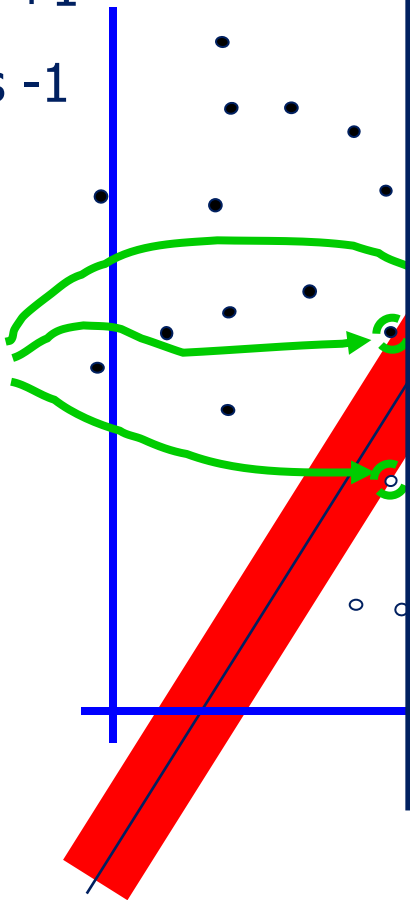
This is the simplest kind of SVM (Called an LSVM)

Linear SVM

Why Maximum Margin?

- denotes +1
- denotes -1

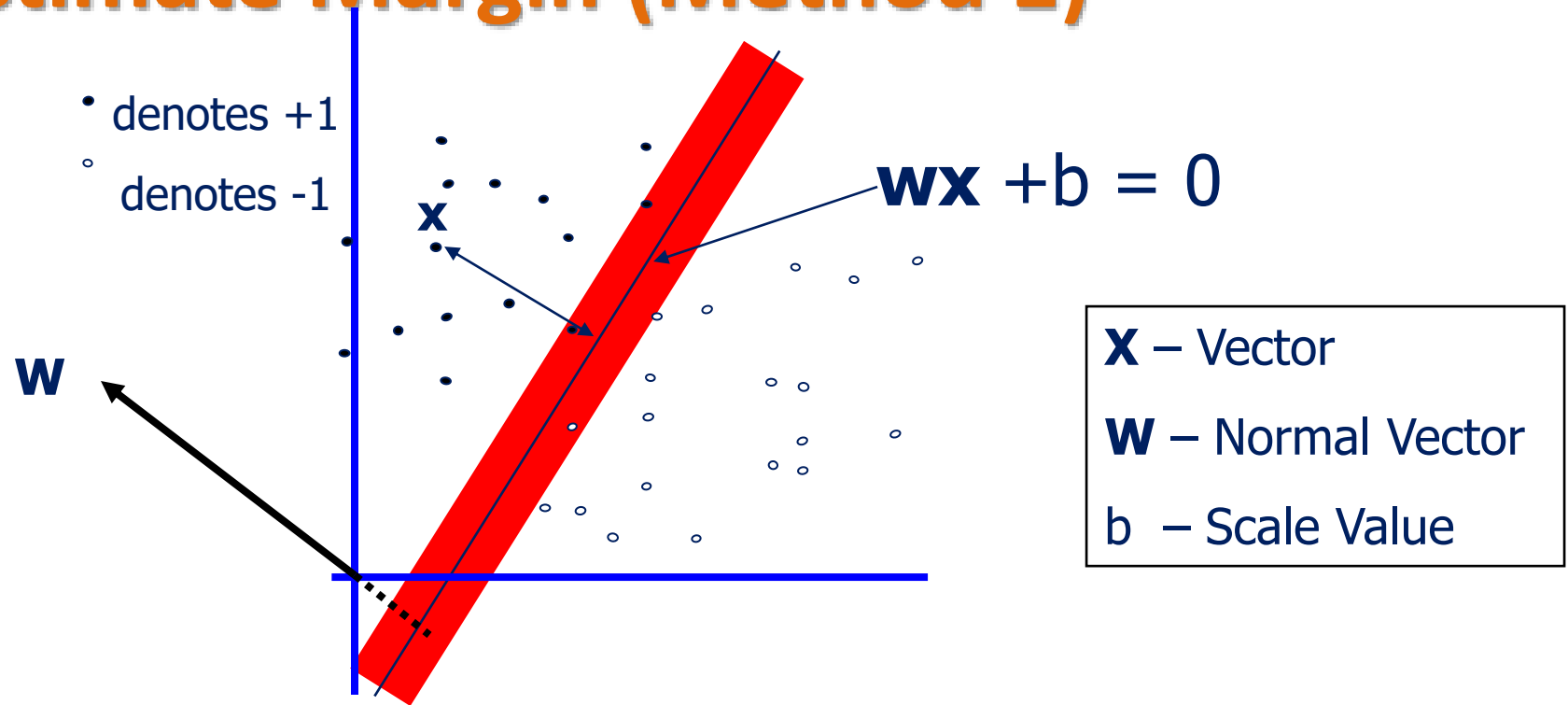
Support Vectors
are those
datapoints that
the margin
pushes up
against



- 1. Intuitively this feels safest.**
- 2. Empirically it works very well.**
3. If we've made a small error in the location of the boundary (it's been jolted in its perpendicular direction) this gives us least chance of causing a misclassification.
4. LOOCV is easy since the model is immune to removal of any non-support-vector datapoints.
5. There's some theory (using VC dimension) that this is a good thing.

ESV117

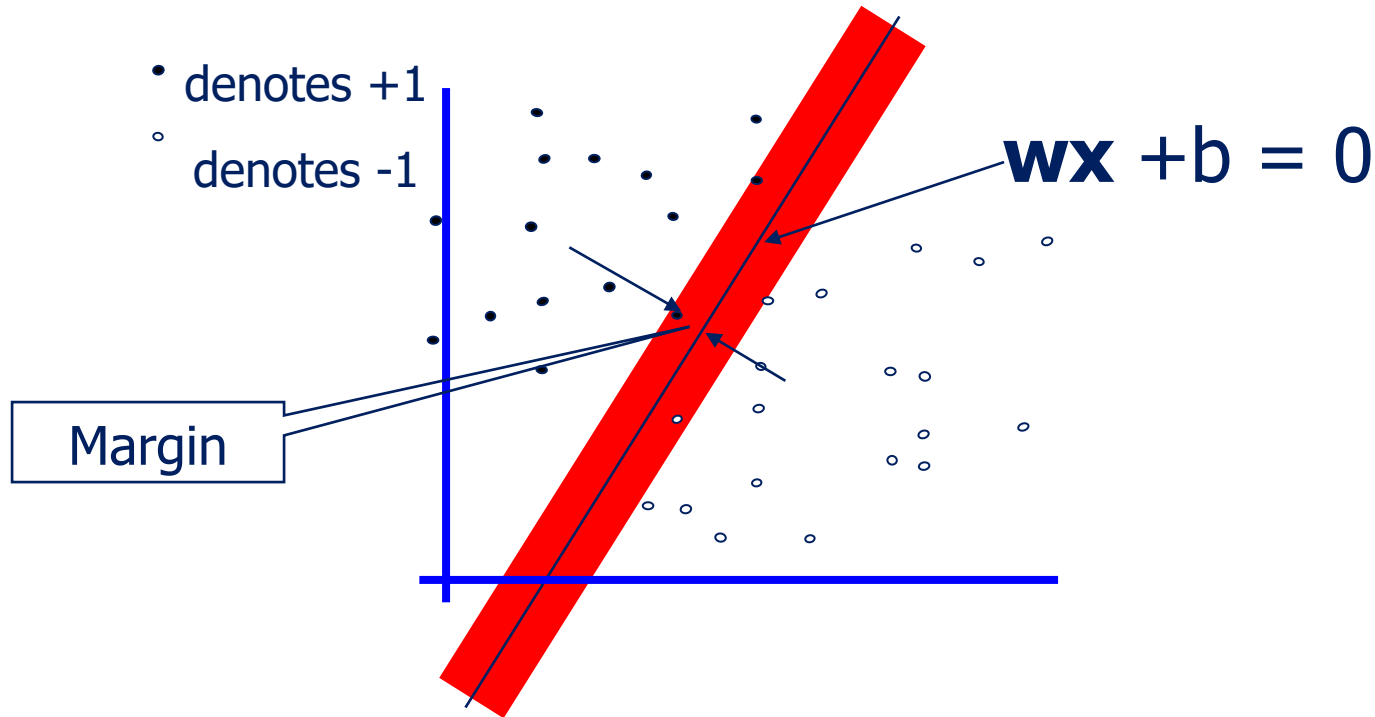
Estimate Margin (Method 1)



- What is the distance expression for a point \mathbf{x} to a line $\mathbf{w}\mathbf{x} + b = 0$?

$$d(\mathbf{x}) = \frac{|\mathbf{x} \cdot \mathbf{w} + b|}{\sqrt{\|\mathbf{w}\|_2^2}} = \frac{|\mathbf{x} \cdot \mathbf{w} + b|}{\sqrt{\sum_{i=1}^d w_i^2}}$$

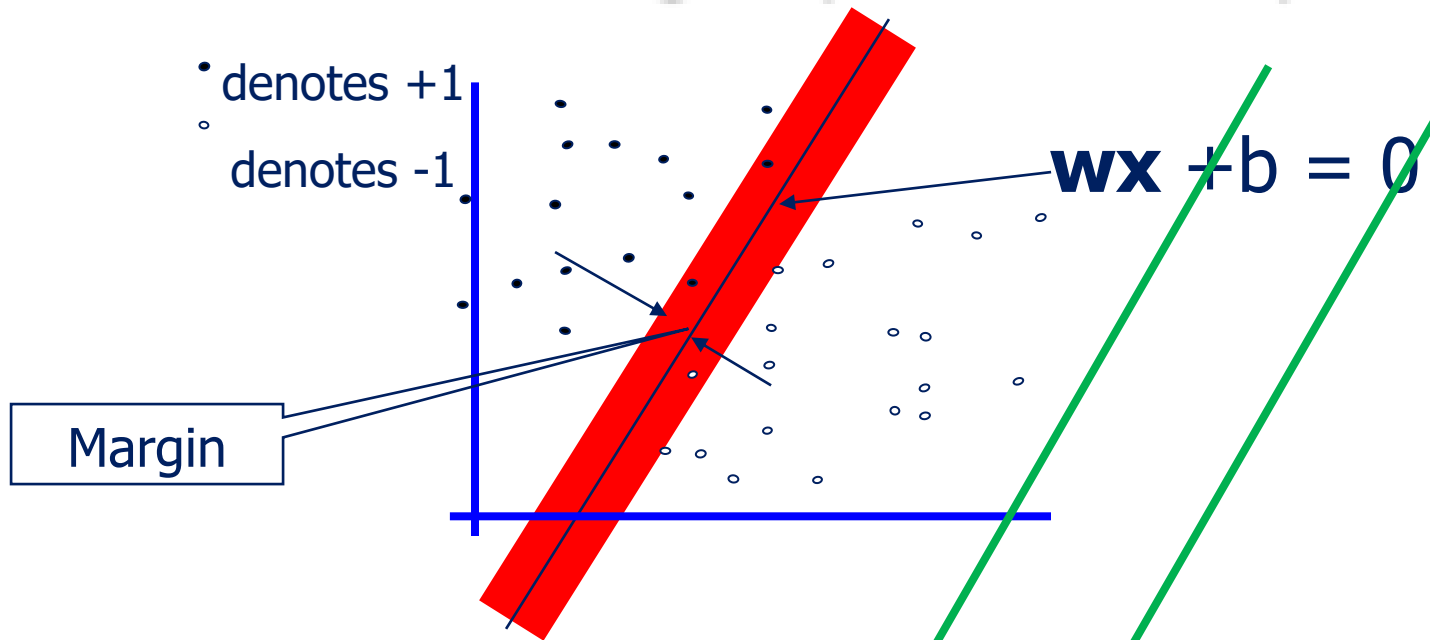
Estimate Margin (Method 1)



- What is the expression for margin?

$$\text{margin} \equiv \arg \min_{\mathbf{x} \in D} d(\mathbf{x}) = \arg \min_{\mathbf{x} \in D} \frac{|\mathbf{x} \cdot \mathbf{w} + b|}{\sqrt{\sum_{i=1}^d w_i^2}}$$

Estimate Margin (Method 1)



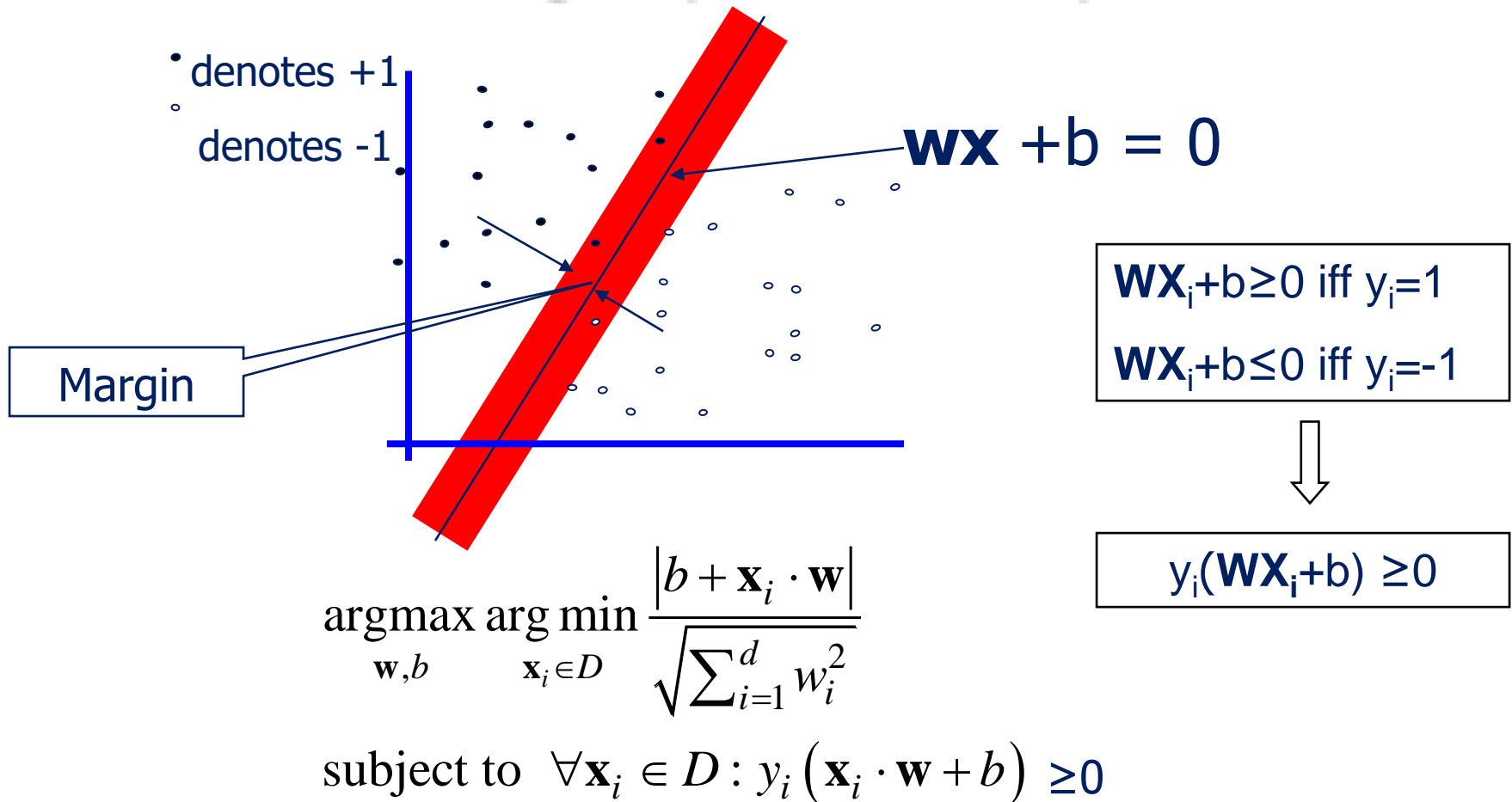
Is this all?

