Learning From Data
Lecture 24
The Optimal Hyperplane and Overfitting

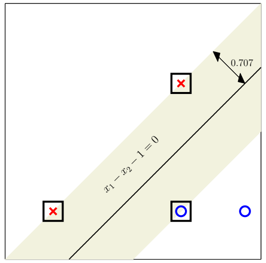
Why is the fattest hyperplane the best?
Non-separable Data

M. Magdon-Ismail
CSCI 4100/6100

RECAP: The Optimal Hyperplane

The Optimal Hyperplane

The fattest hyperplane that separates the data
tolerates most measurement error



1. Can we efficiently find the fattest separator?
2. Is fatter better than thin?

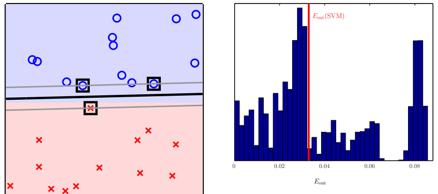
The Algorithm

Quadratic Programming:

$$\underset{b,\mathbf{w}}{\text{minimize}} \quad \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$\text{subject to: } y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1 \quad \text{for } n = 1, \ldots, N.$$

Support vectors: the data points that sit on the cushion.
Using only support vectors, the classifier does not change.



PLA depends on the (random) order of data

Link to Regularization

optimal hyperplane		regularization	
$\underset{b,\mathbf{w}}{\text{minimize}}$	$\tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$	$\underset{\mathbf{w}}{\text{minimize}}$	$E_{\text{in}}(\mathbf{w})$
subject to:	$y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1$ for $n = 1, \ldots, N.$	subject to:	$\mathbf{w}^{\mathrm{T}}\mathbf{w} \leq C.$

	optimal hyperplane	regularization
minimize	$\mathbf{w}^{\mathrm{T}}\mathbf{w}$	E_{in}
subject to	$E_{\text{in}} = 0$	$\mathbf{w}^{\mathrm{T}}\mathbf{w} \leq C$

The optimal hyperplane performs ‘automatic’ regularization.

Evidence that Larger Margin is Better

(1) Experimental: larger margin gives lower E_{out}; bias drops a little and var a lot.

(2) Bound for d_{VC} can be less than $d + 1$ – fat hyperplanes generalize better.

(3) E_{cv} bound does not explicitly depend on d.

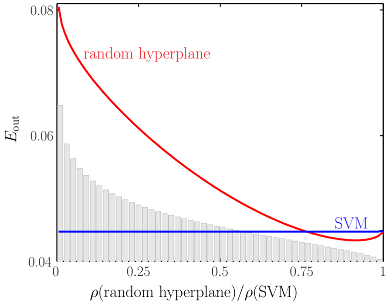
Larger Margin is Better

Generate a random separable data set ($N = 20$)

Select 50,000 random separating hyperplanes h

Compute E_{out} and $\rho(h)/\rho(SVM)$

Average over several thousands of random data sets



(Histogram shows relative frequency of different margins)

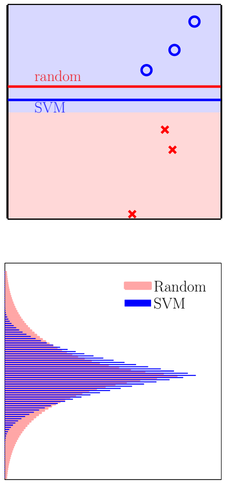
Bigger margin is generally better

Biggest is not best.

