INSOFE
Inspire...Educate...Transform.






Inspire…Educate…Transform.

Predictive Analytics

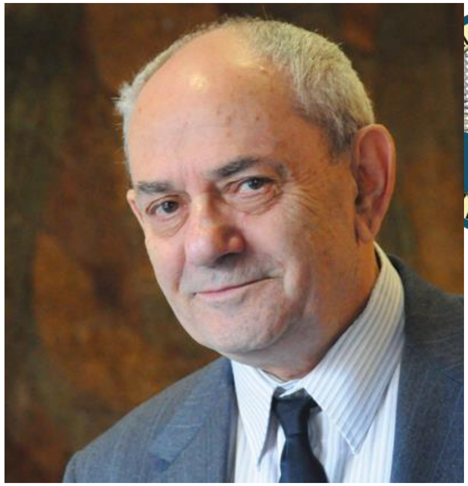
SVM

Dr. Suryaprakash Kompalli
Senior Mentor, INSOFE

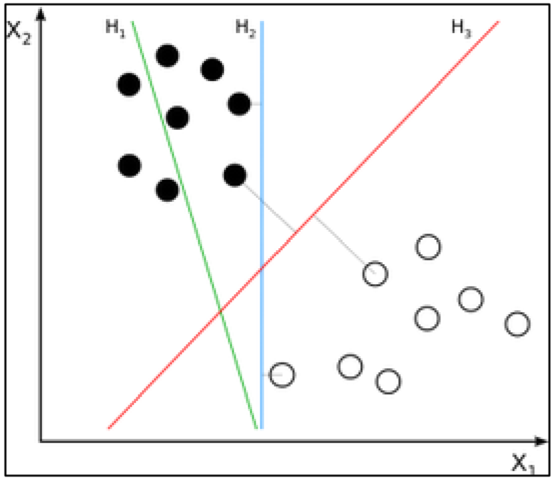
SUPPORT VECTOR MACHINES

SVM – Curtain Raiser

- Support Vector Machines is arguably the most important & interesting recent discovery in Machine Learning.
- SVMs have a clever way to prevent over-fitting
- SVMs have a very clever way to use a huge number of features without requiring nearly as much computation as seems to be necessary.



Vladimir Vapnik
Invented SVM in 1963!!!
Non-linear classifier in 1992!!!



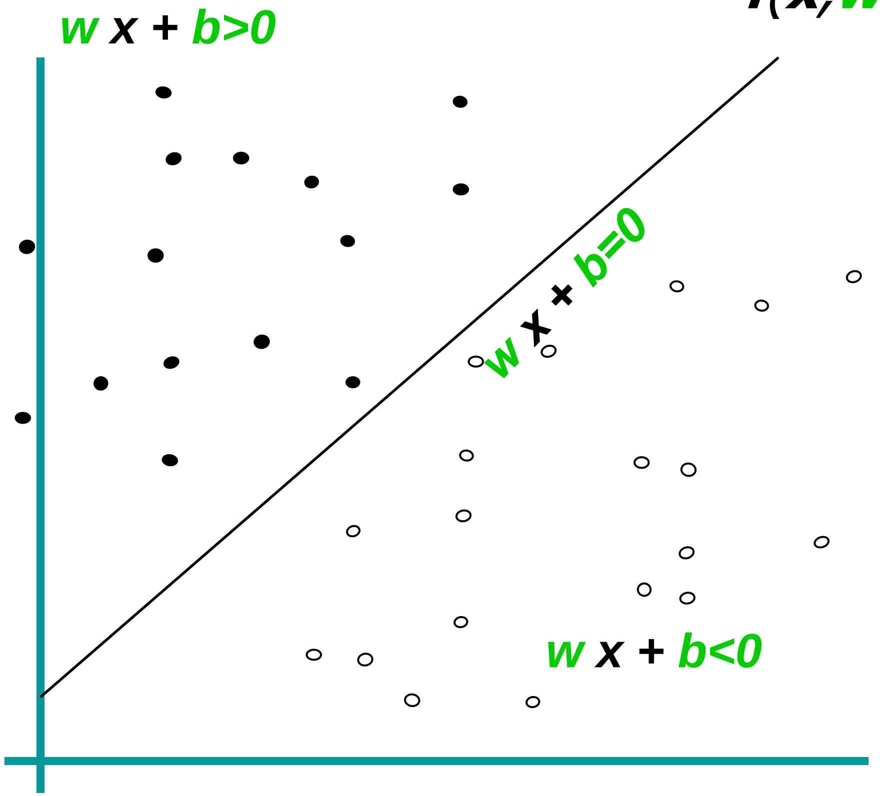
H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximum margin. Source: https://en.wikipedia.org/wiki/Support_vector_machine

CSE 7305c

Linear Classifiers

- • denotes +1
- ◦ denotes -1

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



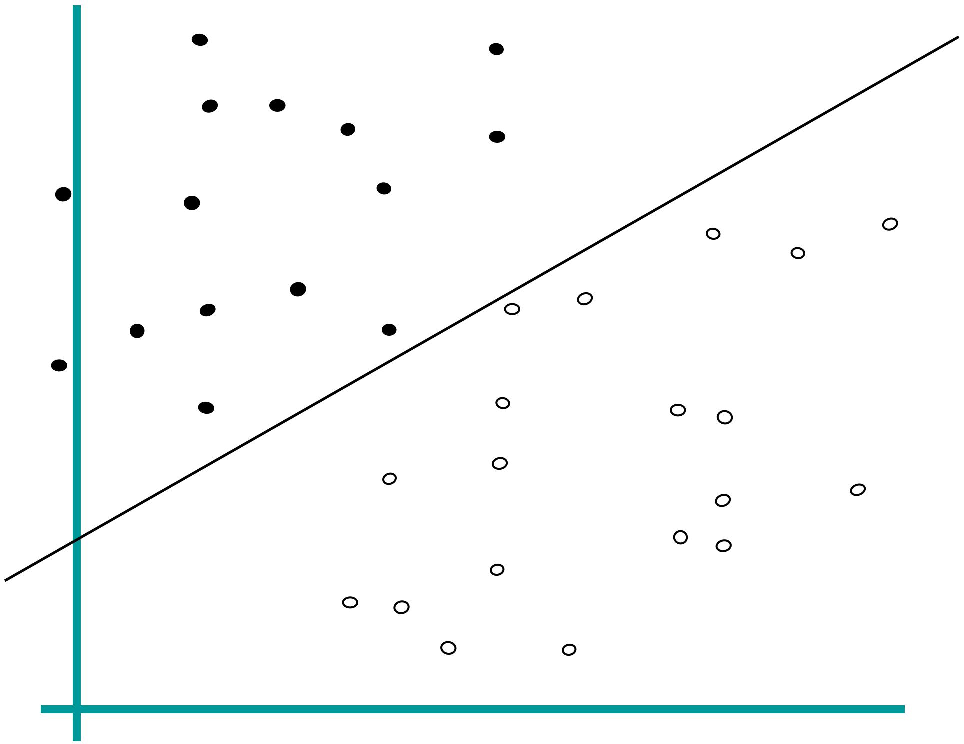
How would you classify this data?

Is a single line the only solution?

Linear Classifiers

- • denotes +1
- ◦ denotes -1

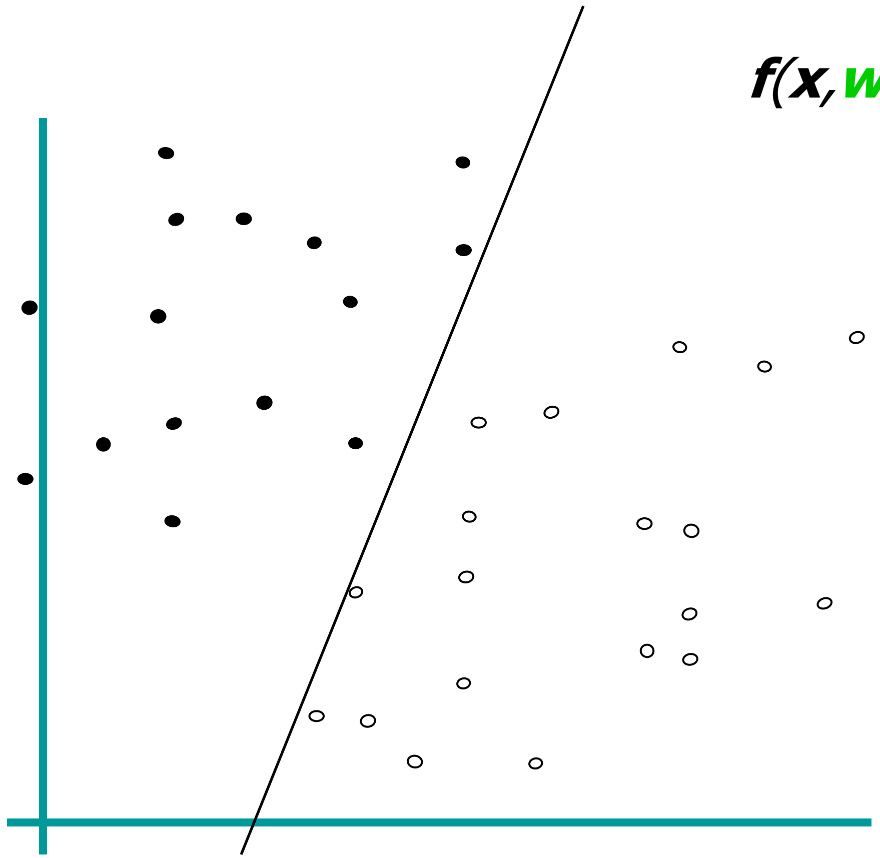
$$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}\,\mathbf{x} + b)$$



How would you classify this data?