$$\operatorname{argmax}_{\mathbf{w}, b} \operatorname{margin}(\mathbf{w}, b, D)$$

$$= \operatorname{argmax}_{\mathbf{w}, b} \operatorname{argmin}_{\mathbf{x}_i \in D} d(\mathbf{x}_i)$$

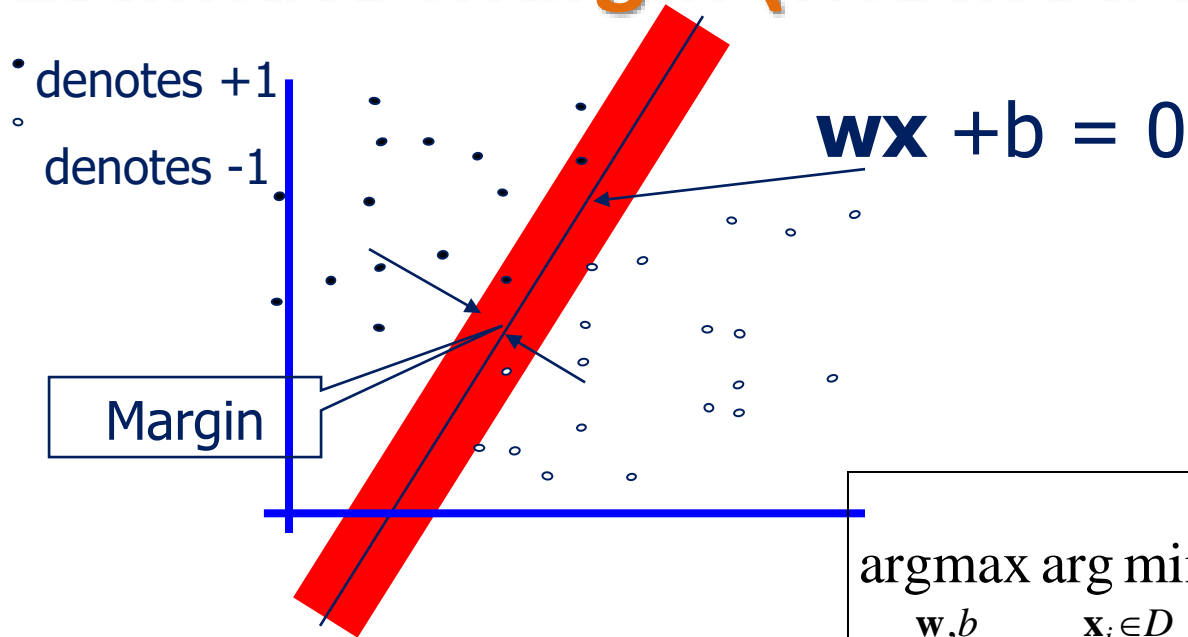
$$= \operatorname{argmax}_{\mathbf{w}, b} \operatorname{argmin}_{\mathbf{x}_i \in D} \frac{|b + \mathbf{x}_i \cdot \mathbf{w}|}{\sqrt{\sum_{i=1}^d w_i^2}}$$

Estimate Margin (Method 1)



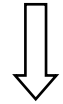
- Min-max problem \rightarrow game problem

Estimate Margin (Method 1)



$$\mathbf{w}\mathbf{x}_i + b \geq 0 \text{ iff } y_i = 1$$

$$\mathbf{w}\mathbf{x}_i + b \leq 0 \text{ iff } y_i = -1$$



$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq 0$$

Strategy:

$$\forall \mathbf{x}_i \in D : |b + \mathbf{x}_i \cdot \mathbf{w}| \geq 1$$

$$\begin{aligned} & \underset{\mathbf{w}, b}{\operatorname{argmax}} \underset{\mathbf{x}_i \in D}{\operatorname{argmin}} \frac{|b + \mathbf{x}_i \cdot \mathbf{w}|}{\sqrt{\sum_{i=1}^d w_i^2}} \\ & \text{subject to } \forall \mathbf{x}_i \in D : y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 0 \end{aligned}$$



$$\begin{aligned} & \underset{\mathbf{w}, b}{\operatorname{argmin}} \sum_{i=1}^d w_i^2 \\ & \text{subject to } \forall \mathbf{x}_i \in D : y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 \end{aligned}$$

Estimate Margin (Method 1)

- How does it come ?

$$\begin{aligned} & \underset{\mathbf{w}, b}{\operatorname{argmax}} \operatorname{argmin}_{\mathbf{x}_i \in D} \frac{|b + \mathbf{x}_i \cdot \mathbf{w}|}{\sqrt{\sum_{i=1}^d w_i^2}} \\ & \text{subject to } \forall \mathbf{x}_i \in D : y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 0 \end{aligned}$$

$$\forall \mathbf{x}_i \in D : |b + \mathbf{x}_i \cdot \mathbf{w}| \geq 1$$

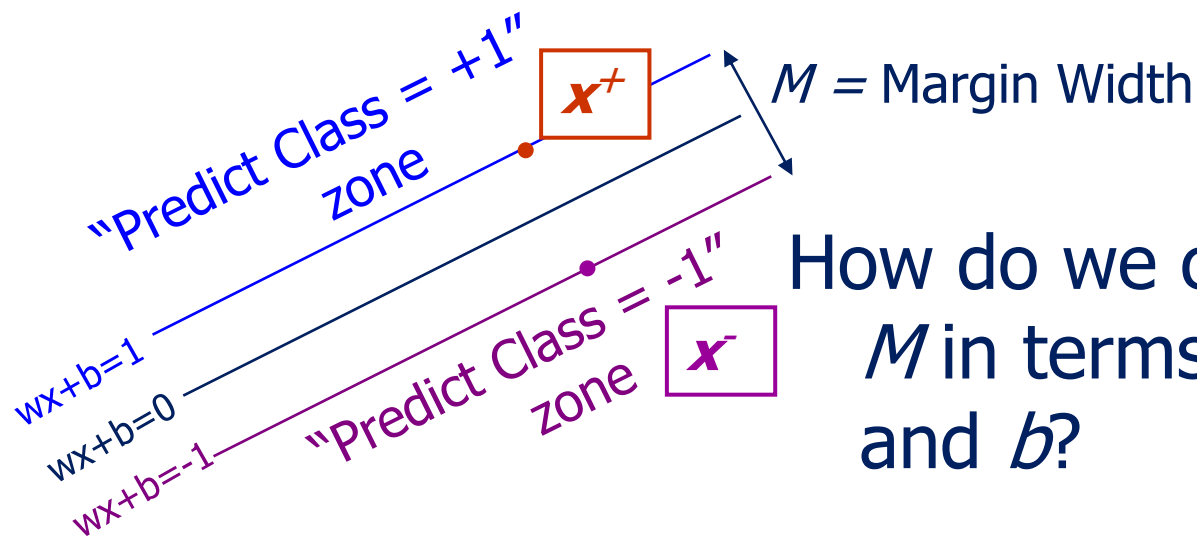


$$\operatorname{argmin}_{\mathbf{x}_i \in D} \frac{|b + \mathbf{x}_i \cdot \mathbf{w}|}{\sqrt{\sum_{i=1}^d w_i^2}} \geq \operatorname{argmin}_{\mathbf{x}_i \in D} \frac{1}{\sqrt{\sum_{i=1}^d w_i^2}} = \frac{1}{\sqrt{\sum_{i=1}^d w_i^2}}$$

$$\begin{aligned} & \underset{\mathbf{w}, b}{\operatorname{argmin}} \sum_{i=1}^d w_i^2 \\ & \text{subject to } \forall \mathbf{x}_i \in D : y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 \end{aligned}$$



Estimate Margin (Method 2)



How do we compute M in terms of \mathbf{w} and b ?

- Plus-plane = $\{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b = +1\}$
- Minus-plane = $\{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b = -1\}$
- What is the distance between these two planes?

Estimate Margin (Method 2)

- Margin can also be defined as distance between two parallel lines.

Given 2 parallel lines with equations

$$ax + by + c_1 = 0$$

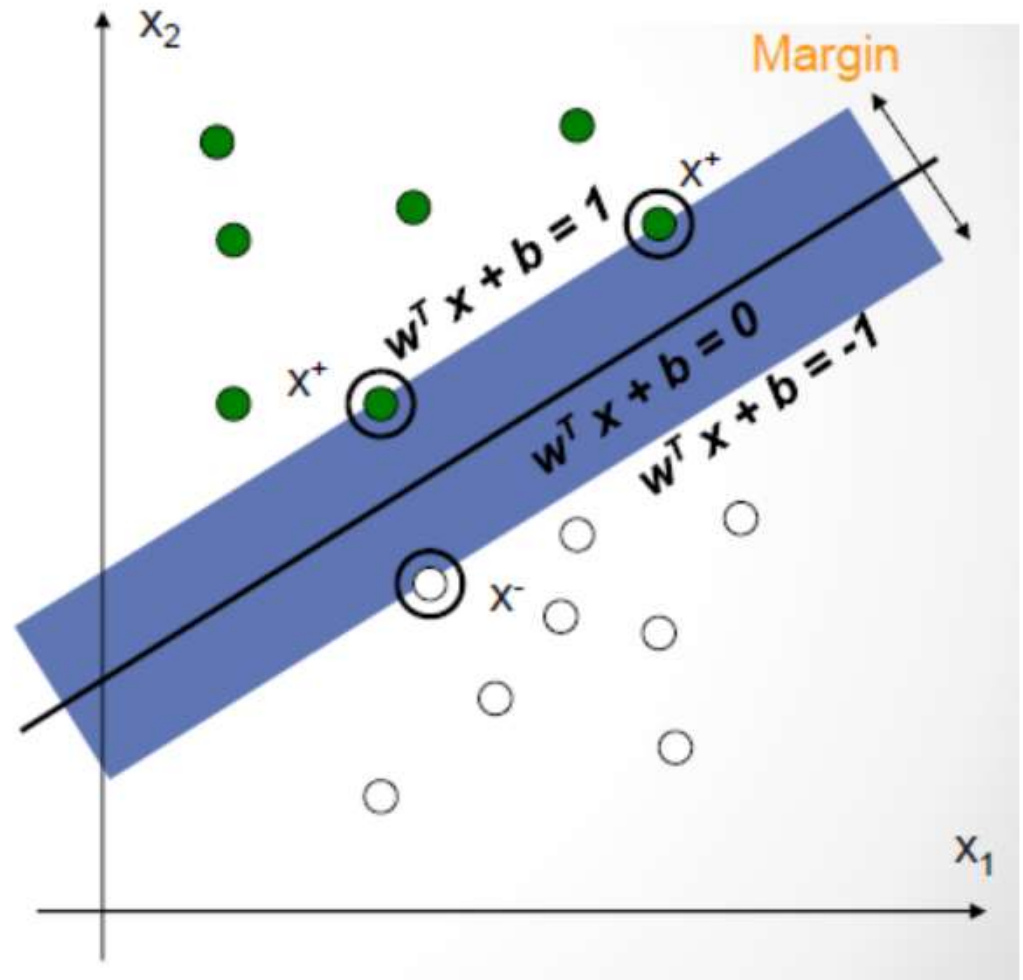
and

$$ax + by + c_2 = 0$$

the distance between them is given by:

$$d = \frac{|c_2 - c_1|}{\sqrt{a^2 + b^2}}$$

$$\frac{2}{\|\mathbf{w}\|}$$



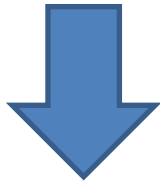
Maximize Margin

$$\text{maximize } \frac{2}{\|\mathbf{w}\|}$$

such that

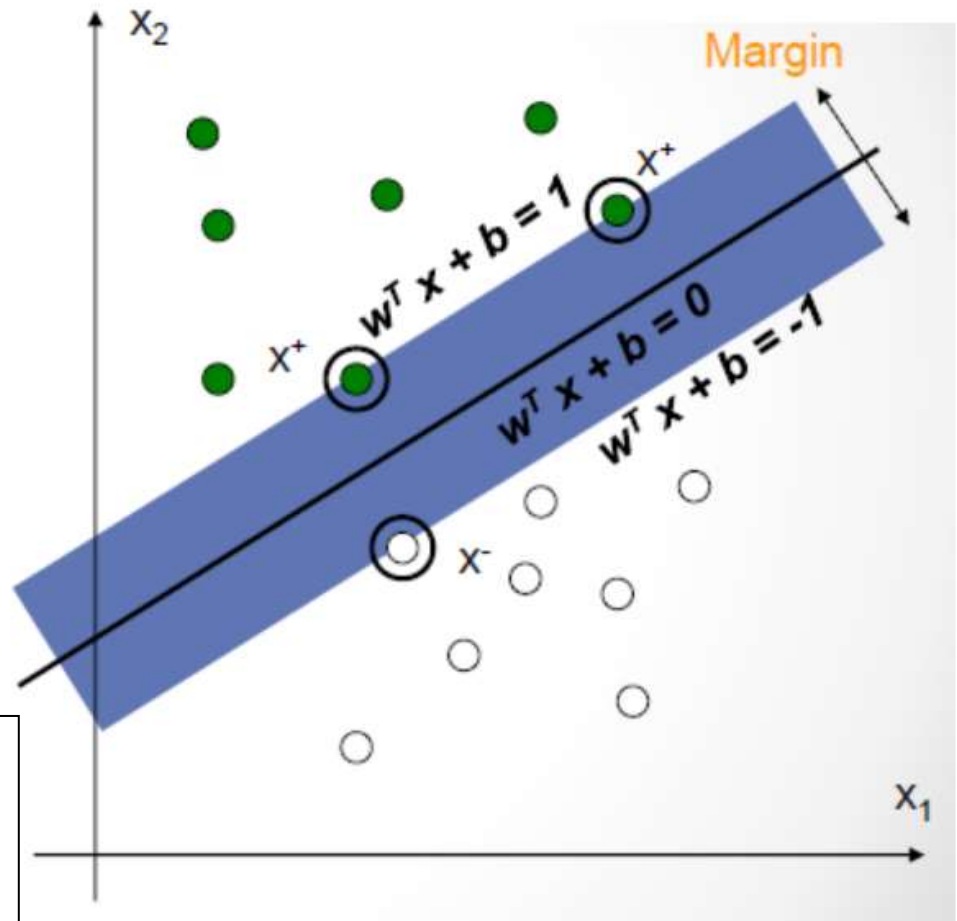
$$\text{For } y_i = +1, \quad \mathbf{w}^T \mathbf{x}_i + b \geq 1$$

$$\text{For } y_i = -1, \quad \mathbf{w}^T \mathbf{x}_i + b \leq -1$$




$$\underset{\mathbf{w}, b}{\operatorname{argmin}} \sum_{i=1}^d w_i^2$$

$$\text{subject to } \forall \mathbf{x}_i \in D : y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1$$