⟵ Data other than support vectors can have role in fine-tuning

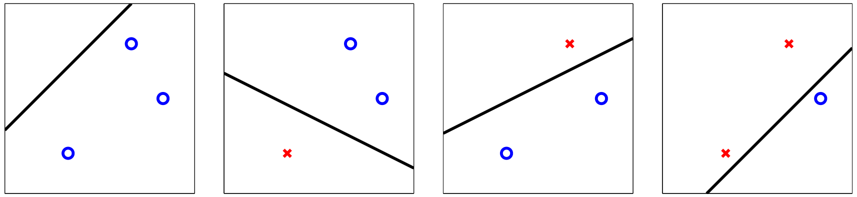
---

Bias and Variance

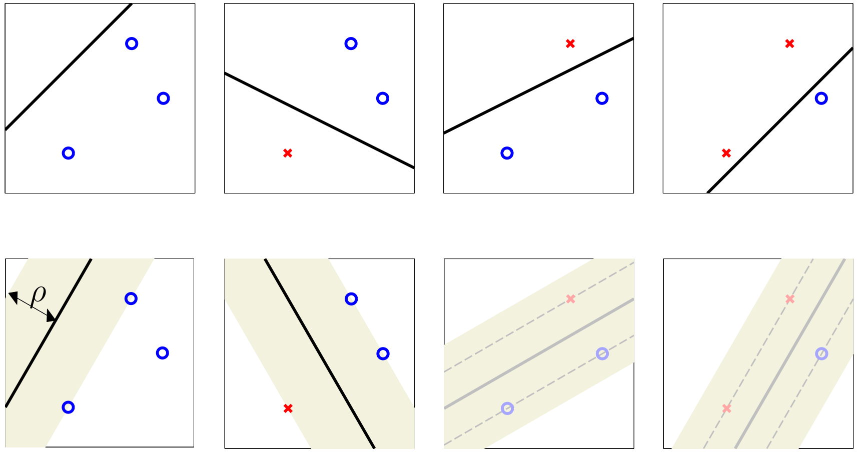


	Random	SVM	
bias	0.02	0.015	**-0.005**
var	0.059	0.038	**-0.021**
E_{out}	0.079	0.053	**-0.026**

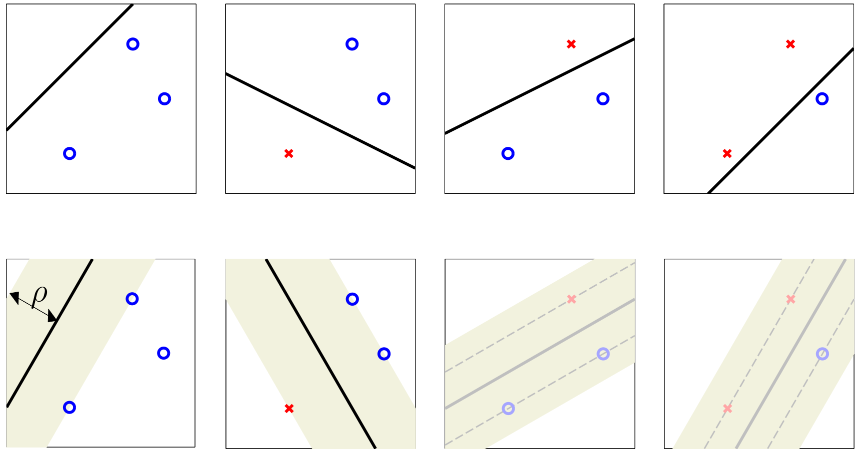
Fat Hyperplanes Shatter Fewer Points



Fat Hyperplanes Shatter Fewer Points



Fat Hyperplanes Shatter Fewer Points

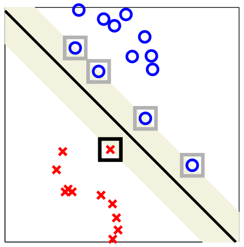


Theorem. $d_{\text{VC}}(\gamma) \leq \left\lceil \frac{R^2}{\gamma^2} \right\rceil + 1$

A Bound on E_{cv}

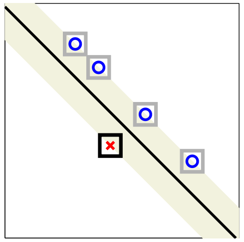
$$E_{\text{cv}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{e}_n$$

(unbiased estimate of $E_{\text{out}}(N-1)$)



$$E_{\text{cv}} \leq \frac{\#\text{ support vectors}}{N}$$

(no explicit dependence on d)



Summary of Hyperplanes and Generalization

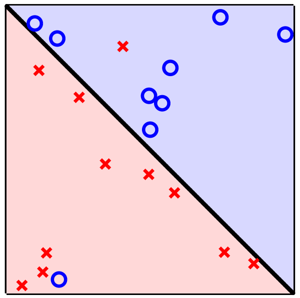
Algorithm For Selecting Separating Hyperplane

General	PLA	SVM
$d_{\text{VC}} = d+1$	$E_{\text{cv}} \leq \frac{R^2}{N\rho^2}$	bias↓ var↓↓ $d_{\text{VC}}(\rho) \leq \left\lceil \frac{R^2}{\rho^2} \right\rceil + 1$ $E_{\text{cv}} \leq \frac{\#\text{ support vectors}}{N}$

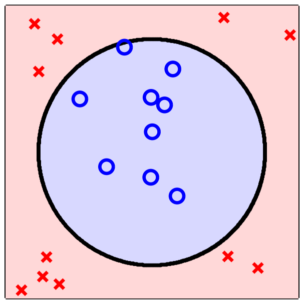
Generalization performance controlled by quantities not explicitly depending on d

---

Non-Separable Data

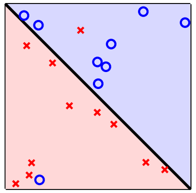


tolerate error



nonlinear transform

Soft Margin SVM



tolerate error

$$\underset{b,\mathbf{w},\boldsymbol{\xi}}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{n=1}^{N}\xi_n$$