Linear Classifiers

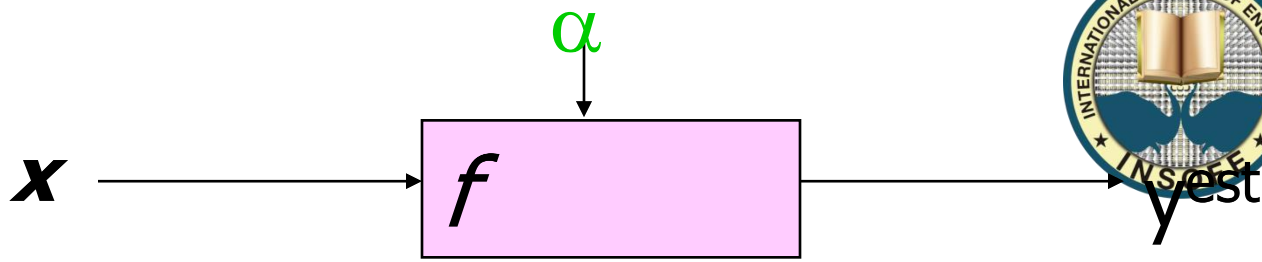
- • denotes +1
- ° denotes -1



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

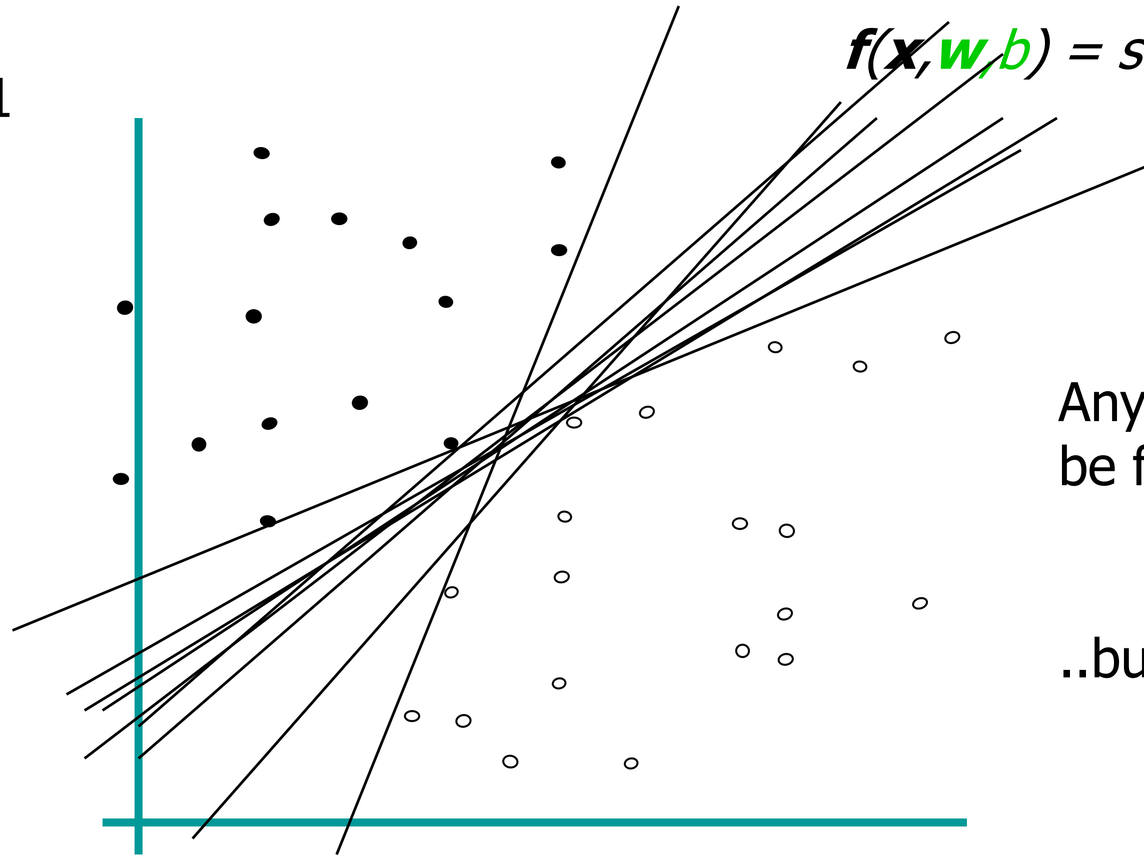
How would you classify this data?

Linear Classifiers



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

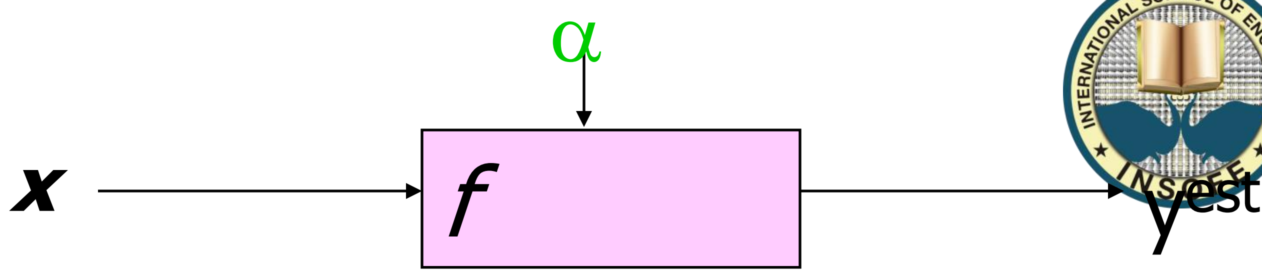
- ● denotes +1
- ○ denotes -1



Any of these would be fine..

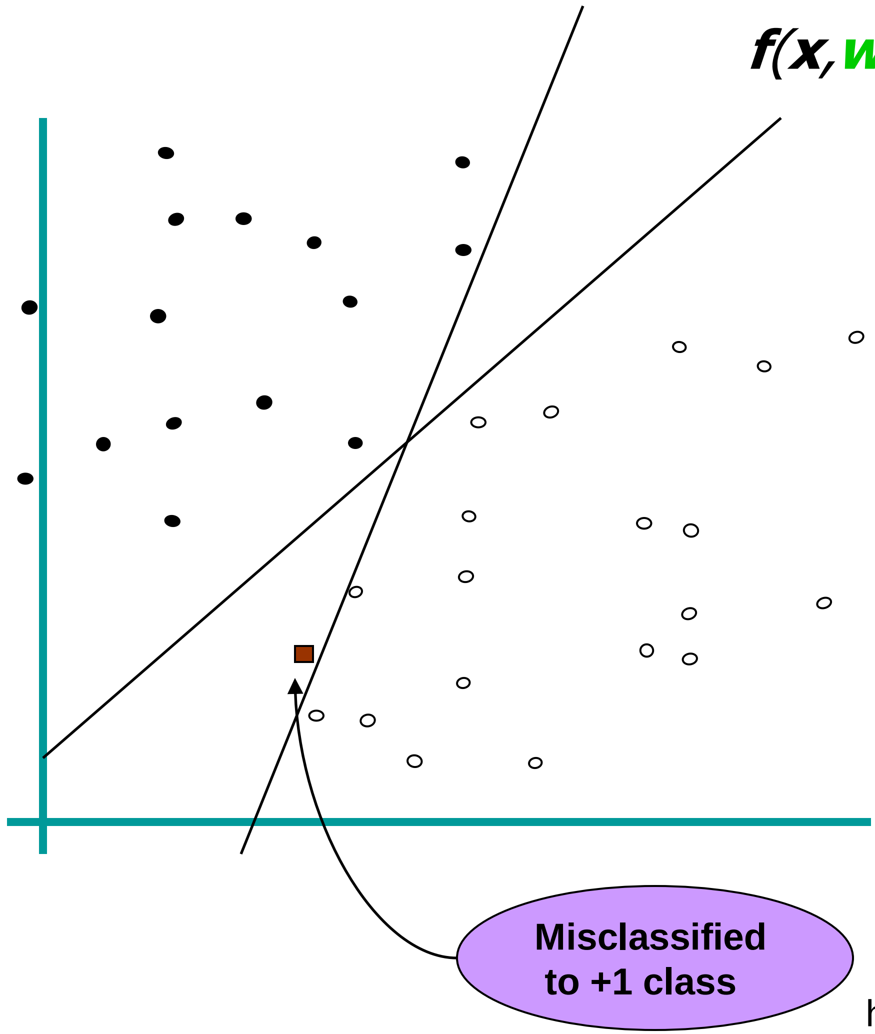
..but which is best?

Linear Classifiers



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

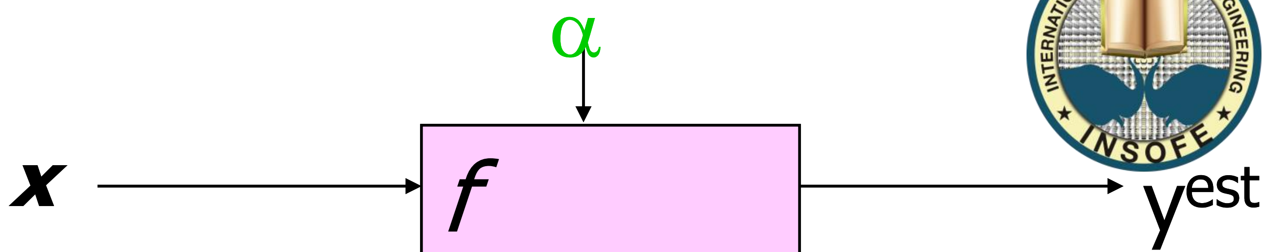
- • denotes +1
- ○ denotes -1



How would you classify this data?

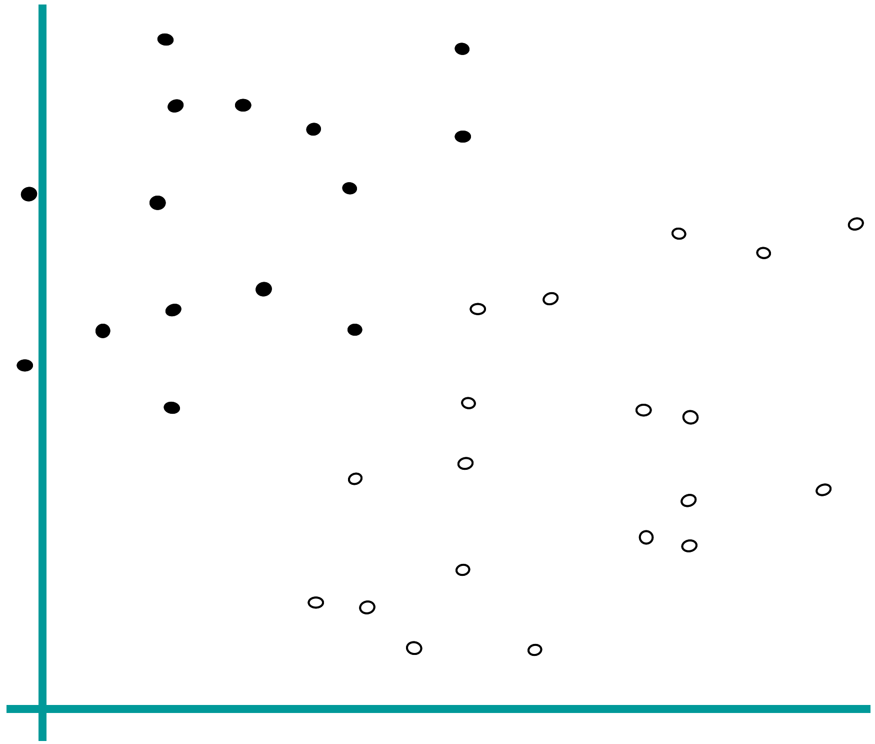
http://www.autonlab.org/tutorials/svm15.pdf