Quadratic Programming

Find $\arg \max_{\mathbf{u}} \quad c + \mathbf{d}^T \mathbf{u} + \frac{\mathbf{u}^T R \mathbf{u}}{2}$  Quadratic criterion


Subject to

$$\begin{aligned} a_{11}u_1 + a_{12}u_2 + \dots + a_{1m}u_m &\leq b_1 \\ a_{21}u_1 + a_{22}u_2 + \dots + a_{2m}u_m &\leq b_2 \\ &\vdots \\ a_{n1}u_1 + a_{n2}u_2 + \dots + a_{nm}u_m &\leq b_n \end{aligned}$$

 n additional linear inequality constraints

And subject to

$$\begin{aligned} a_{(n+1)1}u_1 + a_{(n+1)2}u_2 + \dots + a_{(n+1)m}u_m &= b_{(n+1)} \\ a_{(n+2)1}u_1 + a_{(n+2)2}u_2 + \dots + a_{(n+2)m}u_m &= b_{(n+2)} \\ &\vdots \\ a_{(n+e)1}u_1 + a_{(n+e)2}u_2 + \dots + a_{(n+e)m}u_m &= b_{(n+e)} \end{aligned}$$

 e additional linear equality constraints

Quadratic Programming for Linear SVM

$$\{\vec{w}^*, b^*\} = \min_{\vec{w}, b} \sum_i w_i^2$$

subject to $y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1$ for all training data (\vec{x}_i, y_i)



$$\{\vec{w}^*, b^*\} = \operatorname{argmax}_{\vec{w}, b} \left\{ 0 + \vec{0} \cdot \vec{w} - \vec{w}^T \mathbf{I}_n \vec{w} \right\}$$

$$\left. \begin{array}{l} y_1 (\vec{w} \cdot \vec{x}_1 + b) \geq 1 \\ y_2 (\vec{w} \cdot \vec{x}_2 + b) \geq 1 \\ \dots \\ y_N (\vec{w} \cdot \vec{x}_N + b) \geq 1 \end{array} \right\} \text{inequality constraints}$$

Soft-Margin SVM

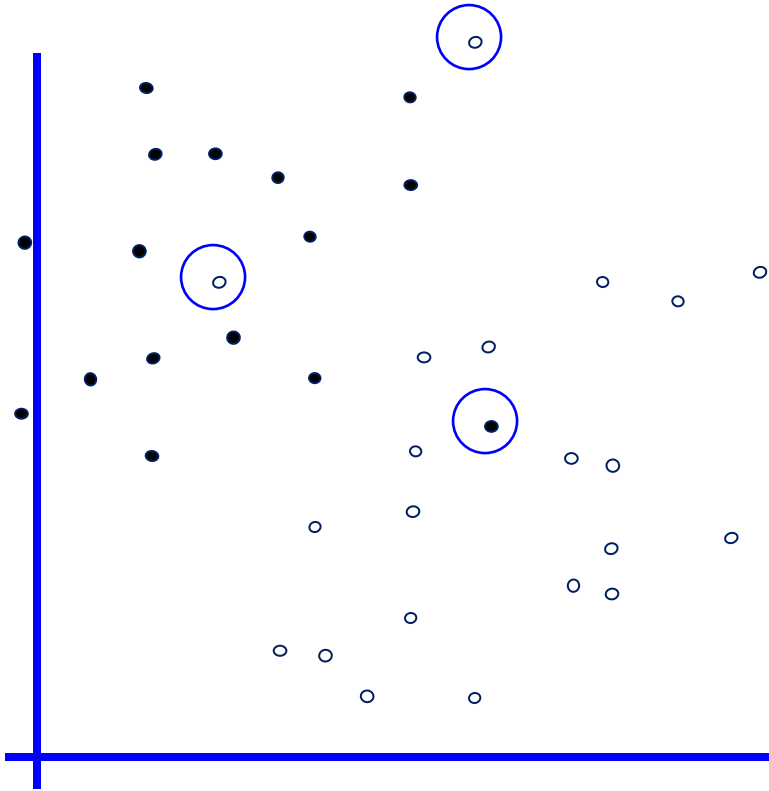
Noisy Data

Not linear separable

This is going to be a problem!

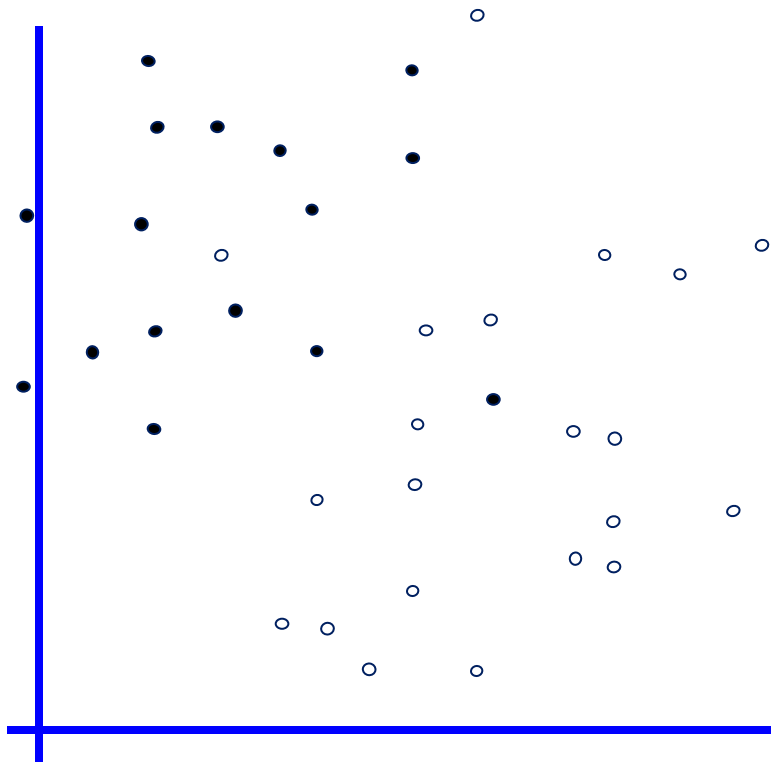
What should we do?

- denotes +1
- denotes -1



Noisy Data

- denotes +1
- denotes -1



This is going to be a problem!

What should we do?

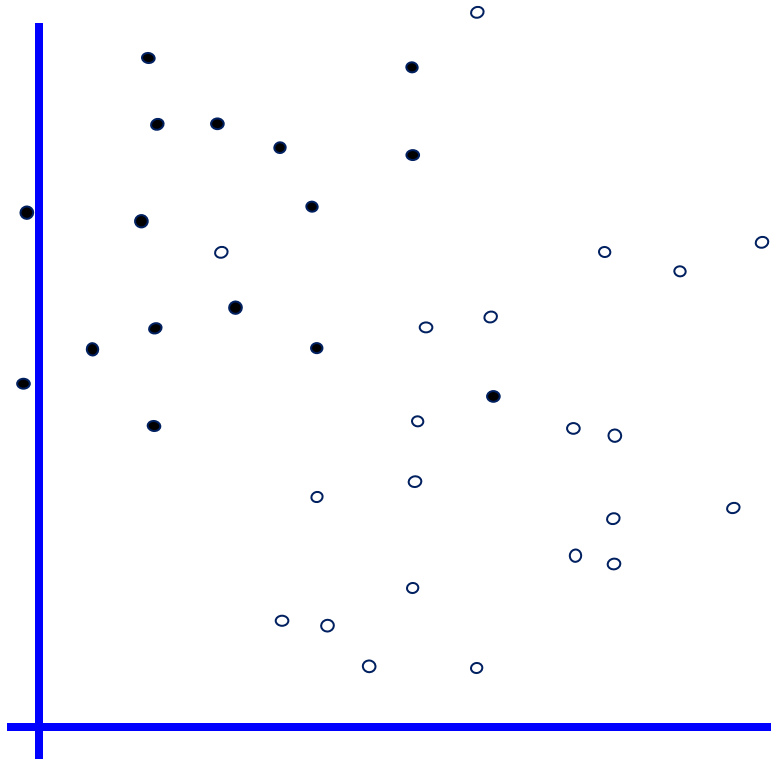
Idea 1:

Find minimum \mathbf{w}, \mathbf{w} , while minimizing number of training set errors.

Problem: Two things to minimize makes for an ill-defined optimization

Noisy Data

- denotes +1
- denotes -1



This is going to be a problem!

What should we do?

Idea 1.1:

Minimize

$$\mathbf{w} \cdot \mathbf{w} + C (\#train\ errors)$$

Tradeoff parameter

There's a serious practical problem that's about to make us reject this approach. Can you guess what it is?

Noisy Data

This is going to be a problem!

What should we do?

Idea 1.1:

Minimize

$$\mathbf{w} \cdot \mathbf{w} + C (\# \text{train errors})$$

Tradeoff parameter

Can't be expressed as a Quadratic Programming problem.

Solving it may be too slow.

(Also, doesn't distinguish between disastrous errors and near misses)

So... any other ideas?

you guess what?

Noisy Data

This is going to be a problem!

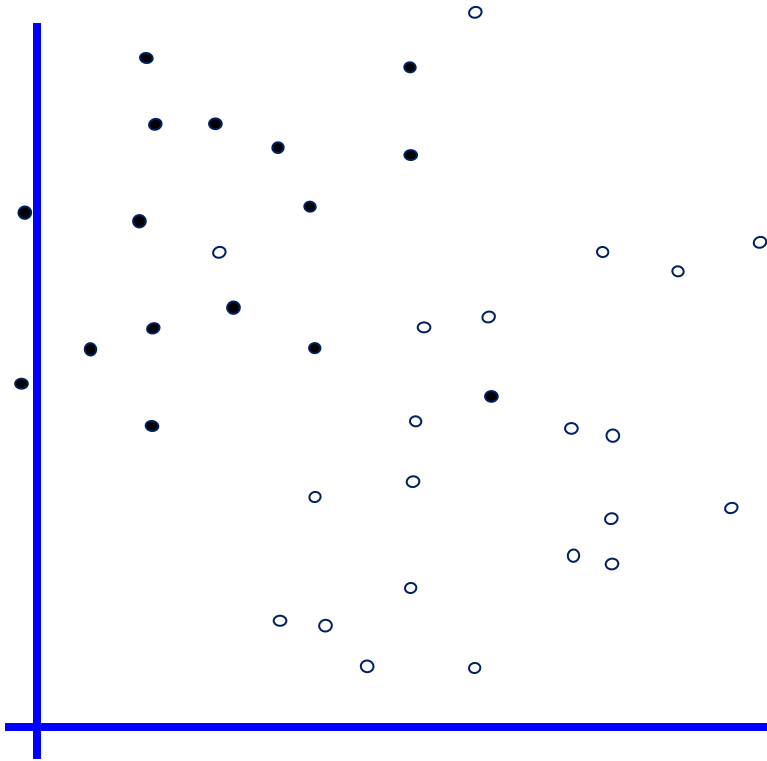
What should we do?

Idea 2.0:

Minimize

$\mathbf{W} \cdot \mathbf{W} + C$ (*distance of error points to their correct place*)

- denotes +1
- denotes -1



SVM for Noisy Data

$$\{\vec{w}^*, b^*\} = \min_{\vec{w}, b} \sum_{i=1}^d w_i^2 + c \sum_{j=1}^N \varepsilon_j$$

$$y_1 (\vec{w} \cdot \vec{x}_1 + b) \geq 1 - \varepsilon_1$$

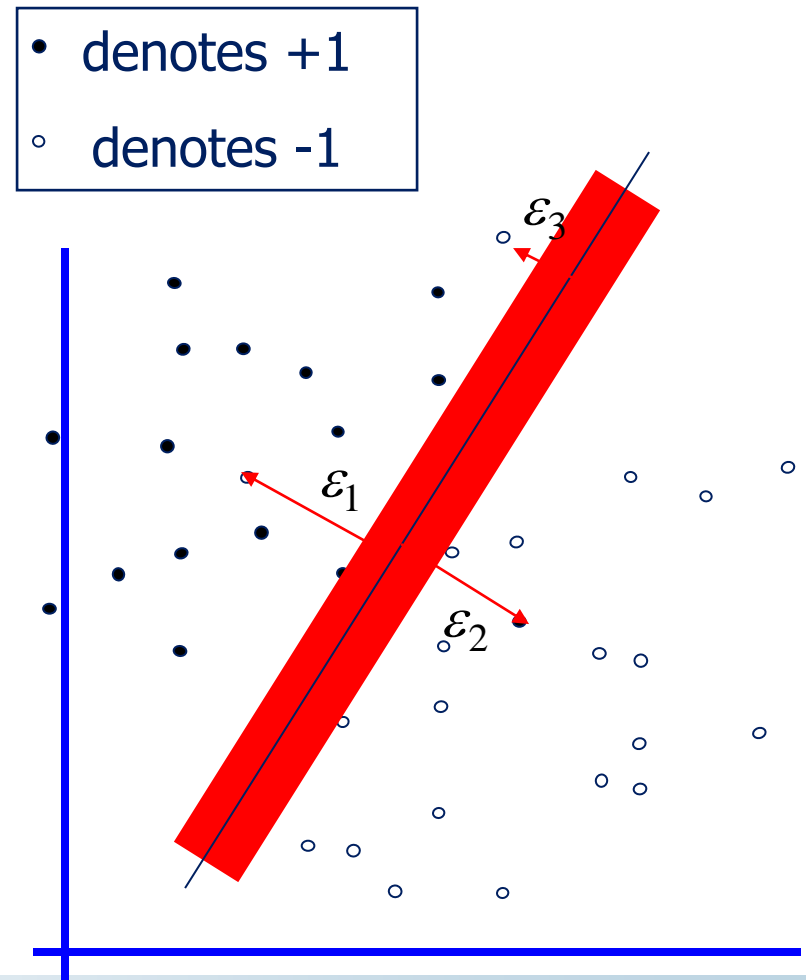
$$y_2 (\vec{w} \cdot \vec{x}_2 + b) \geq 1 - \varepsilon_2$$

...

$$y_N (\vec{w} \cdot \vec{x}_N + b) \geq 1 - \varepsilon_N$$

- Any problem with the above formulism?

What happens when $\varepsilon_i < 0$?



SVM for Noisy Data

$$\{\vec{w}^*, b^*\} = \min_{\vec{w}, b} \sum_{i=1}^d w_i^2 + c \sum_{j=1}^N \varepsilon_j$$

$$y_1 (\vec{w} \cdot \vec{x}_1 + b) \geq 1 - \varepsilon_1, \varepsilon_1 \geq 0$$

$$y_2 (\vec{w} \cdot \vec{x}_2 + b) \geq 1 - \varepsilon_2, \varepsilon_2 \geq 0$$

...