↙ 'soft in-sample error'

$$\text{subject to:} \quad y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1 - \xi_n$$

↙ 'soft' error on $(\mathbf{x}_n, y_n)$

$$\xi_n \geq 0 \qquad \text{for } n = 1, \ldots, N$$

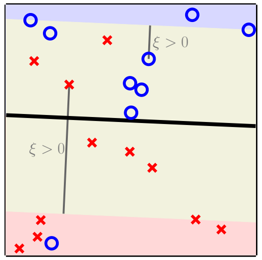
Trades off 'soft in-sample error' $\sum_{n=1}^{N}\xi_n$ with weight norm $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$ ⟵ regularization

C plays the role of a regularization parameter ($\lambda \sim \frac{1}{C}$)

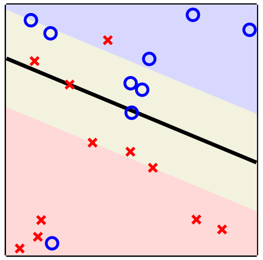
Choice of C is important - similar to choice of λ in regularization

---

Non-Separable Data



$C = 1$



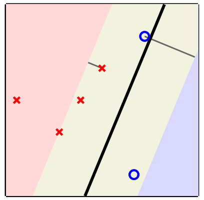
$C = 500$

$$\underset{b,\mathbf{w},\boldsymbol{\xi}}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{n=1}^{N}\xi_n$$

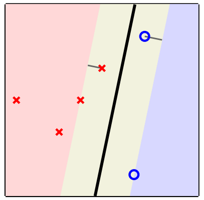
$$\text{subject to:} \quad y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1 - \xi_n$$

$$\xi_n \geq 0 \qquad \text{for } n = 1, \ldots, N$$

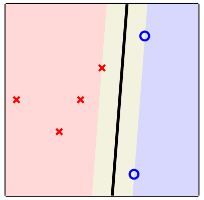
Soft Margin SVM With Separable Data



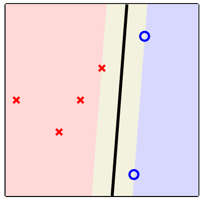
small C



medium C



large C



hard margin

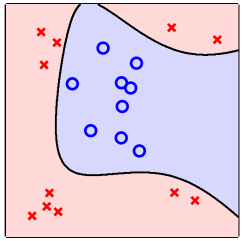
$$\underset{b,\mathbf{w},\boldsymbol{\xi}}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{n=1}^{N}\xi_n$$

$$\text{subject to:} \quad y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1 - \xi_n$$

$$\xi_n \geq 0 \qquad \text{for } n = 1, \ldots, N$$

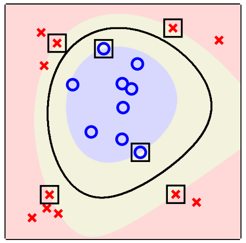
Choice of C is
IMPORTANT

Nonlinear Transform and SVM

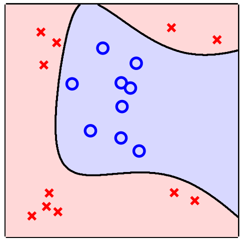


Φ_3+ pseudoinverse algorithm

Nonlinear Transform and SVM

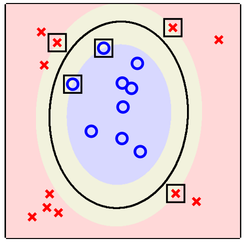


Φ_3+ SVM

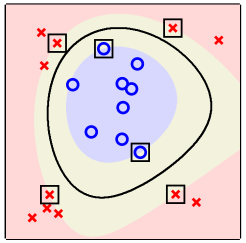


Φ_3+ pseudoinverse algorithm

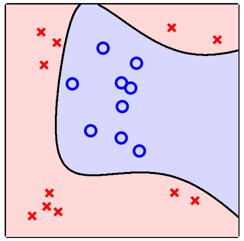
Nonlinear Transform and SVM



Φ_2+ SVM



Φ_3+ SVM



Φ_3+ pseudoinverse algorithm

Observations:

1. Φ_3 has almost 2× the parameters of Φ_2
2. Φ_3-SVM does not display significant overfitting compared to Φ_3-regression
3. #support vectors did not double
4. Can go to higher dimensions if #support vectors stays small or margin stays large

	pseudoinverse regression		SVM	
	linear	nonlinear (ϕ)	linear	nonlinear (ϕ)
overfitting	little	lots	tiny	ok
boundary	linear	complex	linear	complex

Going to Even Higher Dimension

In higher dimension, can control overfitting with # support vectors or margin ρ

What about:

Efficiency?

Infinitely many dimensions?