Classifier Margin



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

- • denotes +1
- ◦ denotes -1



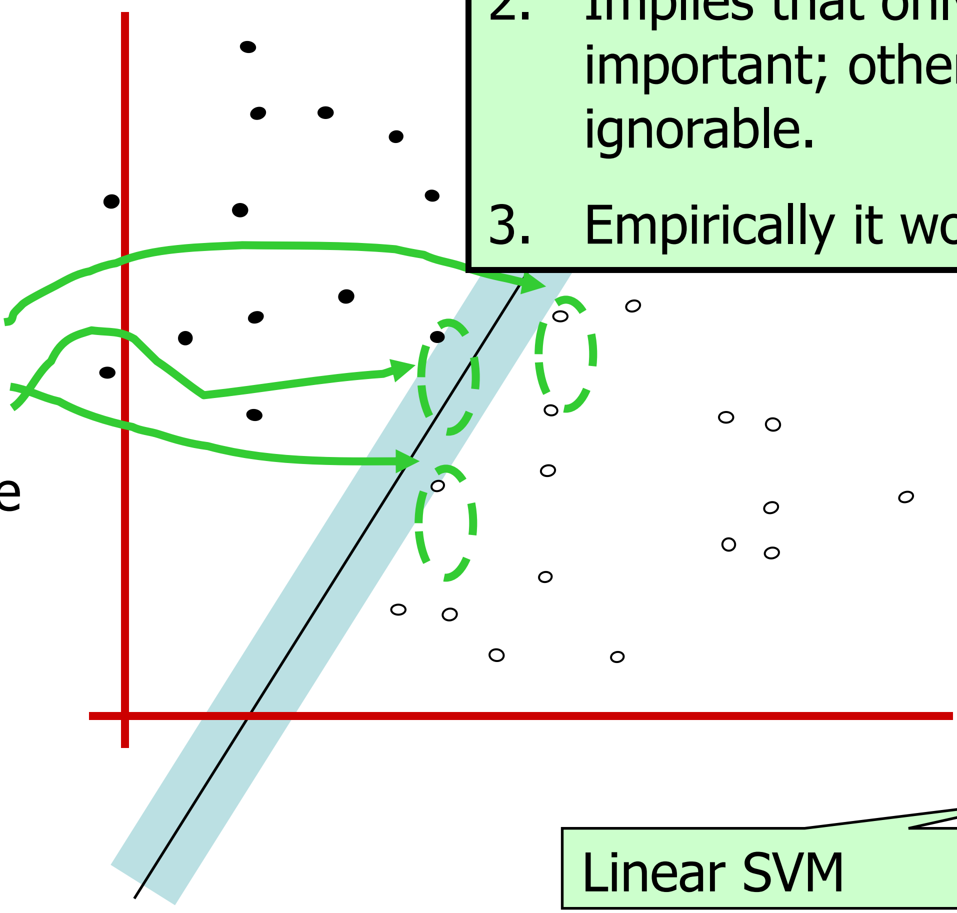
Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

Maximum Margin

$\boldsymbol{x}$ → f → y^{est}, with α feeding into f

1. Maximizing the margin is good according to intuition and PAC theory
2. Implies that only support vectors are important; other training examples are ignorable.
3. Empirically it works very very well.

- ● denotes +1
- ○ denotes -1



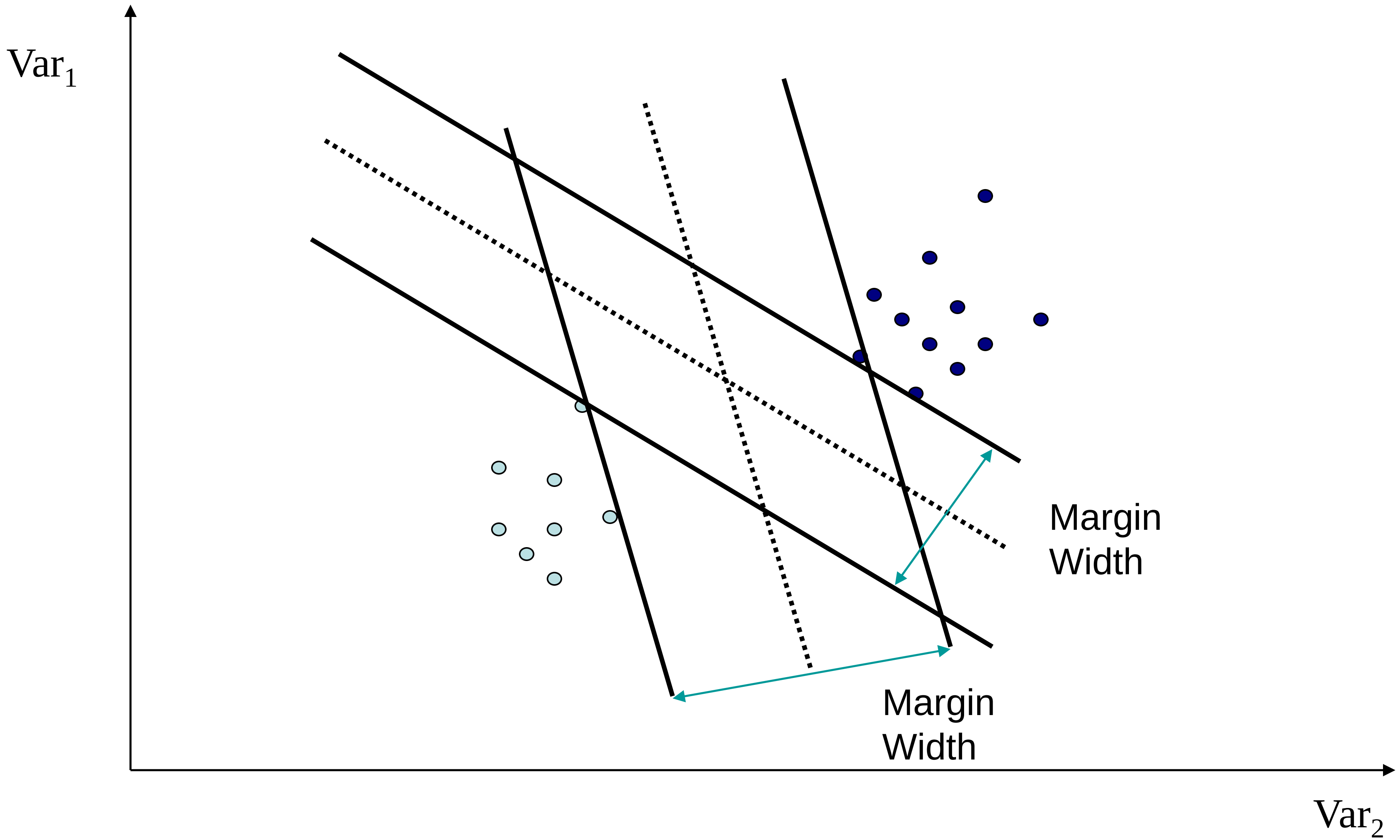
Support Vectors are those datapoints that the margin pushes up against

classifier with the, um, maximum margin.

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

Linear SVM

Maximizing the Margin



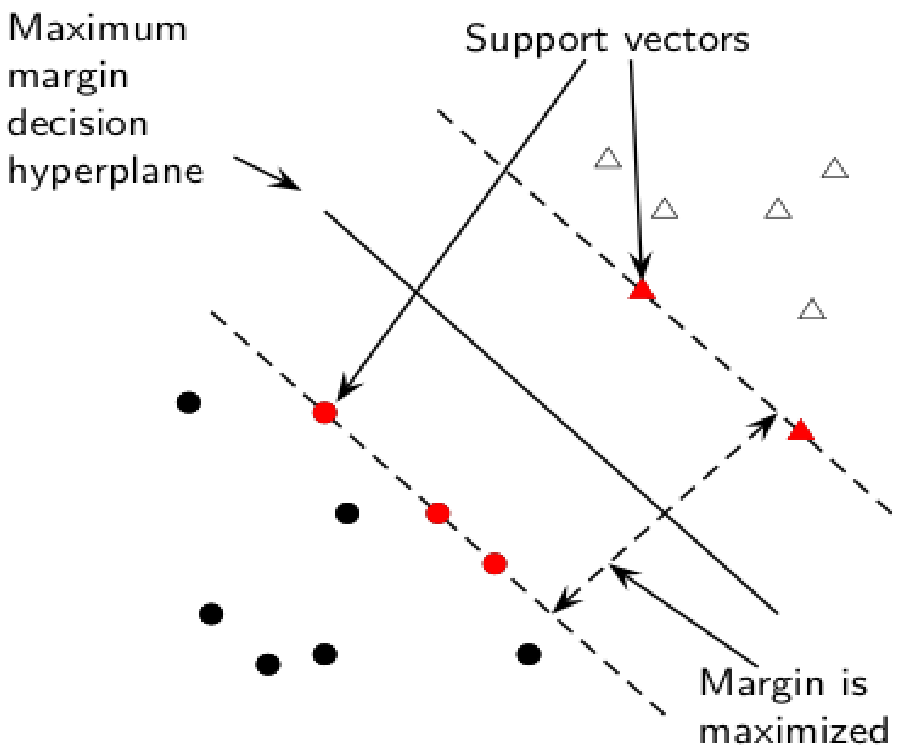
CSE 7305c

Why maximize the margin?

Points near decision surface → uncertain classification decisions (50% either way).

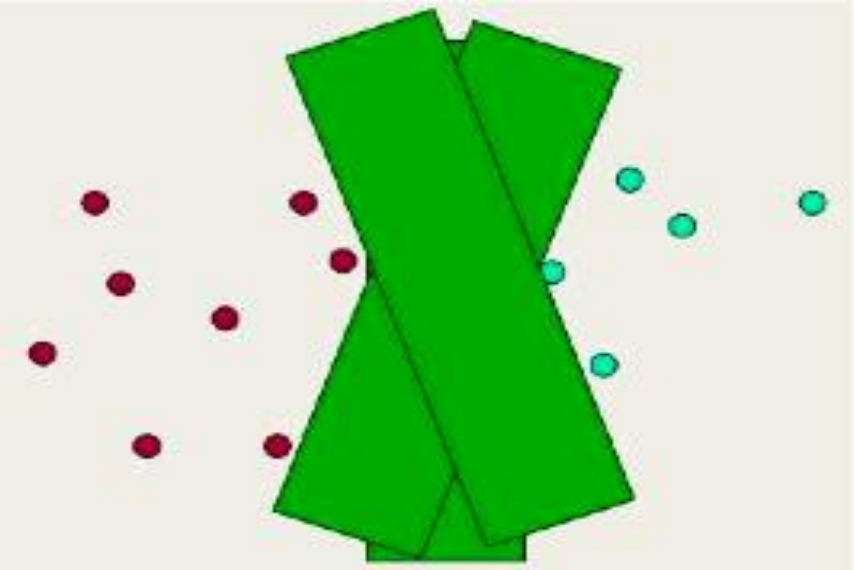
A classifier with a large margin makes no low certainty classification decisions.