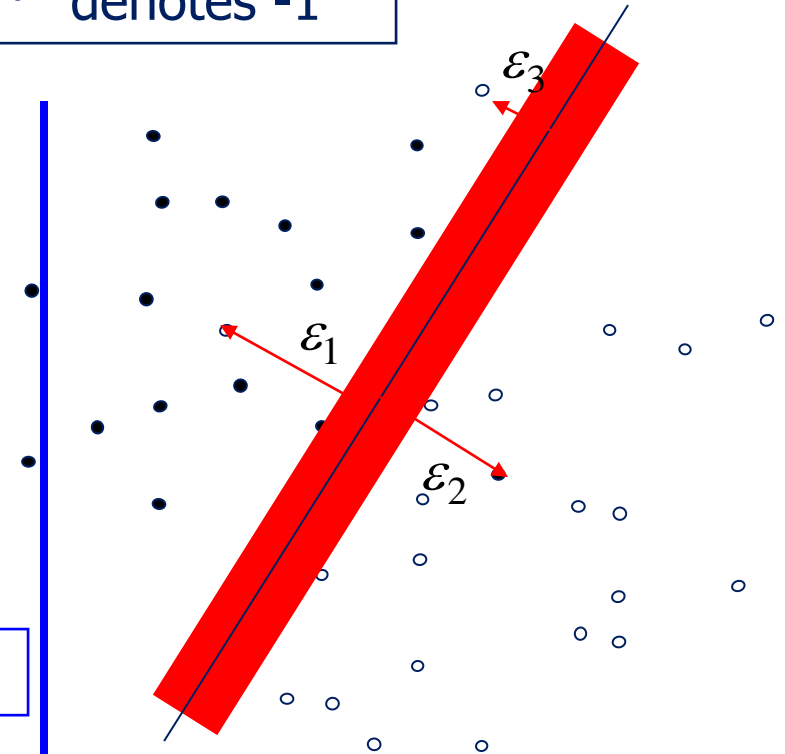
$$y_N (\vec{w} \cdot \vec{x}_N + b) \geq 1 - \varepsilon_N, \varepsilon_N \geq 0$$

- Balance the trade off between margin and classification errors

Describe the Theory

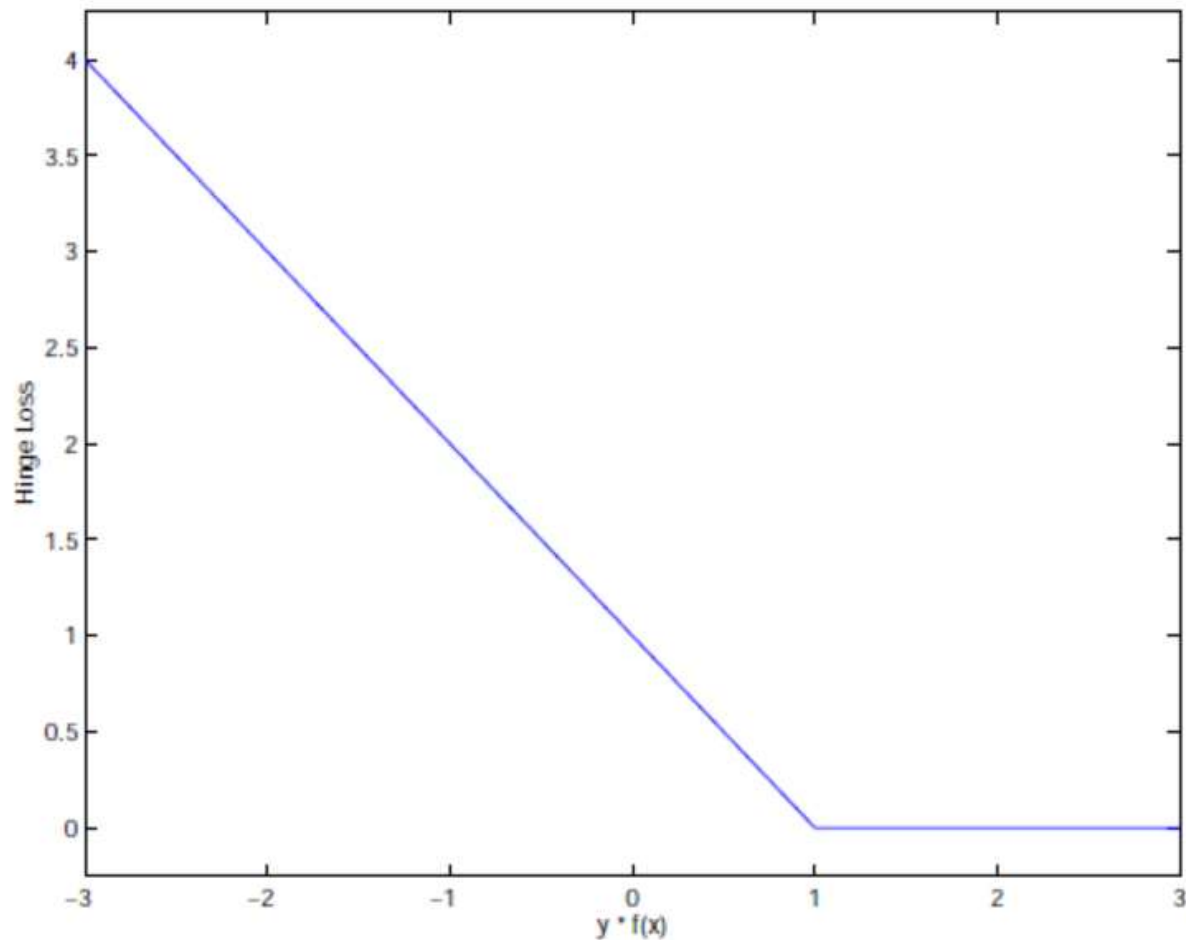
Describe the Mistake

- denotes +1
- denotes -1

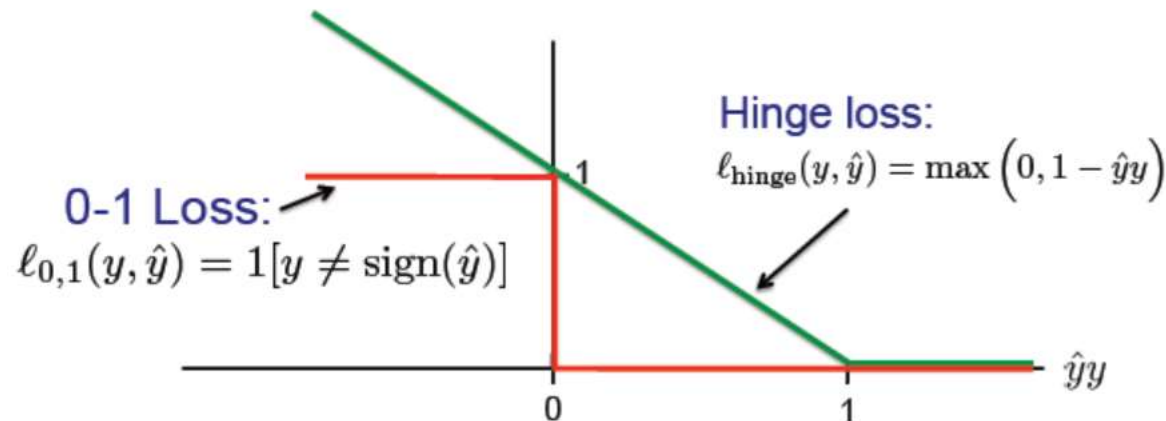


Hinge Loss

$$\text{hinge}(x) = \max(1 - x, 0)$$



Hinge Loss



Hinge loss upper bounds 0/1 loss!

➡ It is the tightest *convex* upper bound on the 0/1 loss

- The SVM is a Tikhonov regularization problem, using the hinge loss:

$$\operatorname{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (1 - y_i f(x_i))_+ + \lambda \|f\|_{\mathcal{H}}^2.$$

Conclusion so far

- Which linear classifier \rightarrow large margin
- How to calculate margin $\rightarrow \frac{2}{\|\mathbf{w}\|}$
- Hard margin SVM \rightarrow QP
- Not linear separable \rightarrow soft-margin SVM
- Still QP
- L2 norm regularization + hinge loss

Dual Problem

Lagrange Multipliers

Minimize $f(x)$

$$\text{subject to } \begin{cases} a(x) \geq 0 \\ b(x) \leq 0 \\ c(x) = 0 \end{cases} \quad \begin{cases} \alpha_1 \geq 0 \\ \alpha_2 \leq 0 \\ \alpha_3 \text{ is unconstrained} \end{cases}$$

We can recover the primal problem by maximizing the Lagrangian with respect to the Lagrange multipliers

$$\max_{\alpha} L(x, \alpha) = \begin{cases} f(x), & \text{if } \begin{cases} a(x) \geq 0 \\ b(x) \leq 0 \\ c(x) = 0 \end{cases} \\ +\infty, & \text{otherwise} \end{cases}$$

So, the Primal problem can be changed into Dual problem

$$\min_x \max_{\alpha} L(x, \alpha) = \max_{\alpha} \min_x L(x, \alpha)$$

$\boxed{\text{Primal}(x)}$ \quad $\boxed{\text{Dual}(\alpha)}$

Karush-Kuhn-Tucker conditions

- For a local minimum

$$\left\{ \begin{array}{l} \text{Stationarity} \quad \nabla f(x^*) - \alpha_1 \nabla a(x^*) - \alpha_2 \nabla b(x^*) - \alpha_3 \nabla c(x^*) = 0 \\ \text{Primal feasibility} \quad \begin{cases} a(x^*) \geq 0 \\ b(x^*) \leq 0 \\ c(x^*) = 0 \end{cases} \\ \text{Dual feasibility} \quad \begin{cases} \alpha_1 \geq 0 \\ \alpha_2 \leq 0 \\ \alpha_3 \text{ is unconstrained} \end{cases} \\ \text{Complementary slackness} \quad \begin{cases} \alpha_1 a(x^*) = 0 \\ \alpha_2 b(x^*) = 0 \\ \alpha_3 c(x^*) = 0 \end{cases} \end{array} \right.$$

Lagrange Transformation

$$\begin{aligned} & \text{Minimize } \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to } 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b) \leq 0 \quad \text{for } i = 1, \dots, n \end{aligned}$$

- The Lagrangian is

$$\mathcal{L} = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{i=1}^n \alpha_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))$$

Lagrangian multipliers

- Setting the gradient of \mathcal{L} w.r.t. \mathbf{w} and b to zero, we have

$$\mathbf{w} + \sum_{i=1}^n \alpha_i (-y_i) \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

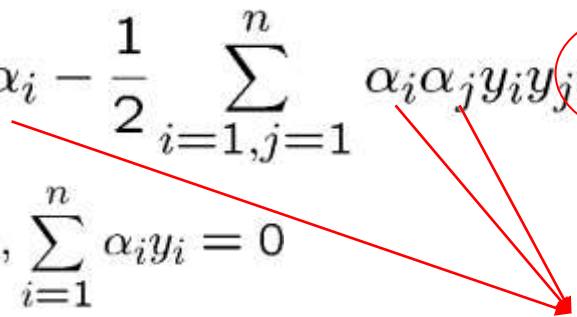
$$\sum_{i=1}^n \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

Dual Problem

- We can transform the problem to its dual

Dot product of \mathbf{X}

$$\begin{aligned} \max. \quad W(\boldsymbol{\alpha}) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to } \alpha_i &\geq 0, \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$


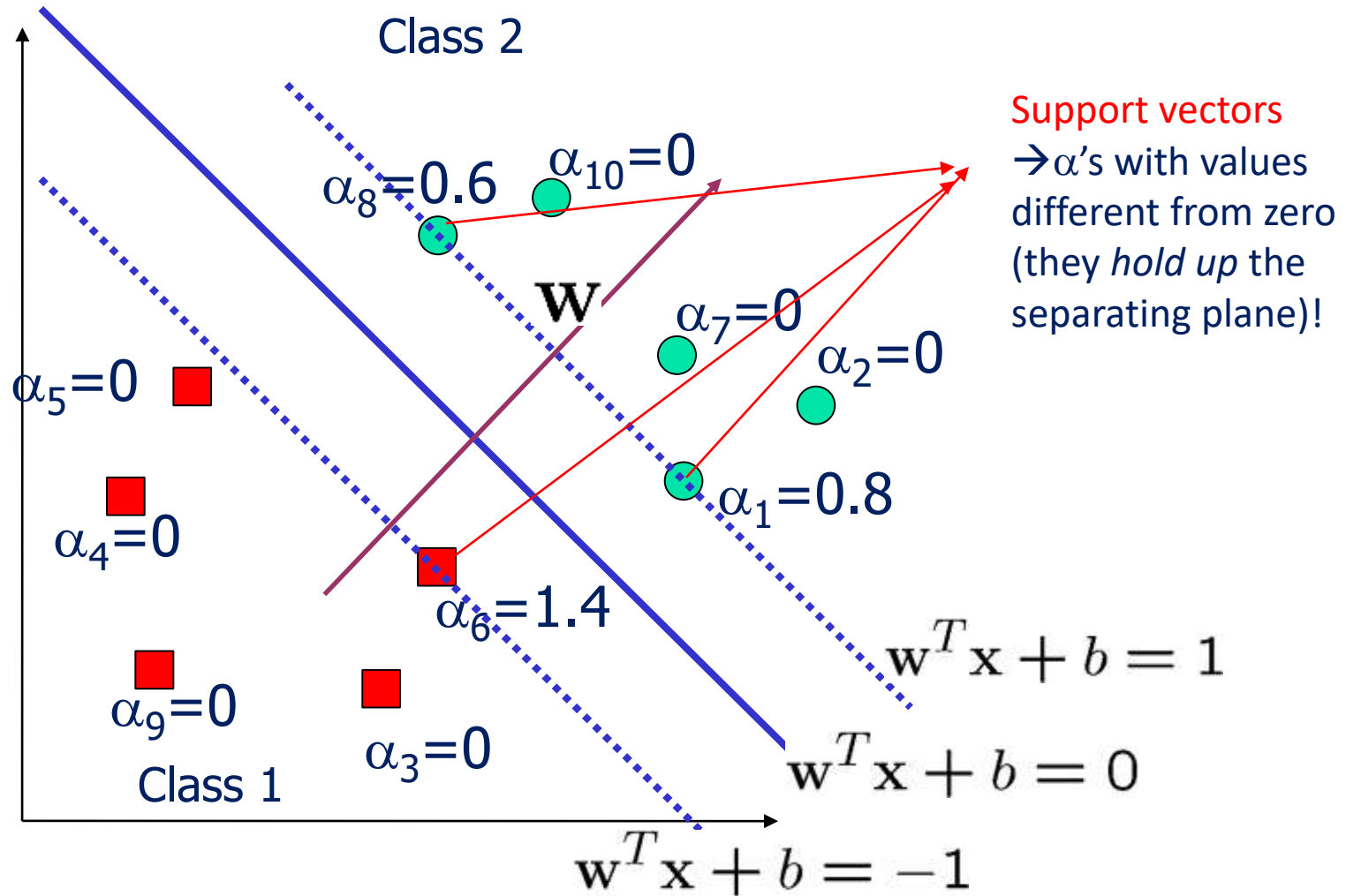
α 's \rightarrow New variables
(Lagrangian multipliers)

- This is a convex quadratic programming (QP) problem
 - Global maximum of α_i can always be found
 - \rightarrow well established tools for solving this optimization problem