Gives classification safety margin w.r.t slight errors in measurement



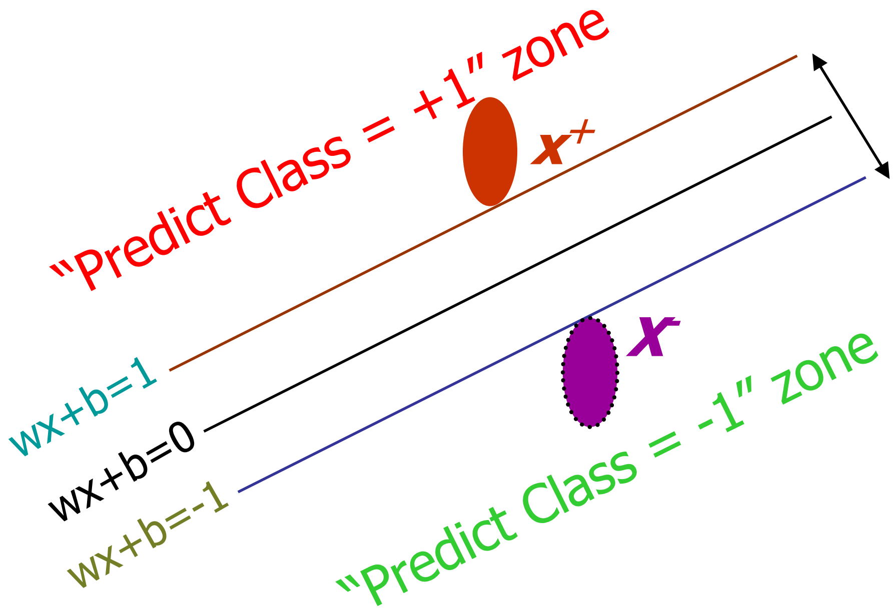
Why maximize the margin?

- SVM classifier: large margin around decision boundary
- Compare to decision hyperplane: place fat separator between classes
- Fewer choices of where it can be put
  - Decreased memory capacity
  - Increased ability to correctly generalize to test data



Linear SVM Mathematically

$\mathbf{w} . (\mathbf{x}^+ \mathbf{-} \mathbf{x}^+) = 2$



M=Margin Width

$\boldsymbol{M} = |\boldsymbol{x}^+ \mathbf{-} \boldsymbol{x}^-|$

Let us express the planes as:

$\mathbf{w} . \mathbf{x}^+ + b = +1$

$\mathbf{w} . \mathbf{x}^- + b = -1$

From above two equations:

$\mathbf{w} . (\mathbf{x}^+ \mathbf{-} \mathbf{x}^-) = 2$

If U and V are vectors on x^+ *plane*

$\mathbf{w} . \mathbf{U} + b = +1$

$\mathbf{w} . \mathbf{V} + b = +1$

$\mathbf{w} . (\mathbf{U-V}) = 0$, *hence* ***w*** *is perpendicular to* x^+

x^+ and x^- are parallel to each other

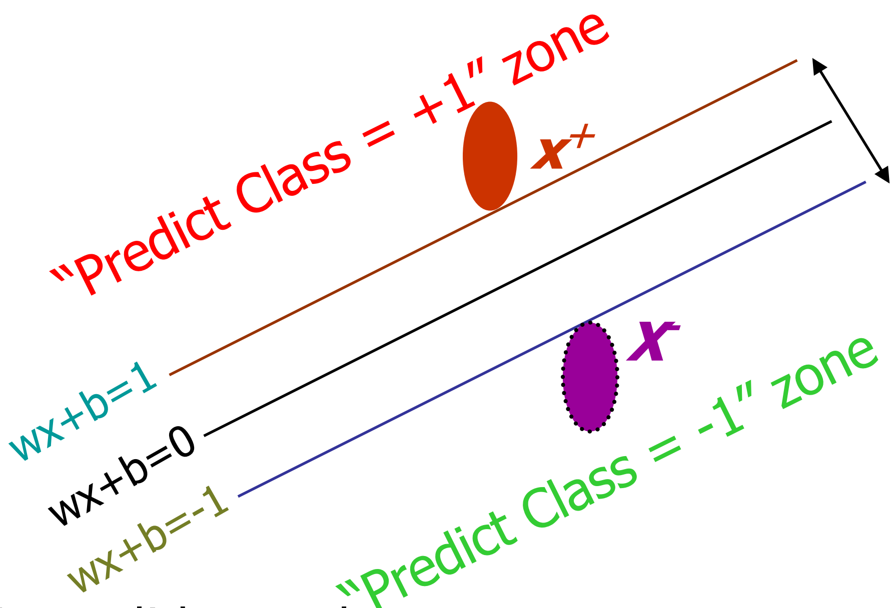
From above two observations, we can say:

$\mathbf{x}^+ = \mathbf{x}^- + k\,\mathbf{w}$

http://www.cs.cmu.edu/~awm/tutorials

CSE 7305c

Linear SVM Mathematically



M=Margin Width

$\boldsymbol{M = |x^+ - x^-|}$

From previous slide, we know:

$\mathbf{w} \, . \, (\mathbf{x}^+ - \mathbf{x}^-) = 2$

$\mathbf{w}^{-1} \, . \, \mathbf{w} \, . \, (\mathbf{x}^+ - \mathbf{x}^-) = \mathbf{w}^{-1} \, . \, 2$

$(\mathbf{x}^+ - \mathbf{x}^-) = \mathbf{w}^{-1} \, . \, 2$ ... *Eq 1*

From Eq 1 and Eq 2:

$\mathbf{w}^{-1} \, . \, 2 = k \, \mathbf{w}$

$k = \frac{w^{-1}.2}{w} = \frac{2}{w.w}$…Eq 3

From previous slide, we know:

$\mathbf{x}^+ = \mathbf{x}^- + k \, \mathbf{w}$

$\mathbf{x}^+ - \mathbf{x}^- = k \, \mathbf{w}$ … *Eq 2*

Margin, $\boldsymbol{M} = |\mathbf{x}^+ - \mathbf{x}^-| = |k\mathbf{w}| = \mathbf{k}|\mathbf{w}| = k\sqrt{w.w}$

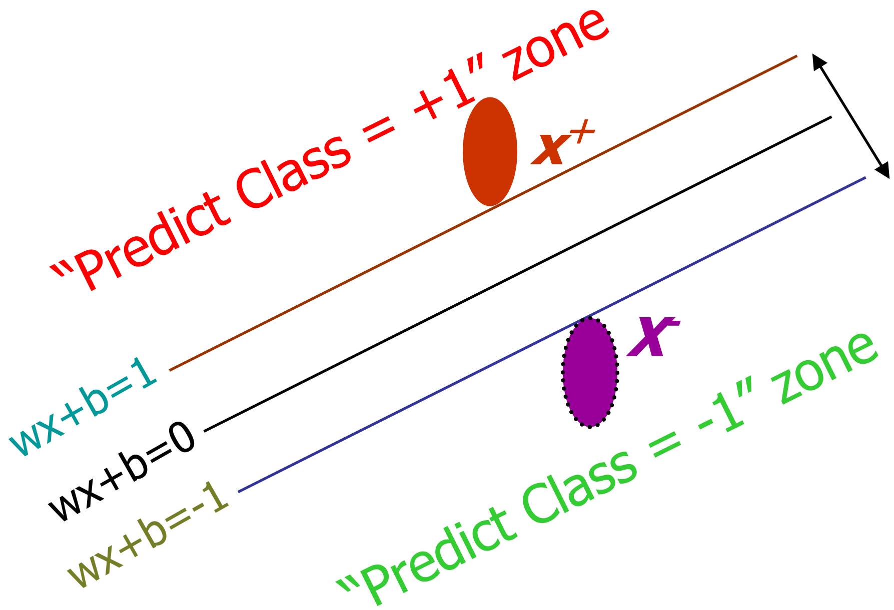
Substituting k from Eq 3 we have:

$$M = \frac{2\sqrt{w.w}}{w.w} = \frac{2}{\sqrt{w.w}}$$

http://www.cs.cmu.edu/~awm/tutorials http://www.autonlab.org/tutorials/svm15.pdf

CSE 7305c

Linear SVM Mathematically



M=Margin Width

$M = |x^+ - x^-|$

Given a set of data points that satisfy:

$\boldsymbol{w} \cdot \boldsymbol{x}^+ + b = +1$

$\boldsymbol{w} \cdot \boldsymbol{x}^- + b = -1$

Find a value of "w" and "b" such that:

$M = \frac{2}{\sqrt{w.w}}$ is maximized

Above is the same as :

$\frac{1}{2}\sqrt{w.w}$ is minimized

http://www.autonlab.org/tutorials/svm15.pdf

CSE 7305c

Linear SVM Mathematically

- Goal: **1) Correctly classify all training data**

$$\left.\begin{aligned} wx_i + b &\geq 1 \\ wx_i + b &\leq -1 \end{aligned}\right\}$$

$$y_i(wx_i + b) \geq 1$$

$$y_i(w.x_i + b) - 1 \geq 0 \text{ for all i}$$

$$\frac{1}{2}\sqrt{w.w}$$

2) Minimize

- **This is a Quadratic Optimization Problem in linear constraints. Solve for w & b.**

Linear SVM Mathematically

- Goal: **1) Correctly classify all training data**

$$\left.\begin{array}{r} w.x_i + b \geq 1 \; if \; y_i = +1 \\ w.x_i + b \leq -1 \; if \; y_i = -1 \end{array}\right\}$$

$$y_i(w.x_i + b) - 1 \geq 0 \text{ for all i}$$

2) Minimize $\frac{1}{2}\sqrt{w.w}$

- **This is a Quadratic Optimization Problem in linear constraints. Solve for w & b.**

Linear (hard-margin) SVM - formulation

- Find *w,b* that solves

$$\min \frac{1}{2}\|w\|^2$$
$$s.t.\ y_i(w \cdot x_i + b) \geq 1,\ \forall x_i$$

- Problem is convex so, there is a unique global minimum value (when feasible)
- Non-solvable if the data is not linearly separable
- Quadratic Programming
  - Very efficient computationally with modern constraint optimization engines (handles thousands of constraints and training instances).

CSE 7305c

Solving the Optimization Problem

Find w, b such that:

$$Minimize\ f(x):\ 1/2\ ||w||^2$$

$$\text{Subject to: } g(x):\ y_i(w.x_i + b) - 1 = 0$$

- **The solution involves constructing a *dual problem* where a *Lagrange multiplier* α_i is associated with every constraint in the primary problem:**

Find $\alpha_1 \ldots \alpha_N$ such that

$$L_D = -\frac{1}{2}\sum_{i=0}^{n} \alpha_i \alpha_j y_i y_j (x_i . x_j) + \sum_{i=0}^{n} \alpha_i \ is\ Maximized$$

$$Subject\ to: \sum_{i=0}^{n} \alpha_i y_i = 0, \alpha_i > 0$$

Once we find $\alpha_1 \ldots \alpha_N$

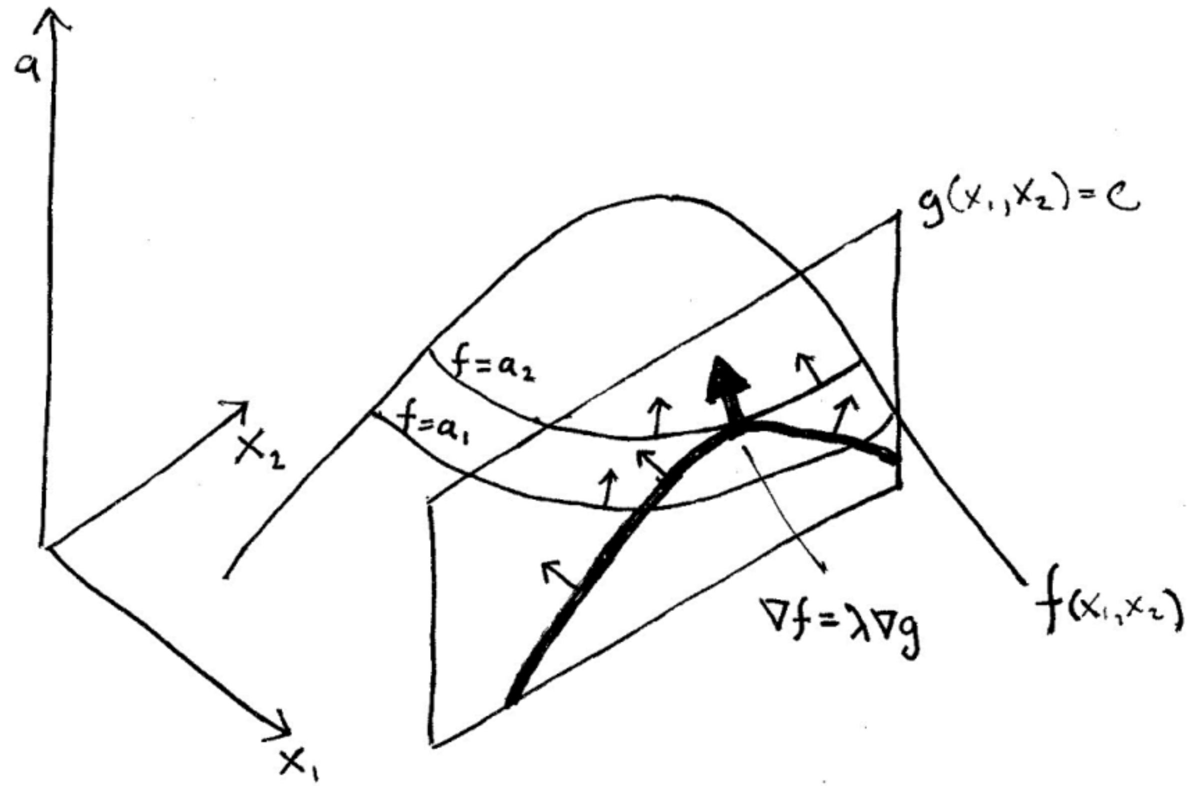
$$Solution\ w = \sum_{i=0}^{n} a_i y_i x_i$$

Derivations

- The next 7 slides are only for reference to understand derivation details !!!

CSE 7305c

Optimization with Lagrange Multipliers



http://tutorial.math.lamar.edu/Classes/CalcIII/LagrangeMultipliers.aspx
http://www.slimy.com/~steuard/teaching/tutorials/Lagrange.html

CSE 7305c

Lagrange Multipliers

- Minimize $U=x^2 + y^2$
- Subject to $g= x^2+y^2+2x-2y+1 = 0$

- Lagrange equation

- $L=U-\lambda g = 0$

CSE 7305c

Method of Lagrange multipliers

$U = x^2 + y^2$
$g = x^2 + y^2 + 2x - 2y + 1 = 0$

- $\frac{\partial L}{\partial x} = 2x - \lambda(2x + 2) = 0;\ x - \lambda(x+1) = 0$
- $\frac{\partial L}{\partial y} = 2y - \lambda(2y - 2) = 0;\ y - \lambda(y - 1) = 0$
- $\frac{\partial L}{\partial \lambda} = -1(x^2 + y^2 + 2x - 2y + 1) = 0$

- From 1 and 2, we get y=-x and substituting in 3, we get x and y values.

CSE 7305c

Practical Use: SVM Training

Note: $\frac{1}{2}\,\mathbf{w}^T\mathbf{w}$ is quadratic

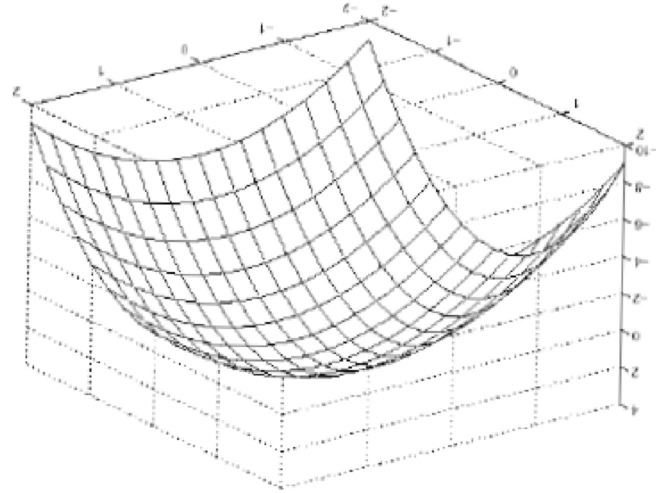
Find $\mathbf{w}$ and b such that

f: $\Phi(\mathbf{w}) = \frac{1}{2}\,\mathbf{w}^T\mathbf{w}$ is minimized and for all $\{(\mathbf{x}_i, y_i)\}$

g: $y_i\,(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

f is a quadratic function, and a paraboloid, and is known to have unique solution.

f, g can be converted into Lagrangian form



Example: paraboloid $2+x^2+2y^2$ s.t. $x+y=1$

Reference: svm-notes-long-08

CSE 7305c

Practical Use: SVM Training

$$Minimize\ f(x):\ 1/2\,||w||^2$$
$$\text{Subject to: } g(x):\ y_i(w.x_i+b)-1=0$$

Convert to Lagrange form:

$$\min L = 1/2\,||w||^2 - \sum_{i=0}^{n} \alpha_i[y_i(w.x_i+b)-1]$$

$$\min L = 1/2\,||w||^2 - \sum_{i=0}^{n} \alpha_i y_i(w.x_i+b) + \sum_{i=0}^{n} \alpha_i$$

Derivatives at Minima =0

$$\frac{\partial L}{\partial w}:\ w - \sum_{i=0}^{n} \alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b}: \sum_{i=0}^{n} \alpha_i y_i = 0$$

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

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

CSE 7305c

Practical Use: SVM Training

$$Minimize\ f(x):\ 1/2\,||w||^2$$
$$\text{Subject to: } g(x):\ y_i(w.x_i + b) - 1 = 0$$

Convert to Lagrange form:

$$\min L = 1/2\,||w||^2 - \sum_{i=0}^{n} \alpha_i[y_i(w.x_i + b) - 1]$$

$$\min L = 1/2\,||w||^2 - \sum_{i=0}^{n} \alpha_i y_i(w.x_i + b) + \sum_{i=0}^{n} \alpha_i$$

$$w = \sum_{i=0}^{n} \alpha_i y_i x_i \qquad \sum_{i=0}^{n} \alpha_i y_i = 0$$

$$\max L_D = -\frac{1}{2}\sum_{i=0}^{n} \alpha_i \alpha_j y_i y_j (x_i . x_j) + \sum_{i=0}^{n} \alpha_i$$

$$Such\ that: \sum_{i=0}^{n} \alpha_i y_i = 0, \alpha_i > 0$$

Substitute these values to get the dual solution:

CSE 7305c

Practical Use: SVM Training

$$\text{Minimize } f(x)\text{: } 1/2\,||w||^2$$
$$\text{Subject to: } g(x)\text{: } y_i(w.x_i + b) - 1 = 0$$

Convert to Lagrange form:

$$\min L = 1/2\,||w||^2 - \sum_{i=0}^{n} a_i y_i (w.x_i + b) + \sum_{i=0}^{n} a_i$$

$$\text{Solution } w = \sum_{i=0}^{n} a_i y_i x_i \qquad \sum_{i=0}^{n} a_i y_i = 0$$

$$\max L_D = -\frac{1}{2}\sum_{i=0}^{n} \alpha_i \alpha_j y_i y_j (x_i.x_j) + \sum_{i=0}^{n} \alpha_i$$

$$\text{Such that: } \sum_{i=0}^{n} \alpha_i y_i = 0, \alpha_i > 0$$

Why is the dual solution used?

Ans 1: Easier to compute in case of Kernel methods.