- **Note:**

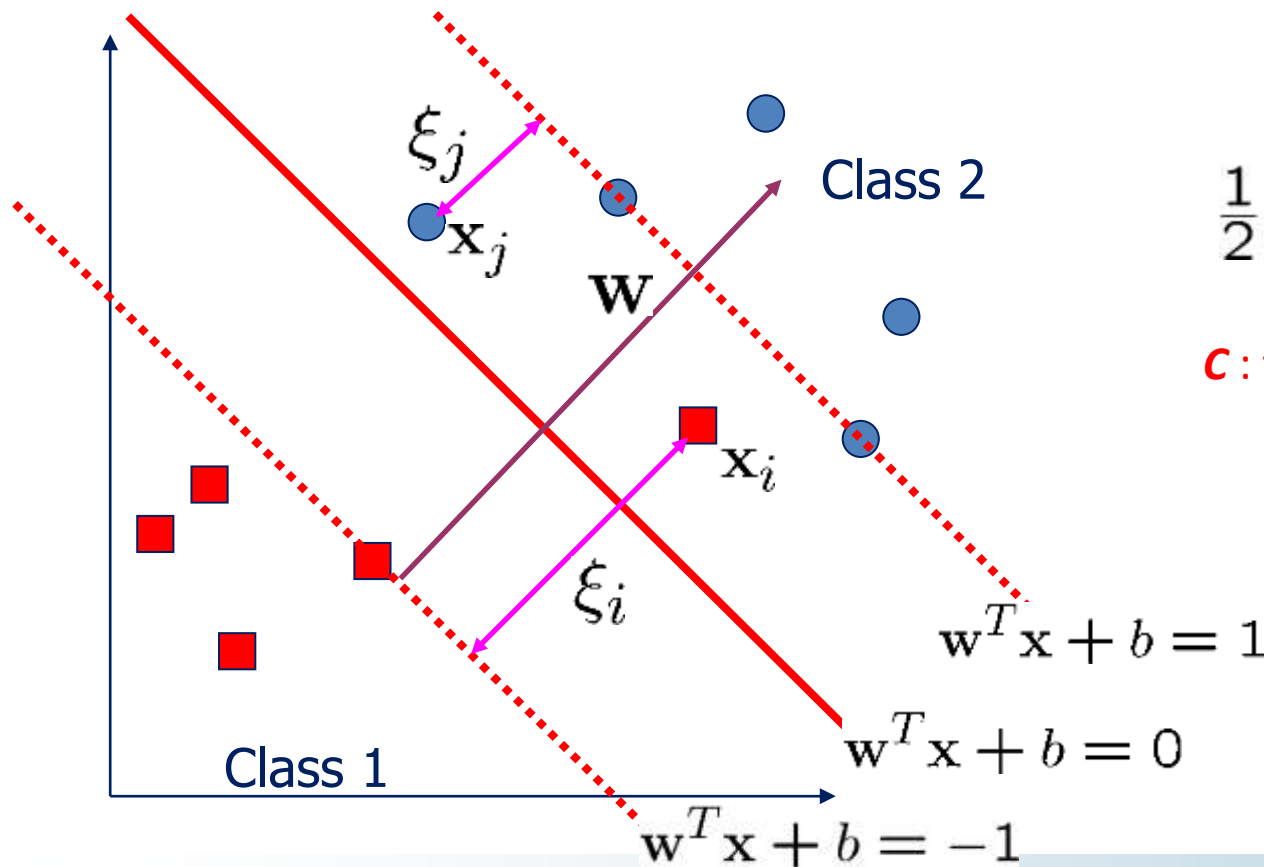
$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

Support Vectors



Soft-Margin Case

- We allow “error” ξ_i in classification; it is based on the output of the discriminant function $\mathbf{w}^T \mathbf{x} + b$
- ξ_i approximates the number of misclassified samples



New objective function:

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

C : tradeoff parameter between error and margin; chosen by the user; large C means a higher penalty to errors

Soft-Margin Case

- Lagrangian Problem

$$L = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^n \gamma_i \xi_i$$

$$\alpha_i \geq 0 \quad \gamma_i \geq 0$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \alpha_i + \gamma_i = C \Rightarrow 0 \leq \alpha_i \leq C$$

Dual Problem for Soft-Margin SVM

- The dual of the problem is

$$\begin{aligned} \max. \quad W(\boldsymbol{\alpha}) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to } & C \geq \alpha_i \geq 0, \quad \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$

- \mathbf{w} is also recovered as $\mathbf{w} = \sum_{j=1}^s \alpha_{t_j} y_{t_j} \mathbf{x}_{t_j}$
- The only difference with the linear separable case is that there is an upper bound C on α_i
- Once again, a QP solver can be used to find α_i efficiently!!!

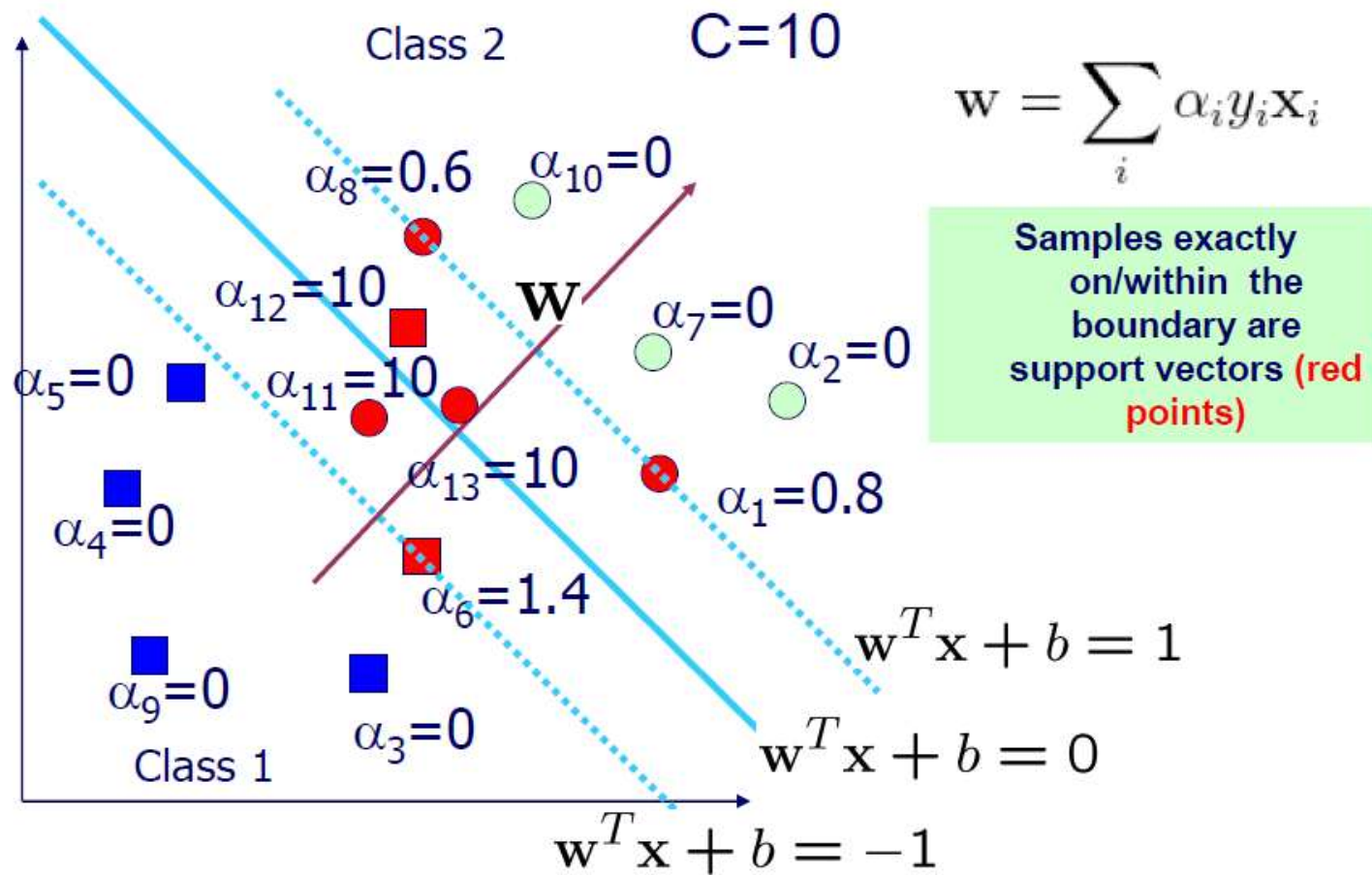
KKT conditions

- When optimum is achieved, KKT conditions are satisfied

$$\begin{cases} \alpha_i(y_i f(x_i) - 1 + \xi_i) = 0 \\ \gamma_i \xi_i = 0 \\ \alpha_i + \gamma_i = C \Rightarrow 0 \leq \alpha_i \leq C \end{cases}$$

$$\begin{cases} \alpha_i = 0 & \Rightarrow y_i f(x_i) \geq 1 & \Rightarrow \text{Samples outside the boundary} \\ 0 < \alpha_i < C & \Rightarrow y_i f(x_i) = 1 & \Rightarrow \text{Samples on the boundary} \\ \alpha_i = C & \Rightarrow y_i f(x_i) \leq 1 & \Rightarrow \text{Samples within the boundary} \end{cases}$$

Support Vectors

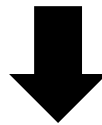


Finding the bias b

- Find the bias b based on

$$0 < \alpha_i < C \Rightarrow y_i f(\mathbf{x}_i) = 1$$

$$f(\mathbf{z}) = \sum_{j=1}^s \alpha_j y_j \mathbf{x}_j^T \mathbf{z} + b$$

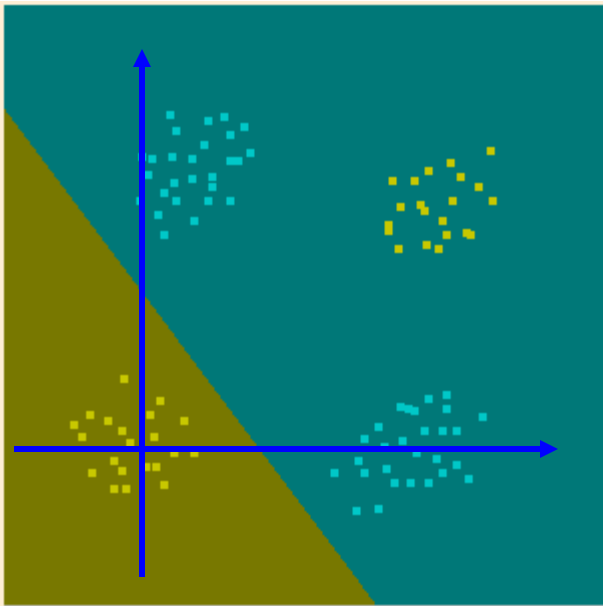


$$b = y_i - \sum_{j=1}^s \alpha_j y_j \mathbf{x}_j^T \mathbf{x}_i \quad \forall 0 < \alpha_i < C$$

Kernel Methods

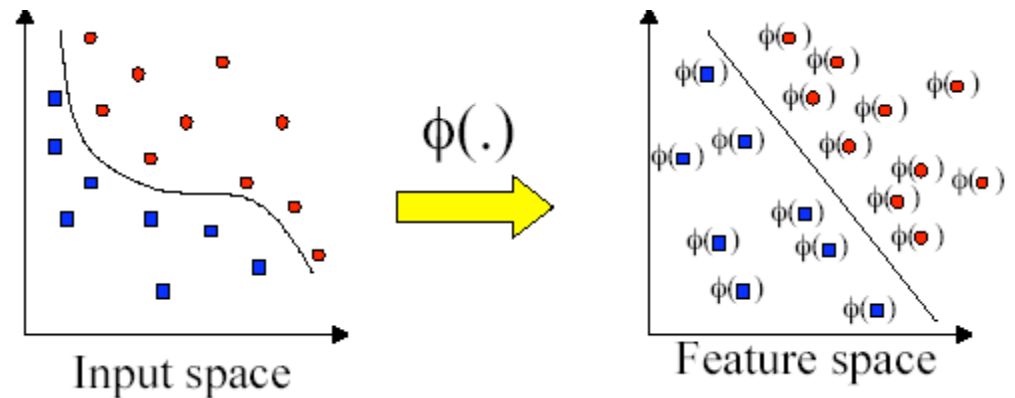
Feature Transformation ?

- The problem is non-linear
- Find some trick to transform the input
- Linear separable after Feature Transformation
- What Features should we use ?



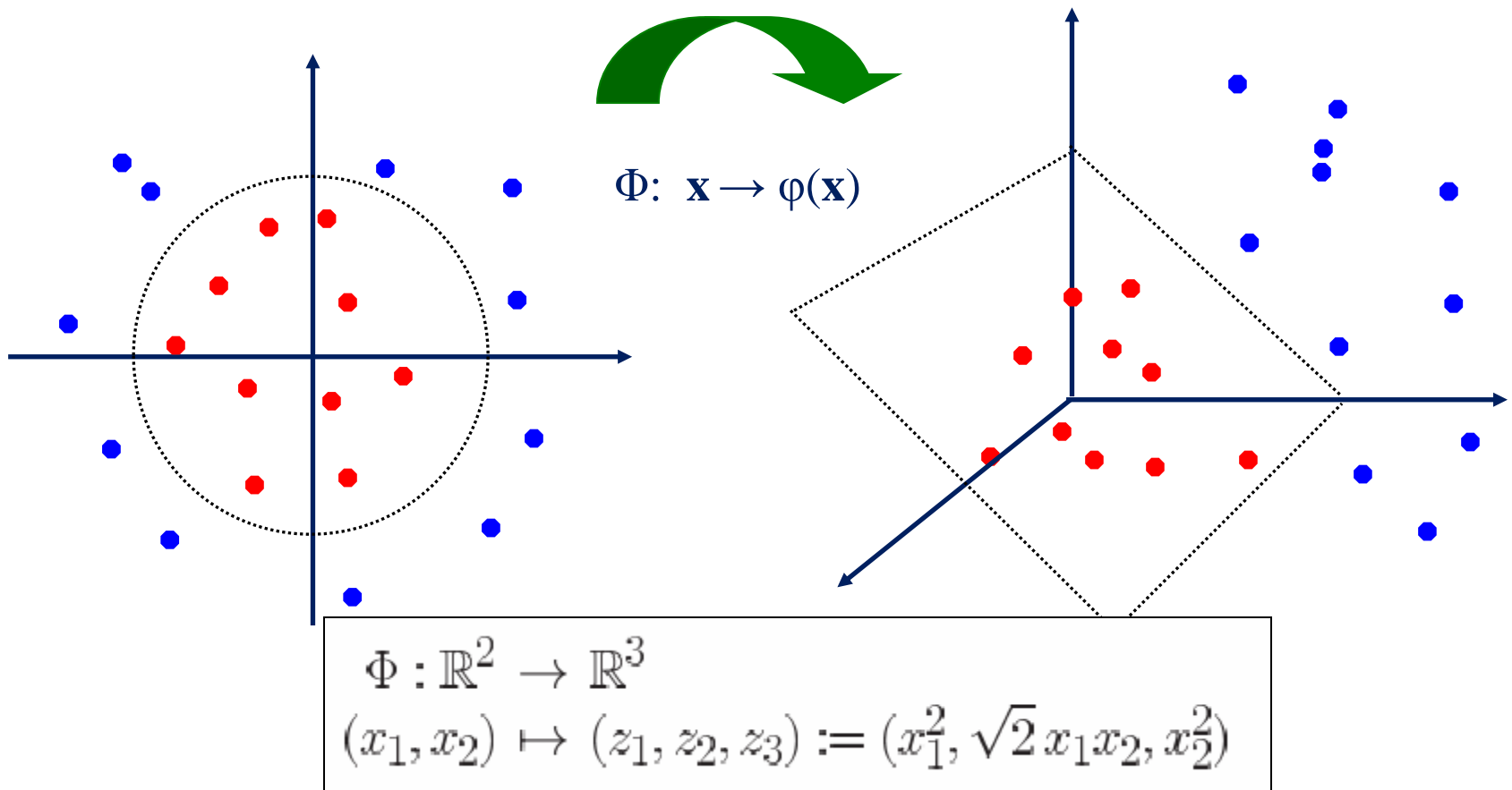
XOR Problem

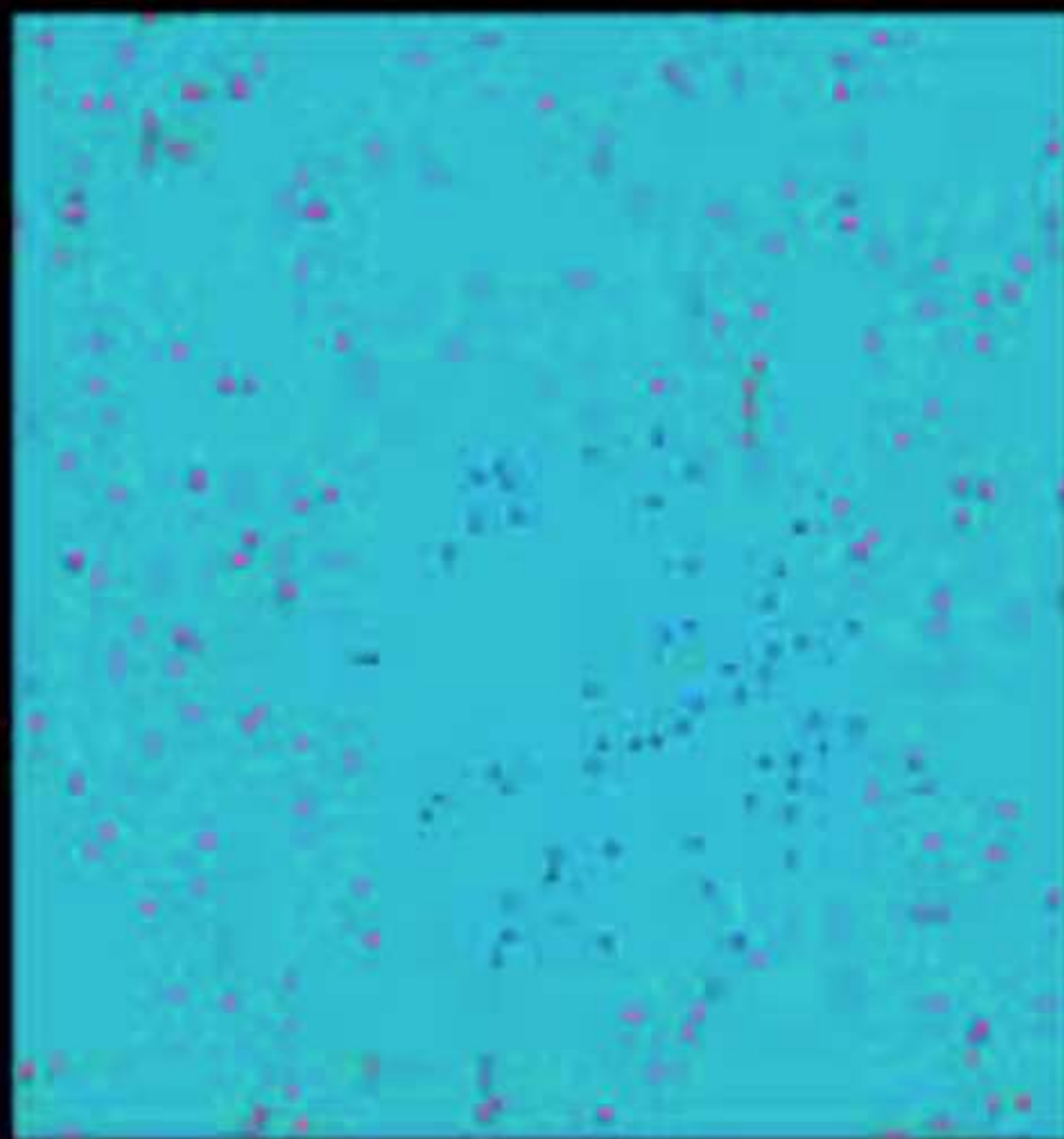
Basic Idea :



Non-linear SVMs: Feature spaces

- General idea: the original input space can always be mapped to some higher-dimensional feature space where the training set is separable:





Kernel Trick

Recall:

$$\text{maximize} \quad \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j x_i x_j$$

$$\text{subject to} \quad C \geq \alpha_i \geq 0, \sum_{i=1}^N \alpha_i y_i = 0$$

Note that data only appears as dot products

Since data is only represented as dot products, we need not do the mapping explicitly.

Introduce a Kernel Function K such that:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

Kernel function – a function that can be applied to pairs of input data to evaluate dot products in some corresponding feature space

Example Transform

- Consider the following transformation

$$\phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$\phi\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) = (1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1y_2)$$

- Define the kernel function $K(\mathbf{x}, \mathbf{y})$ as

$$\begin{aligned} \langle \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right), \phi\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) \rangle &= (1 + x_1y_1 + x_2y_2)^2 \\ &= K(\mathbf{x}, \mathbf{y}) \end{aligned}$$

$$K(\mathbf{x}, \mathbf{y}) = (1 + x_1y_1 + x_2y_2)^2$$

- The inner product $\phi(.)\phi(.)$ can be computed by K without going through the map $\phi(.)$ explicitly!!!

Examples of Kernel Function

- Polynomial kernel with degree d

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^d$$

- Radial basis function kernel with width σ

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / (2\sigma^2))$$

– Closely related to radial basis function neural networks

- Sigmoid with parameter κ and θ

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\kappa \mathbf{x}^T \mathbf{y} + \theta)$$

– It does not satisfy the Mercer condition on all κ and θ

- Research on different kernel functions in different applications is very active

Modification Due to Kernel Function

- Change **all inner products to kernel functions**
- For training,

Original

$$\max. W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^n \alpha_i y_i = 0$$

With kernel
function

$$\max. W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^n \alpha_i y_i = 0$$

Modification Due to Kernel Function

- For testing, the new data \mathbf{z} is classified as *class 1* if $f \geq 0$ and as *class 2* if $f < 0$

Original

$$\mathbf{w} = \sum_{j=1}^s \alpha_{t_j} y_{t_j} \mathbf{x}_{t_j}$$

$$f = \mathbf{w}^T \mathbf{z} + b = \sum_{j=1}^s \alpha_{t_j} y_{t_j} \mathbf{x}_{t_j}^T \mathbf{z} + b$$

With kernel
function

$$\mathbf{w} = \sum_{j=1}^s \alpha_{t_j} y_{t_j} \phi(\mathbf{x}_{t_j})$$

$$f = \langle \mathbf{w}, \phi(\mathbf{z}) \rangle + b = \sum_{j=1}^s \alpha_{t_j} y_{t_j} K(\mathbf{x}_{t_j}, \mathbf{z}) + b$$

Modification Due to Kernel Function

- Find the bias b

Original

$$b = y_i - \sum_{j=1}^s \alpha_j y_j x_j^T x_i \quad \forall 0 < \alpha_i < C$$

With kernel function

$$b = y_i - \sum_{j=1}^s \alpha_j y_j k(x_j, x_i) \quad \forall 0 < \alpha_i < C$$

Example

- Suppose we have 5 1D data points
– $x_1=1, x_2=2, x_3=4, x_4=5, x_5=6$, with 1, 2, 6 as class 1 and 4, 5 as class 2 $\Rightarrow y_1=1, y_2=1, y_3=-1, y_4=-1, y_5=1$
- We use the polynomial kernel of degree 2
– $K(x,y) = (xy+1)^2$
– C is set to 100

- We first find α_i ($i=1, \dots, 5$) by

$$\max. \quad \sum_{i=1}^5 \alpha_i - \frac{1}{2} \sum_{i=1}^5 \sum_{j=1}^5 \alpha_i \alpha_j y_i y_j (x_i x_j + 1)^2$$

$$\text{subject to } 100 \geq \alpha_i \geq 0, \quad \sum_{i=1}^5 \alpha_i y_i = 0$$

Example

- By using a QP solver, we get

$$\alpha_1=0, \alpha_2=2.5, \alpha_3=0, \alpha_4=7.333, \alpha_5=4.833$$

–Verify that the constraints are indeed satisfied

–The support vectors are $\{x_2=2, x_4=5, x_5=6\}$

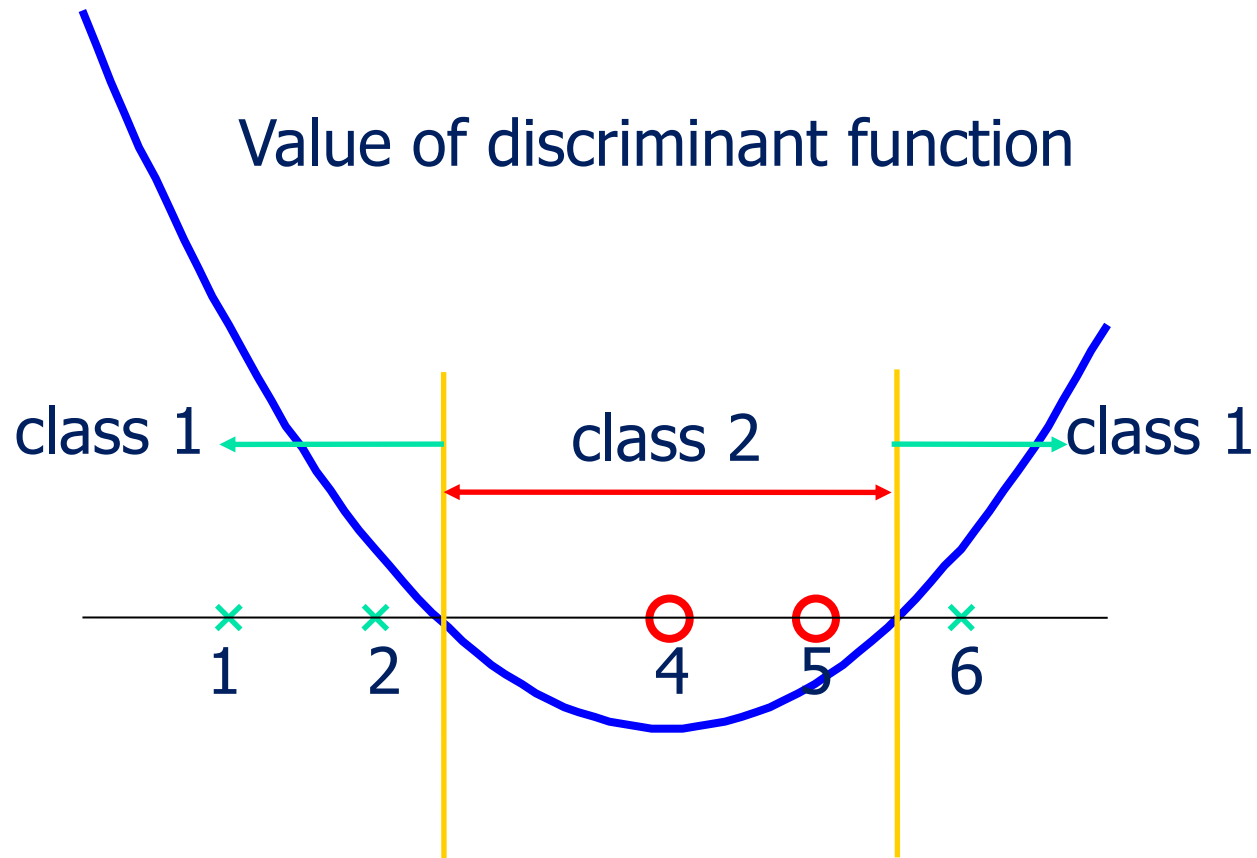
- The discriminant function is

$$\begin{aligned} f(y) &= 2.5(1)(2y+1)^2 + 7.333(-1)(5y+1)^2 + 4.833(1)(6y+1)^2 + b \\ &= 0.6667x^2 - 5.333x + b \end{aligned}$$

- b is recovered by solving $f(2)=1$ or by $f(5)=-1$ or by $f(6)=1$,
as x_2, x_4, x_5 lie on $y_i(\mathbf{w}^T \phi(z) + b) = 1$ and all give $b=9$

→ $f(y) = 0.6667x^2 - 5.333x + 9$

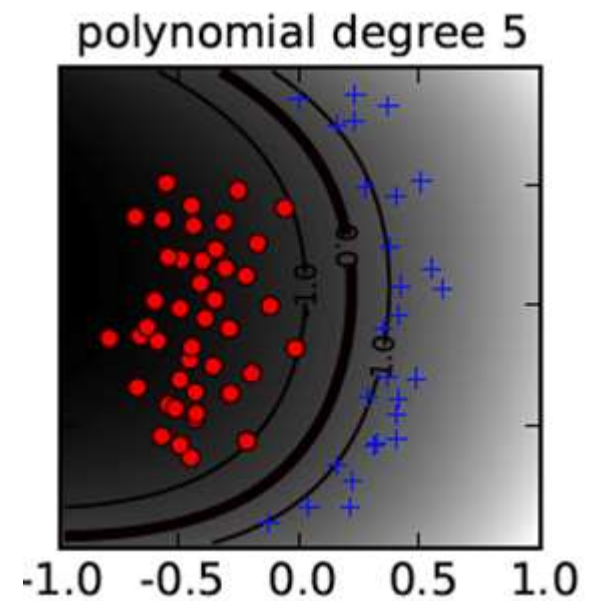
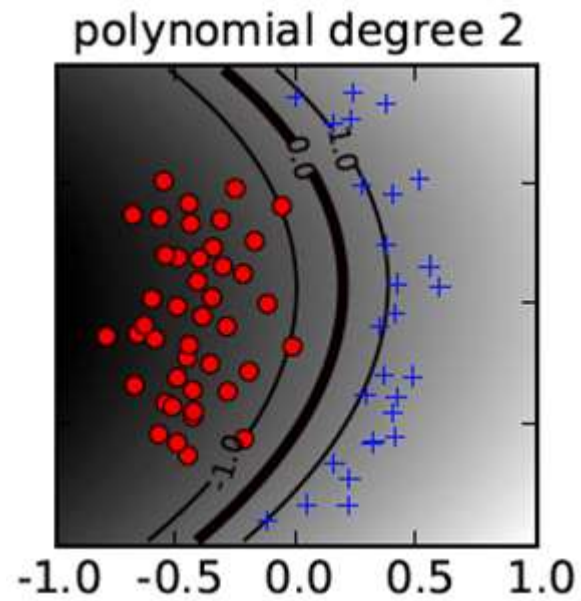
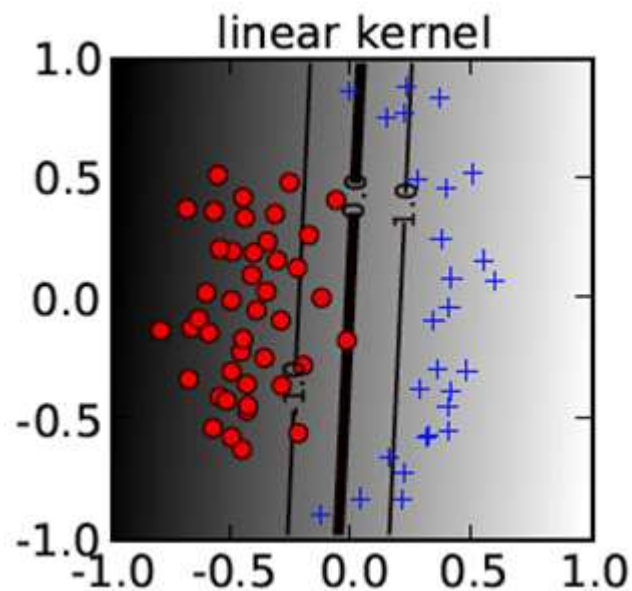
Example