Ans 2: Kind of related to Answer 1. If you take partial derivative of $\max L_D WRT \alpha$, the solution will be:

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

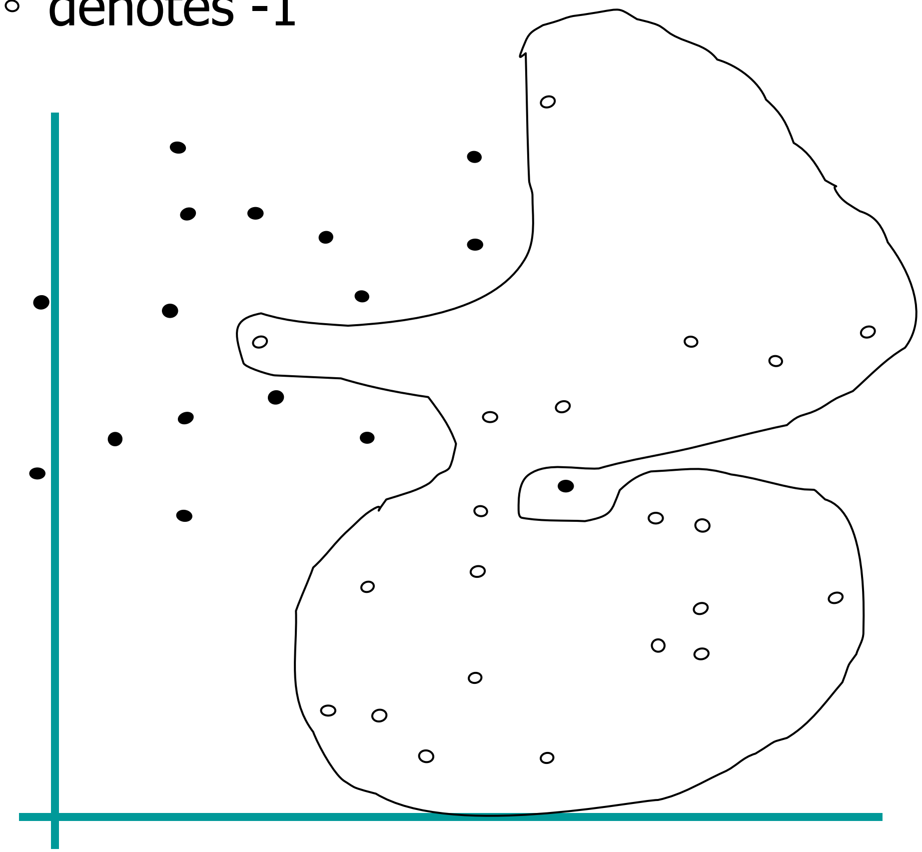
The way to solve is to find different values of α and see which ones will give maximum. The original solution is much worse

Reference: svm15.pdf

CSE 7305c

Dataset with noise

- • denotes +1
- ◦ denotes -1



- **Hard Margin:** **So far we require all data points be classified correctly**

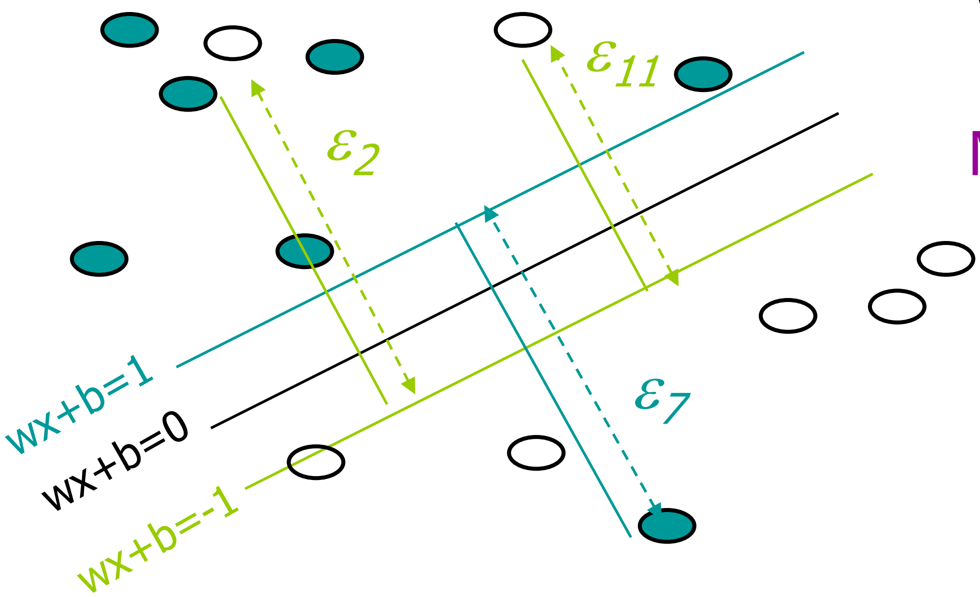
  - No training error
- **What if the training set is noisy?**

  - Solution 1: **use very powerful kernels**

OVERFITTING!

Soft Margin Classification

Slack variables **can be added to allow misclassification of difficult or noisy examples.**



What should our quadratic optimization criterion be?

Minimize

$$\frac{1}{2}\mathbf{w}.\mathbf{w} + C\sum_{k=1}^{R}\varepsilon_k$$

CSE 7305c

Hard vs. Soft Margin SVM

- Soft-Margin also <u>always</u> has a solution
- Soft-Margin is more robust to outliers
  - Smoother surfaces (in the non-linear case)
- Hard-Margin does not require to guess the cost parameter (requires no parameters at all)

CSE 7305c

Hard vs Soft Margin SVM

- **The old formulation:**

Find $\mathbf{w}$ and b such that
$\mathbf{\Phi}(\mathbf{w}) = ½\ \mathbf{w}^T\mathbf{w}$ is minimized and for all $\{(\mathbf{x_i}, y_i)\}$
$y_i\ (\mathbf{w}^T\mathbf{x_i} + \mathrm{b}) \geq 1$

- **The new formulation incorporating slack variables:**

Find $\mathbf{w}$ and b such that
$\mathbf{\Phi}(\mathbf{w}) = ½\ \mathbf{w}^T\mathbf{w} + C\Sigma\xi_i$ is minimized and for all $\{(\mathbf{x_i}, y_i)\}$
$y_i\ (\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for all i

- **Parameter *C* can be viewed as a way to control overfitting.**

Linear SVMs: Overview

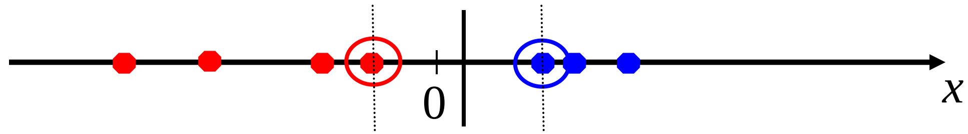
- **The classifier is a *separating hyperplane.***
- **Most "important" training points are support vectors; they define the hyperplane.**
- **Quadratic optimization algorithms can identify which training points x_i are support vectors with non-zero Lagrangian multipliers α_i.**
- **Both in the dual formulation of the problem and in the solution, training points appear only inside dot products:**

Find $\alpha_1 \dots \alpha_N$ such that
$Q(\alpha) = \Sigma\alpha_i - \frac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j \boxed{\mathbf{x}_i^T\mathbf{x}_j}$ is maximized and
(1) $\Sigma\alpha_i y_i = 0$
(2) $0 \leq \alpha_i \leq C$ for all α_i

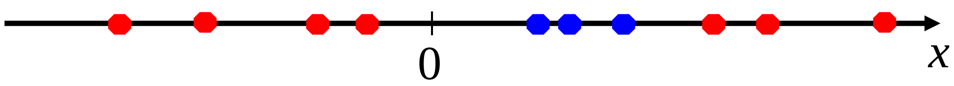
$$f(\mathbf{x}) = \Sigma\alpha_i y_i \boxed{\mathbf{x}_i^T\mathbf{x}} + b$$

Non-linear SVMs

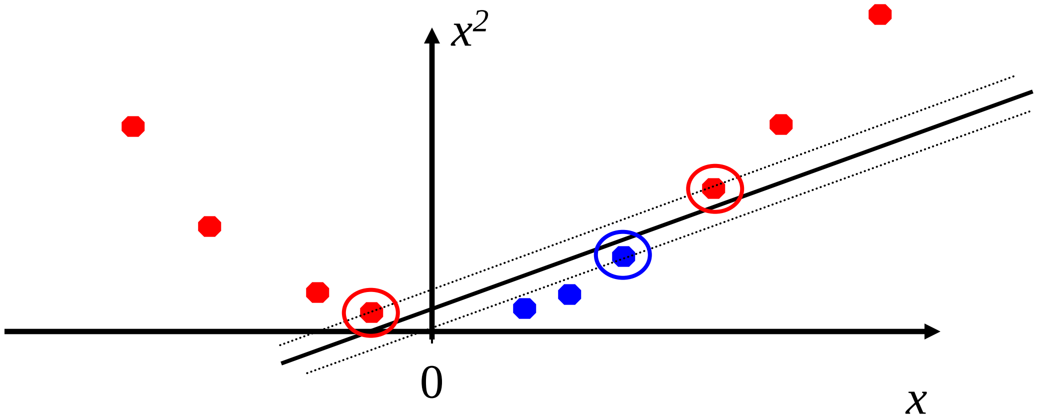
- Datasets that are linearly separable with some noise work out great:



- But what are we going to do if the dataset is just too hard?

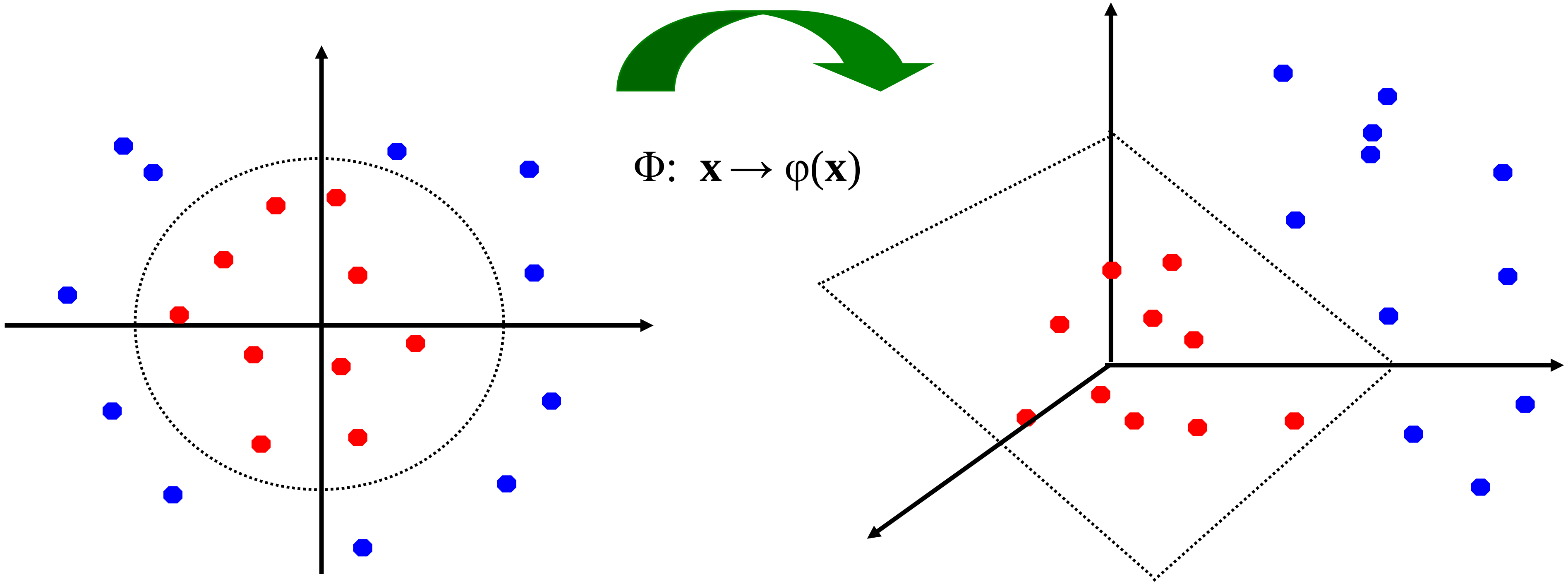


- How about… mapping data to a higher-dimensional space:



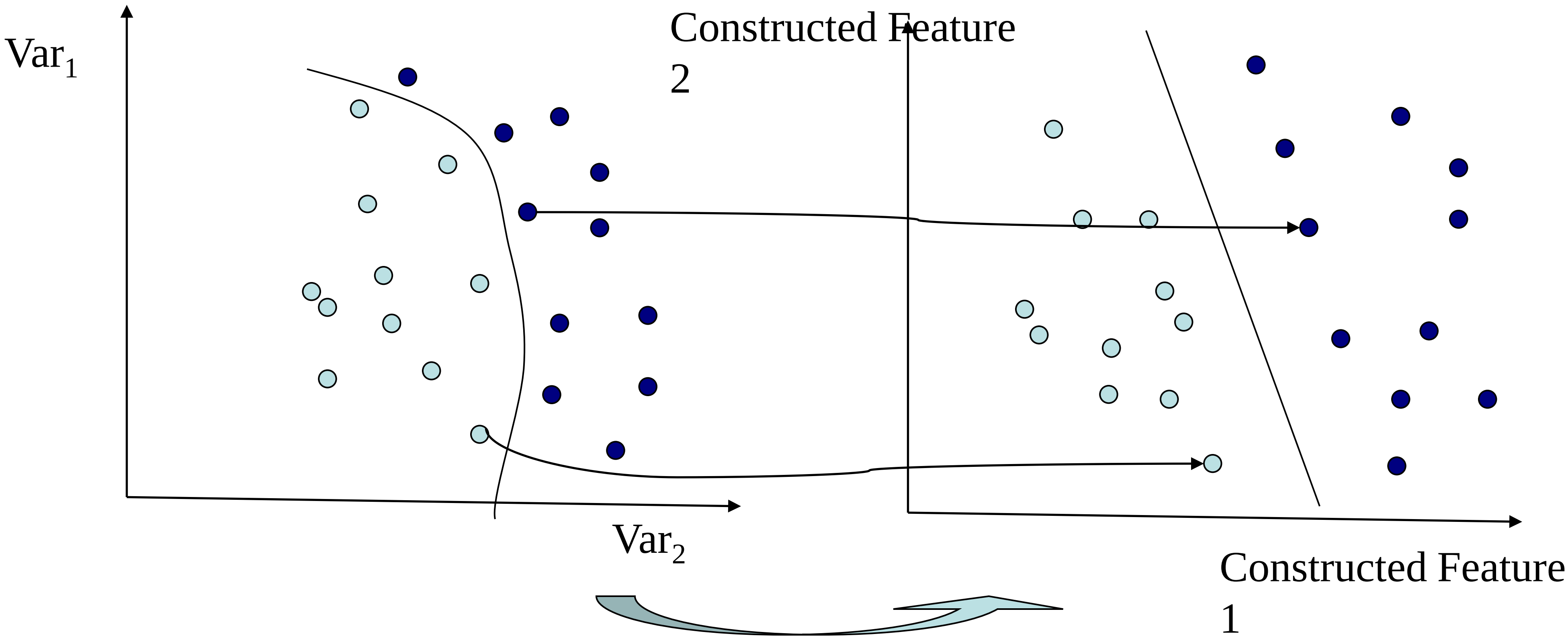
Non-linear SVMs: Feature spaces

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



Φ: $\mathbf{x} \rightarrow \varphi(\mathbf{x})$

Linear Classifiers in High-Dimensional Spaces



Find function Φ(x) to map to a different space

CSE 7305c

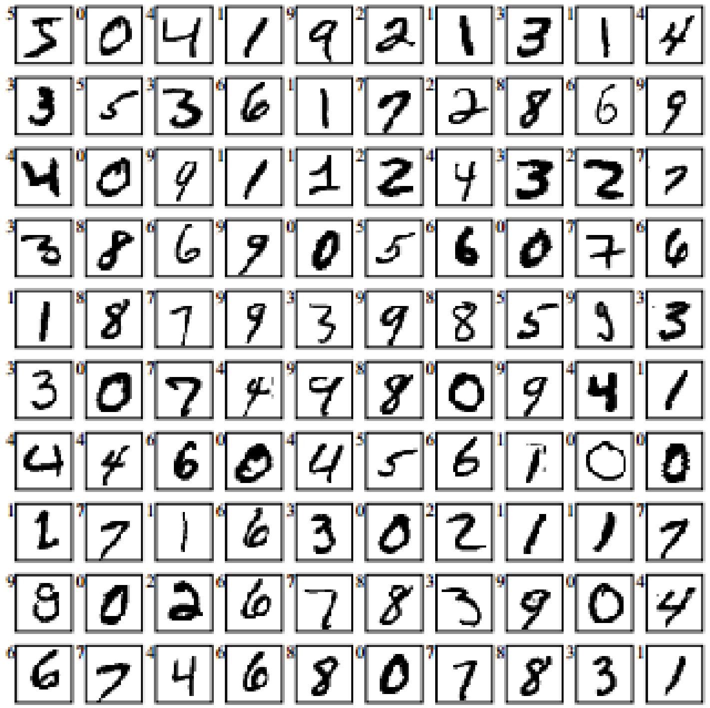
Some problems need non-linear SVMs.

For example MNIST hand-writing recognition.

60,000 training examples, 10000 test examples, 28x28.

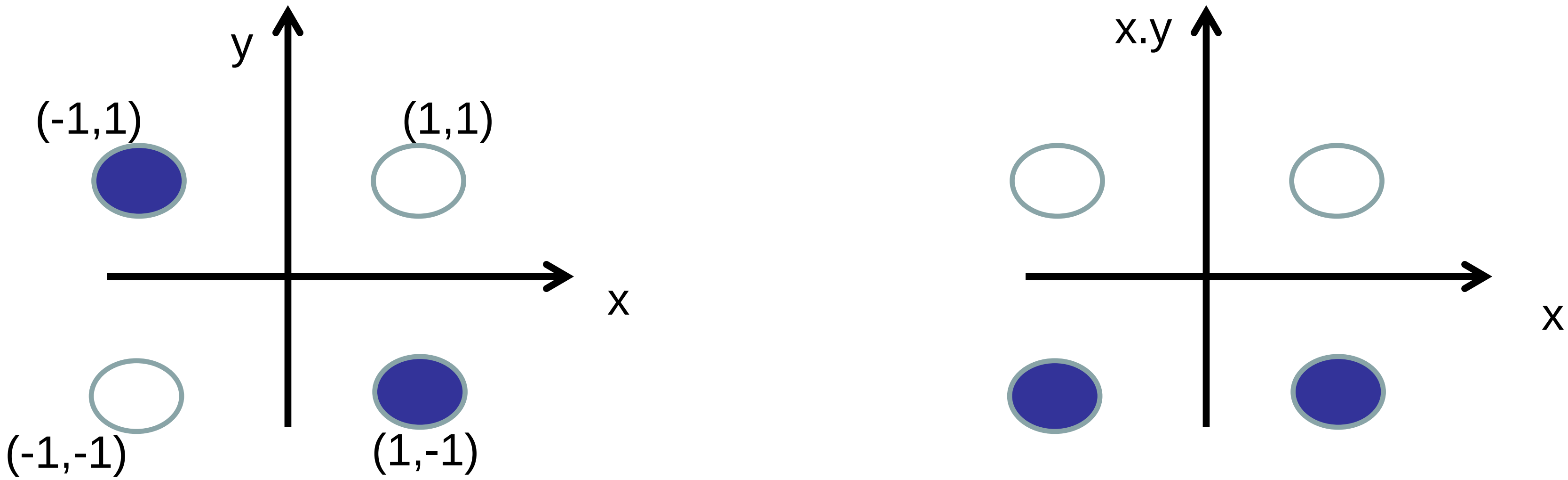
Linear SVM has around 8.5% test error.

Polynomial SVM has around 1% test error.



CSE 7305c

Linear Separation of XOR



- XOR is not linearly separable in x,y space
- Linearly separable in x.y space
  - The kernel here will be K(x,y) = x.y

CSE 7305c

Reference: http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture18.pdf, Duda and Hart, chapter 3.

Nonlinear SVM - Overview

- SVM locates a separating hyperplane in the feature space and classify points in that space
- It does not need to represent the space explicitly, simply by defining a kernel function
- The kernel function plays the role of the dot product in the feature space.

CSE 7305c

A Little More on Kernel Functions

Find **w** and b such that

$\mathbf{\Phi}(\mathbf{w}) = \frac{1}{2}\, \mathbf{w}^T\mathbf{w} + C\Sigma\xi_i$

is minimized and for all $\{(\mathbf{x_i}, y_i)\}$

$y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i$

and $\xi_i \geq 0$ for all i

Find $\alpha_1 \ldots \alpha_N$ such that

$\mathbf{Q}(\alpha) = \Sigma\alpha_i - \frac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j \boxed{\mathbf{x_i}^T\mathbf{x_j}}$ **is maximized and**

(1) $\Sigma\alpha_i y_i = 0$

(2) $0 \leq \alpha_i \leq C$ **for all** α_i

$f(\mathbf{x}) = \Sigma\alpha_i y_i \boxed{\mathbf{x_i}^T\mathbf{x}} + b$

- If we map the input vectors into a very high-dimensional feature space, surely the task of finding the maximum-margin separator becomes computationally intractable?
  - The mathematics is all linear, which is good.
  - But the vectors have a huge number of components. Taking the scalar product of two vectors is very expensive.
- We can keep things tractable by using "the kernel trick"

CSE 7305c

The Kernel Trick

- All computations needed for the maximum-margin separator are in terms of scalar products between pairs of data points (in the high-dimensional feature space).

- Scalar products are the only part of the computation that depends on the dimensionality of the high-dimensional space.

  – So if we had a fast way to do the scalar products we would not have to pay a price for solving the learning problem in the high-D space.

- The kernel trick is a way of doing scalar products a whole lot faster than is usually possible.

  – It relies on choosing a way of mapping to the high-dimensional feature space that allows fast scalar products.