Choosing Kernel Functions

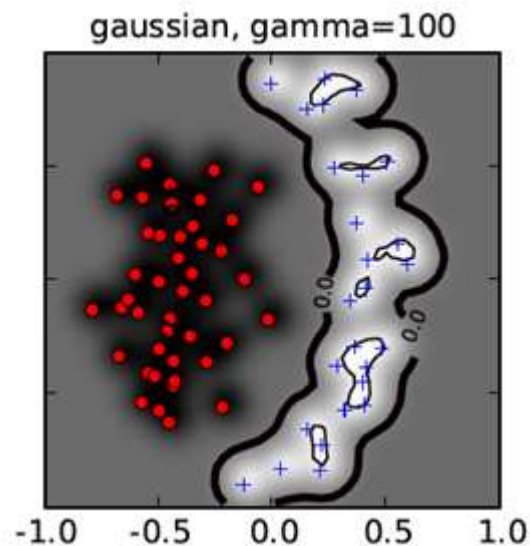
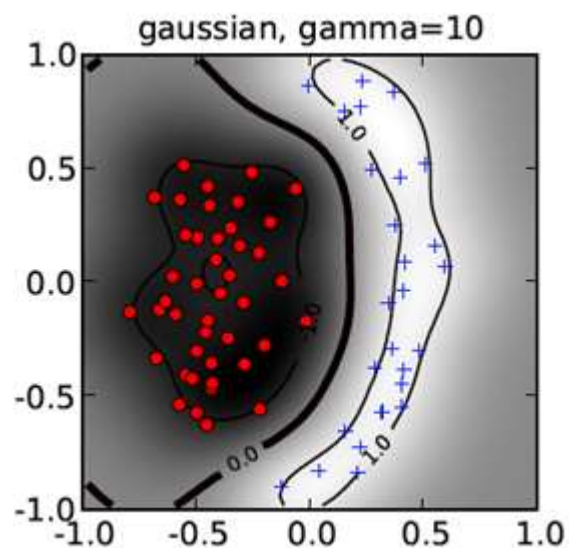
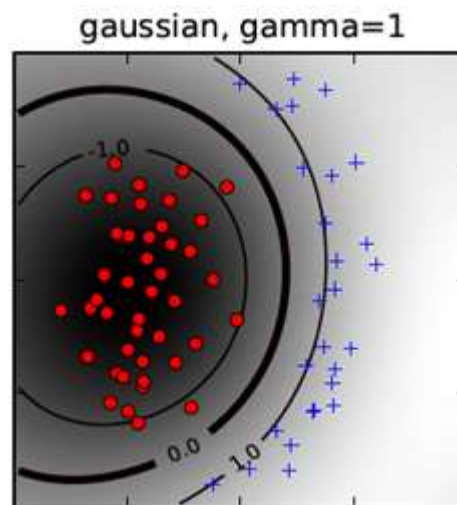
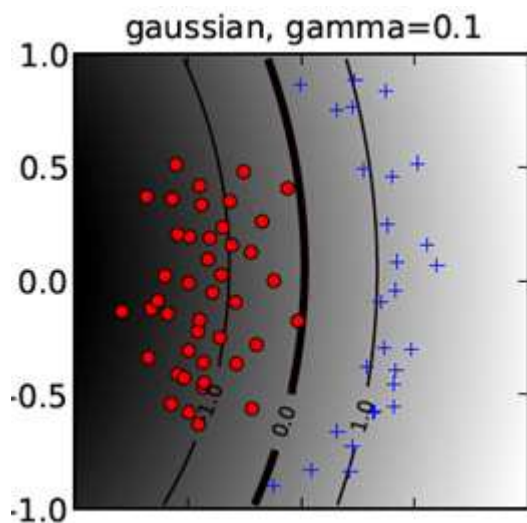
- Probably the most tricky part of using SVM.
- The kernel function is important because it creates the kernel matrix, which summarizes all the data
- Many principles have been proposed (diffusion kernel, Fisher kernel, string kernel, ...)
- There is even research to estimate the kernel matrix from available information
- Multiple Kernel Learning
- In practice, a low degree polynomial kernel or RBF kernel with a reasonable width is a good initial try
- Note that SVM with RBF kernel is closely related to RBF neural networks, with the centers of the radial basis functions automatically chosen for SVM

Polynomial Kernel



RBF Kernel

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / (2\sigma^2))$$



Conclusion and Discussion

Steps in SVM

- Prepare data matrix $\{(x_i, y_i)\}$
- Select a Kernel function
- Select the error parameter C
- “Train” the system (to find all α_i and b)
- New data can be classified using α_i and Support Vectors

Weakness

- Training (Testing) is quite slow compared to ANN
 - Because of Constrained Quadratic Programming
- Essentially a binary classifier
 - However, there are some tricks to evade this.
- Very sensitive to noise. Why?
 - A few off data points can completely throw off the algorithm
- Biggest Drawback: The choice of Kernel function.
 - There is no “set-in-stone” theory for choosing a kernel function for any given problem (still in research...)
 - Once a kernel function is chosen, there is only ONE modifiable parameter, the error penalty C .

Strengths

- Training is relatively easy
 - don't have to deal with local minimum like in ANN
 - SVM solution is always global and unique
- Unlike ANN, doesn't suffer from “curse of dimensionality”
 - How? Why? We have infinite dimensions?!
 - Maximum Margin Constraint: DOT-PRODUCTS!
- Less prone to overfitting
- Simple, easy to understand geometric interpretation.
 - No large networks to mess around with.

Kernelize Logistic Regression (LR)

$$p(y | \vec{x}) = \frac{1}{1 + \exp(-y\vec{x} \cdot \vec{w})}$$

$$l_{reg}(\vec{\alpha}) = \sum_{i=1}^N \log \frac{1}{1 + \exp(-y\vec{x} \cdot \vec{w})} - c \sum_{k=1}^N w_k^2$$

What is the difference between SVM and LR?

- ❑ SVM: L2 norm regularization + hinge loss
- ❑ LR: L2 norm regularization + sigmoid probability (entropy loss)

Merits and drawbacks?

How can we introduce the nonlinearity into the logistic regression?

Kernelize Logistic Regression

$$\vec{x} \rightarrow \vec{\phi}(\vec{x}), \vec{w} = \sum_{i=1}^N \alpha_i \vec{\phi}(\vec{x}_i)$$

$$K(\vec{w}, \vec{x}) = \sum_{i=1}^N \alpha_i K(\vec{x}_i, \vec{x})$$

$$p(y | \vec{x}) = \frac{1}{1 + \exp(-yK(\vec{x}, \vec{w}))} = \frac{1}{1 + \exp\left(-y \sum_{i=1}^N \alpha_i K(\vec{x}_i, \vec{x})\right)}$$

$$l_{reg}(\vec{\alpha}) = \sum_{i=1}^N \log \frac{1}{1 + \exp\left(-y_i \sum_{j=1}^N \alpha_j K(\vec{x}_j, \vec{x}_i)\right)} - c \sum_{i,j=1}^N \alpha_i \alpha_j K(\vec{x}_i, \vec{x}_j)$$

- Representation Theorem
- **Kernelization** of many algorithms (PCA, LDA ...)
 - From linear to non-linear without changing the algorithm

SVM vs Deep Learning

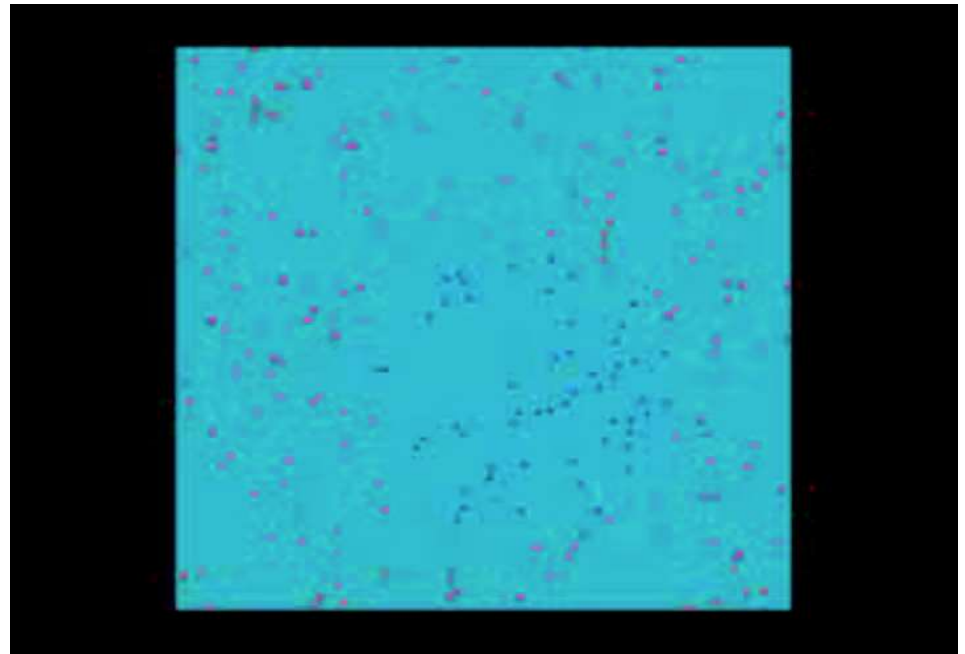
SVM vs Deep Learning

$$K(x, y) = \phi(x)^T \phi(y)$$

SVM (implicit): define K
similarity of data

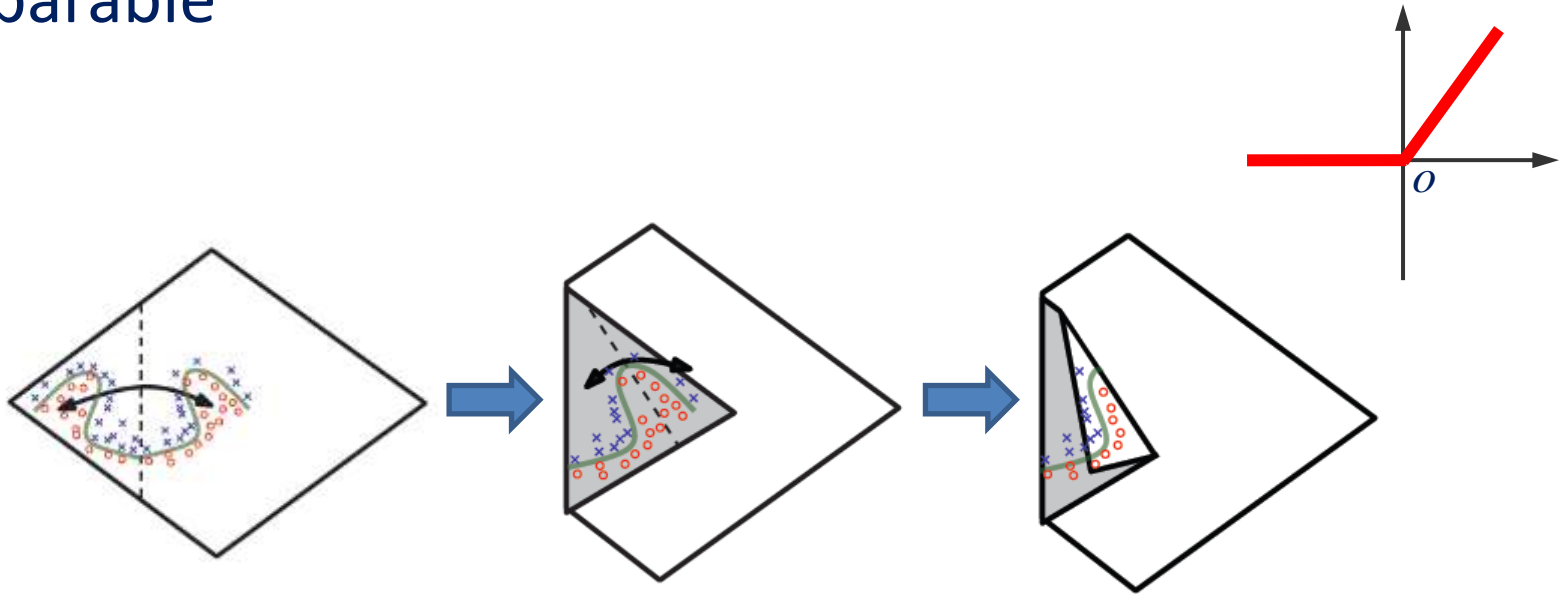


Deep Learning (explicit): define ϕ
learn representation from data

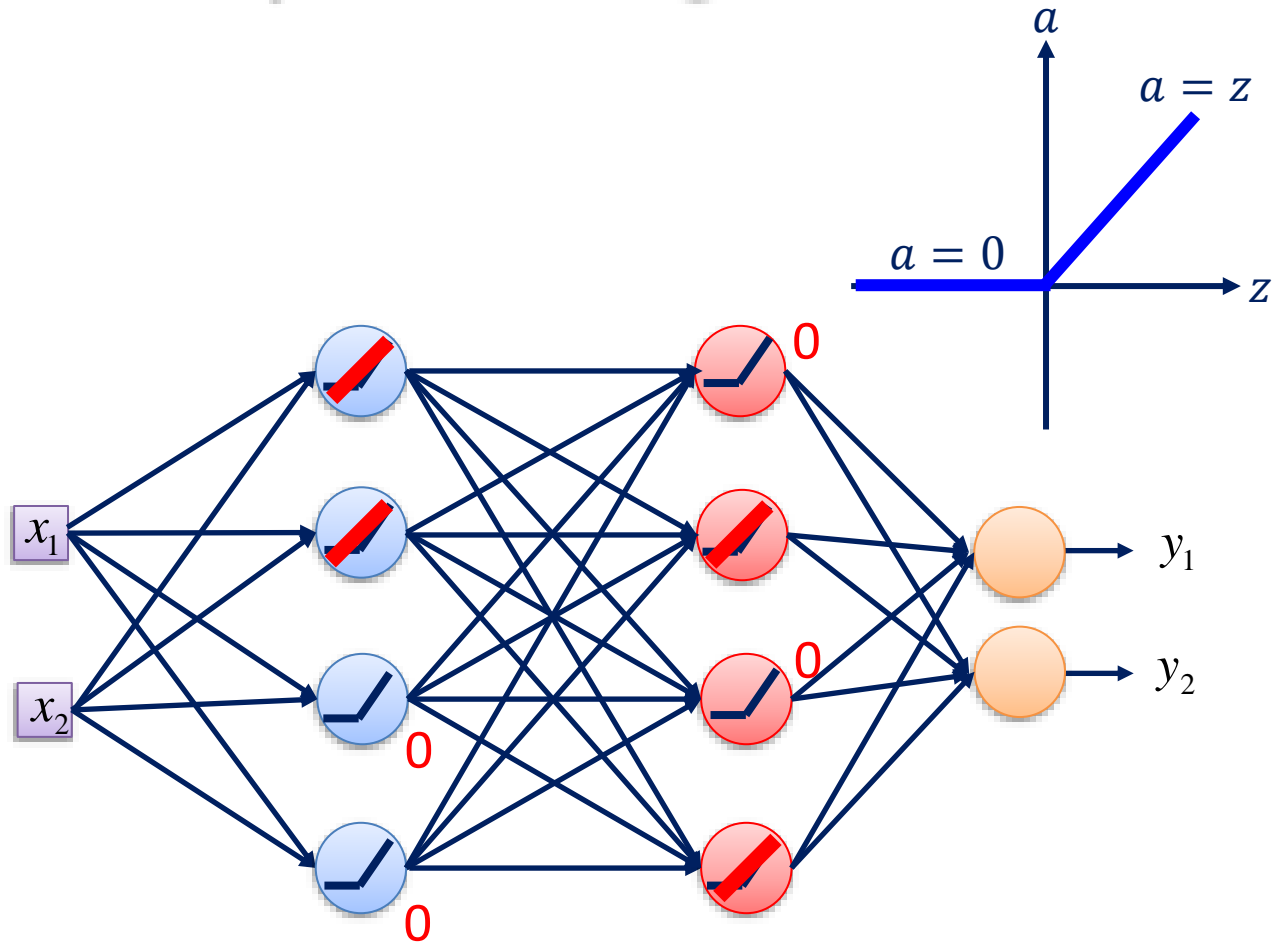


SVM vs Deep Learning

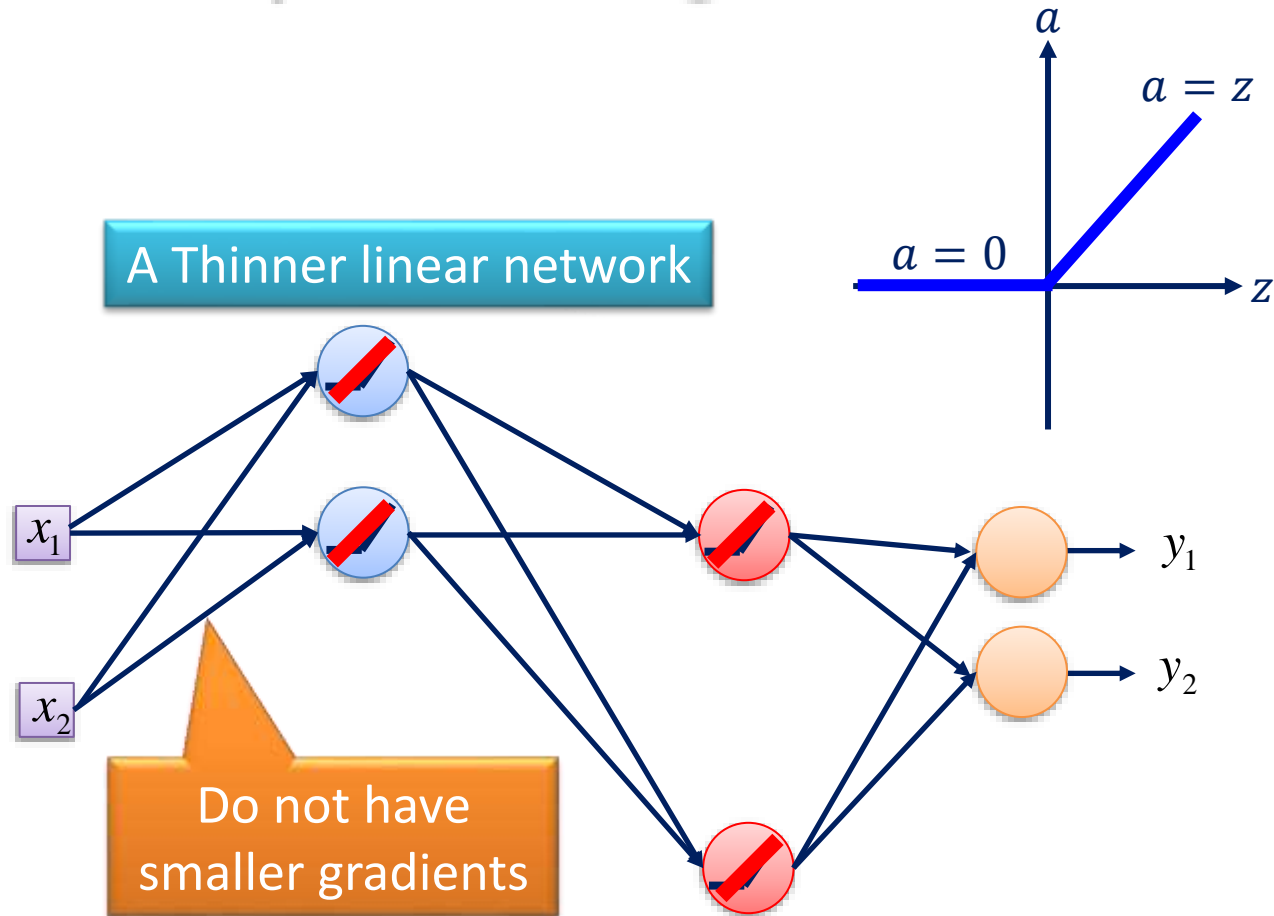
- Repeat: Linear transform + ReLU
- With the increasing of depth, data become linear separable



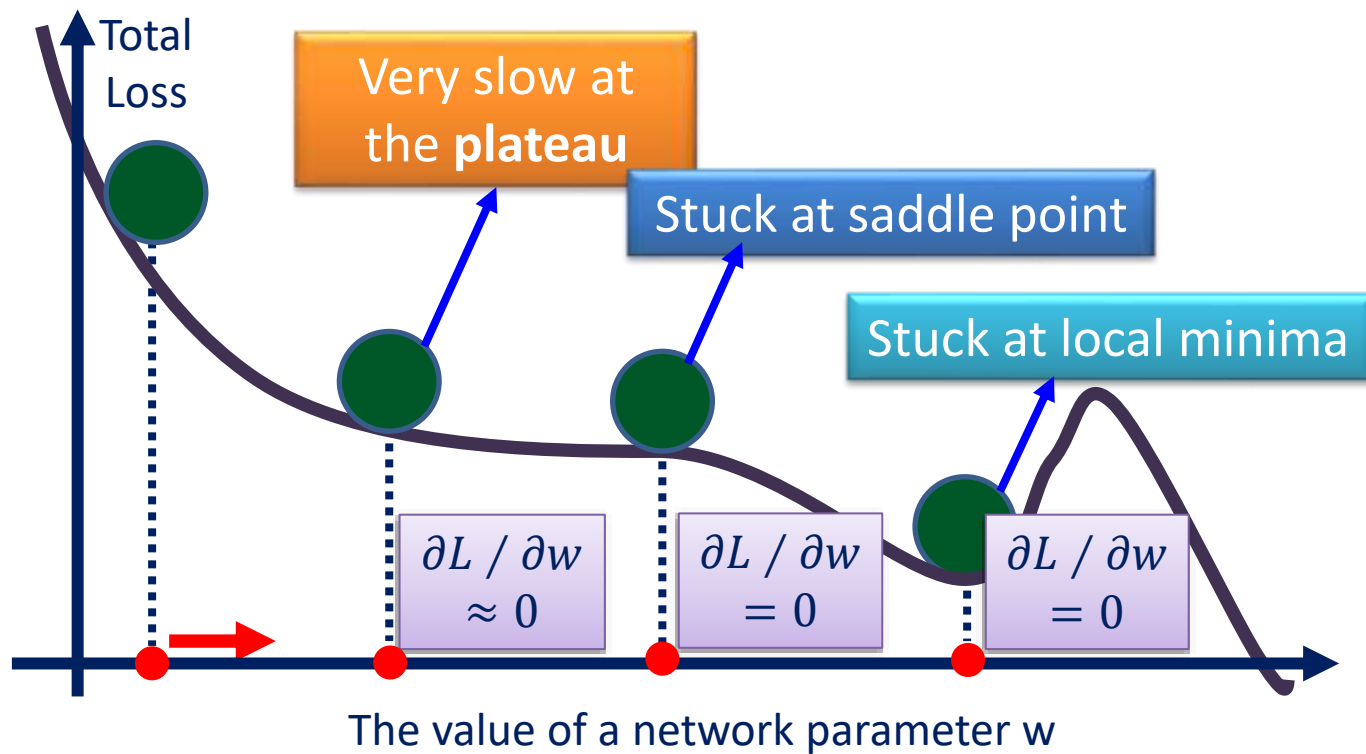
SVM vs Deep Learning



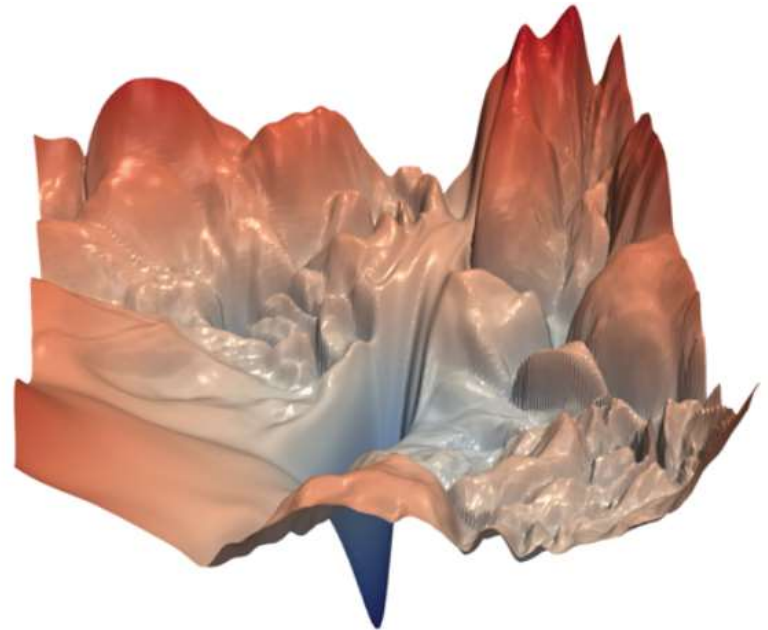
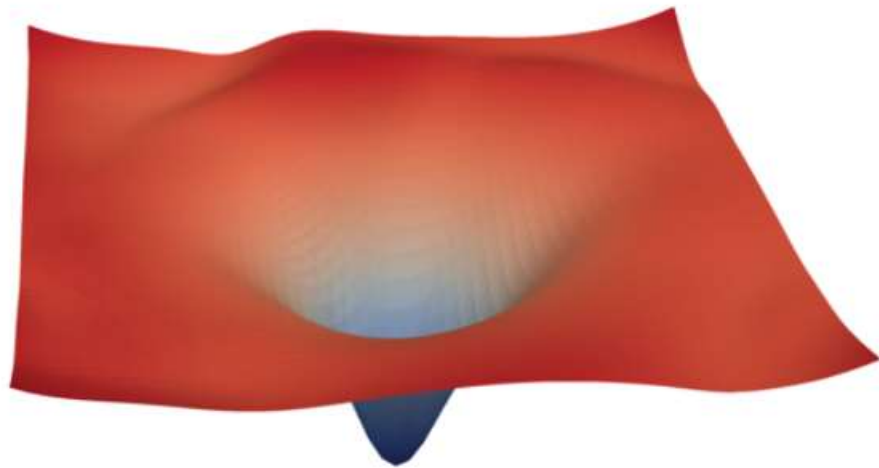
SVM vs Deep Learning



SVM vs Deep Learning

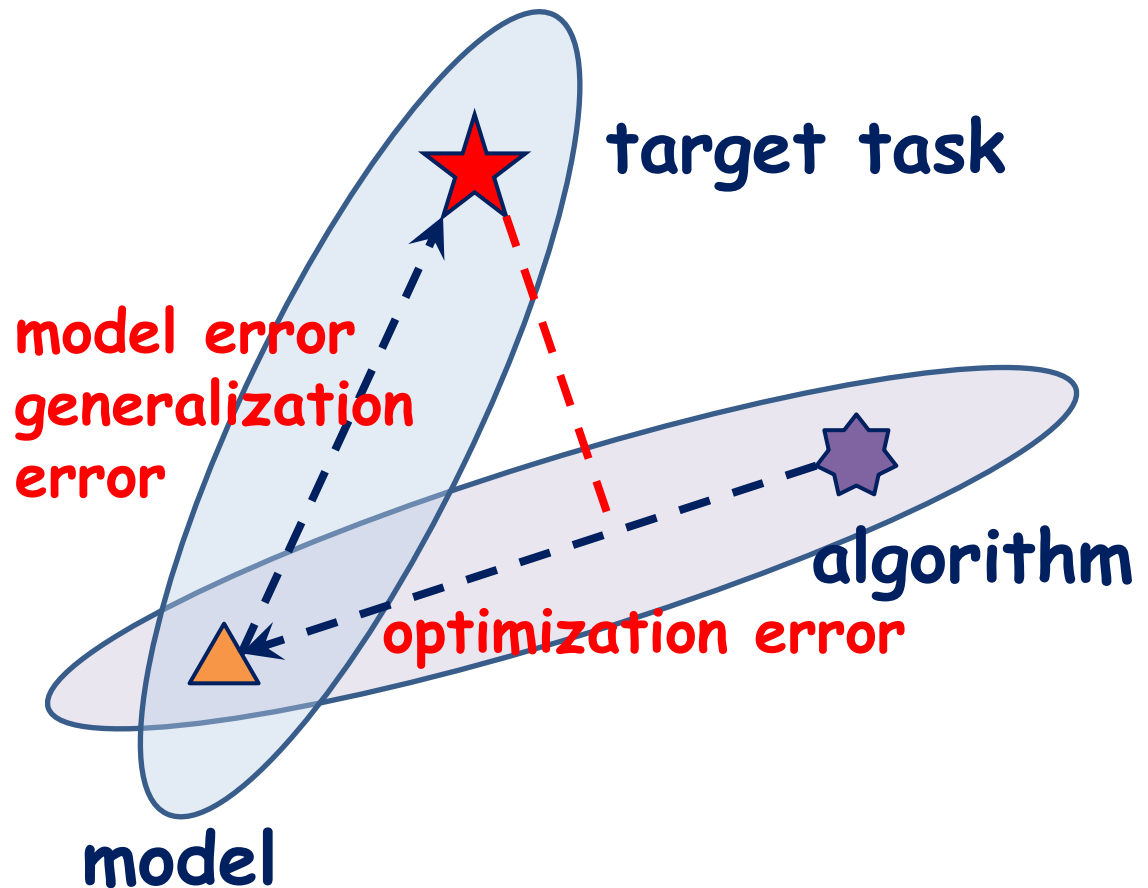


SVM vs Deep Learning



SVM vs Deep Learning

- Machine learning & optimization
- Task \rightarrow model \rightarrow algorithm
- Good optimization \neq Good generalization
- ML = 优化 & 泛化



Thank you!
Q&A