CSE 7305c

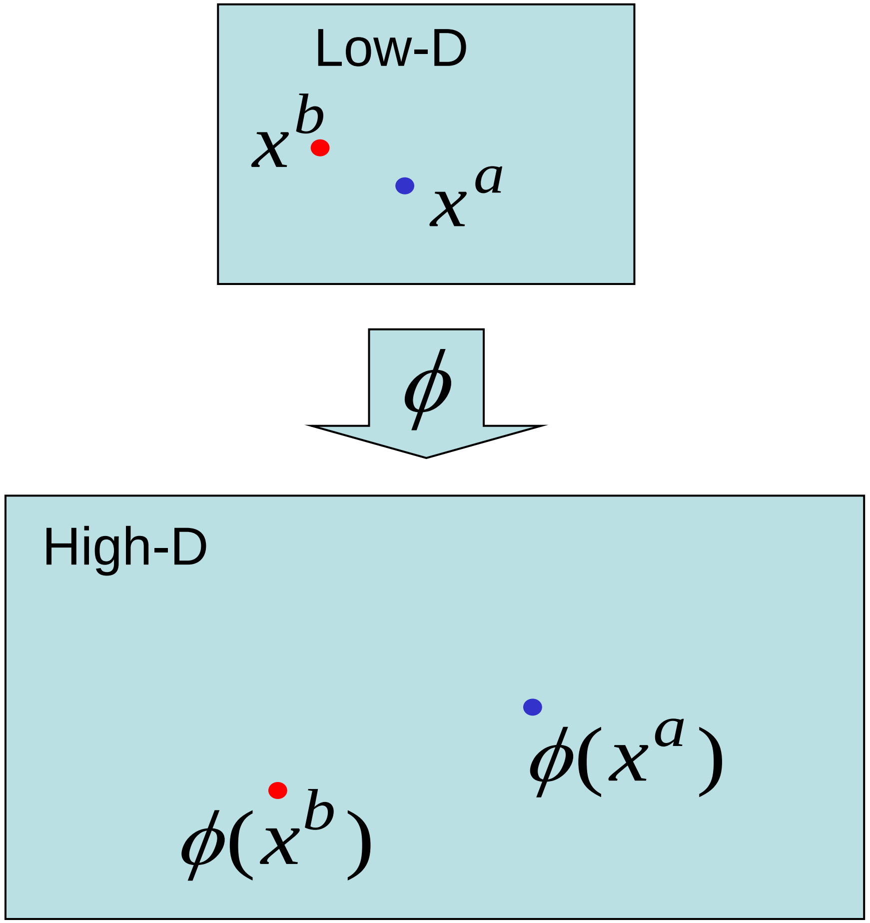
The kernel trick

- For many mappings from a low-D space to a high-D space, there is a simple operation on two vectors in the low-D space that can be used to compute the scalar product of their two images in the high-D space.

$$K(x^a, x^b) = \phi(x^a).\phi(x^b)$$

Letting the kernel do the work

doing the scalar product in the obvious way



CSE 7305c

Examples of Kernel Functions

- Linear: $K(\mathbf{x_i},\mathbf{x_j}) = \mathbf{x_i}^T\mathbf{x_j}$

- Polynomial of power p: $K(\mathbf{x_i},\mathbf{x_j}) = (1 + \mathbf{x_i}^T\mathbf{x_j})^p$

- Gaussian (radial-basis function network):

$$K(\mathbf{x_i},\mathbf{x_j}) = \exp\left(-\frac{\left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2}{2\sigma^2}\right)$$

- Sigmoid: $K(\mathbf{x_i},\mathbf{x_j}) = \tanh(\beta_0\mathbf{x_i}^T\mathbf{x_j} + \beta_1)$

Many Kernel Functions Possible

Linear Kernel
Polynomial Kernel
Gaussian Kernel
Exponential Kernel
Laplacian Kernel
ANOVA Kernel
Hyperbolic Tangent (Sigmoid) Kernel
Rational Quadratic Kernel
Multiquadric Kernel
Inverse Multiquadric Kernel
Circular Kernel
Spherical Kernel

Wave Kernel
Power Kernel
Log Kernel
Spline Kernel
B-Spline Kernel
Bessel Kernel
Cauchy Kernel
Chi-Square Kernel
Histogram Intersection Kernel
Generalized Histogram Intersection Kernel
Generalized T-Student Kernel
Bayesian Kernel
Wavelet Kernel

http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html

CSE 7305c

Properties of SVM

- Flexibility in choosing a similarity function
- Efficiency of solution when dealing with large data sets

  - only support vectors are used to specify the separating hyperplane
- Ability to handle large feature spaces

  - complexity does not depend on the dimensionality of the feature space
- Overfitting can be controlled by soft margin approach
- Nice math property: a simple convex optimization problem which is guaranteed to converge to a single global solution

CSE 7305c

Multi class SVM

- We studied 2 class SVMs
- Multi class is possible using Kesler’s construction
  - Input data modified as below:

$$n_{y2} = \begin{bmatrix} y \\ -y \\ 0 \\ \ldots \\ 0 \end{bmatrix} n_{y3} = \begin{bmatrix} y \\ 0 \\ -y \\ \ldots \\ 0 \end{bmatrix} n_{ym} = \begin{bmatrix} y \\ 0 \\ 0 \\ \ldots \\ -y \end{bmatrix}$$

y is a sample of class 1
If original problem has “m” classes, n samples, d dimensions
Total number of samples will become (m-1).n
Dimensionality will become m.d

  - SVM is trained on this data

CSE 7305c

http://actipret.acin.tuwien.ac.at/Sites/FrancHlavacICPR2002.pdf

Multi class SVM

- Kesler’s construction not practical:
  - Sparsely populated classes
  - Similar/Confusing classes
- Train several 2 class SVMs, perform voting
- Train 1 vs all other SVMs, select best result
- Decision tree SVM !!!

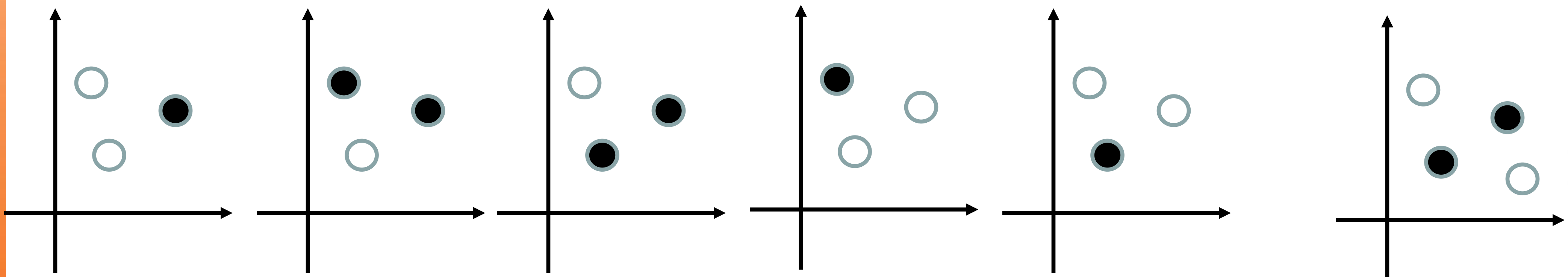
http://www.lib.kobe-u.ac.jp/repository/90000228.pdf

CSE 7305c

VC Dimension

How many dimension?

- Vapnic Chervonenkis Dimension
  - Explanation in plain English: Given a hyperplane in N dimensions, how many points can it separate? Ans: N+1



A line can be drawn to separate the five point sets on left, but not for points on right

http://www.autonlab.org/tutorials/vcdim08.pdf

https://alliance.seas.upenn.edu/~cis520/wiki/index.php?n=Recitations.VCDim
http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture18.pdf

CSE 7305c

VC Dimension

- Given a classifier whose error rate on R training samples is known, and whose VC dimension is h, it is shown that error rate on an unknown test data:

$$Test\ error, e \leq Train\ error + \sqrt{\frac{h\left(\log\left(\frac{2R}{h}\right)+1\right)-\log\left(\frac{\eta}{4}\right)}{R}}$$

Probability of error rate being e: 1- η

A linear classifier operating in h-1 dimensions has VC dimension of h

http://www.liaolin.com/Courses/vc-dimension.pdf http://www.autonlab.org/tutorials/vcdim08.pdf

CSE 7305c

Resources

- An excellent tutorial on VC-dimension and Support Vector Machines:
  C.J.C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):955-974, 1998.

- The VC/SRM/SVM Bible:
  Statistical Learning Theory by Vladimir Vapnik, Wiley-Interscience; 1998

http://www.kernel-machines.org/software

- Quadratic programming basics:
  - http://www.akiti.ca/QuadProgEx0Constr.html
  - http://www.solver.com/linear-quadratic-programming

- Non-convex SVM:
  - http://web.mit.edu/seyda/www/Papers/TPAMI11_Nonconvex.pdf

CSE 7305c

International School of Engineering

2-56/2/19, Khanamet, Madhapur, Hyderabad - 500 081

For Individuals: +91-9177585755 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

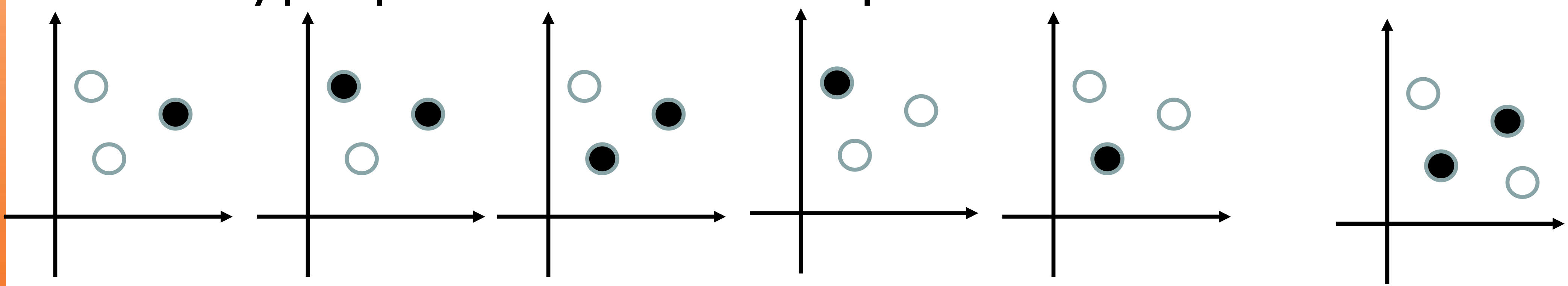
SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.

VC Dimension

- Vapnic Chervonenkis Dimension
  - Explanation in plain English: Given a data set of N points in two classes, what is the dimension which will contain at least one hyperplane that can separate the two classes?



A line can be drawn to separate the five point sets on left, but not for points on right

http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture18.pdf