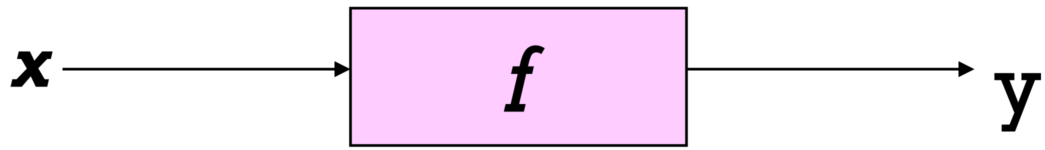
LINEAR CLASSIFIER

FIND A FUNCTION TO CLASSIFY HIGH VALUE CUSTOMERS



High Value Customers

Salary	Nb Orders
150	70
300	100
200	80
120	100

Low Value Customers

Salary	Nb Orders
40	80
220	20
100	20
175	10

Task: Find $\alpha_1, \alpha_2, \alpha_3$:

High value customer $\alpha_1 \bullet salary + \alpha_2 \bullet orders - \alpha_3 > 0$

Low value customer $\alpha_1 \bullet salary + \alpha_2 \bullet orders - \alpha_3 < 0$

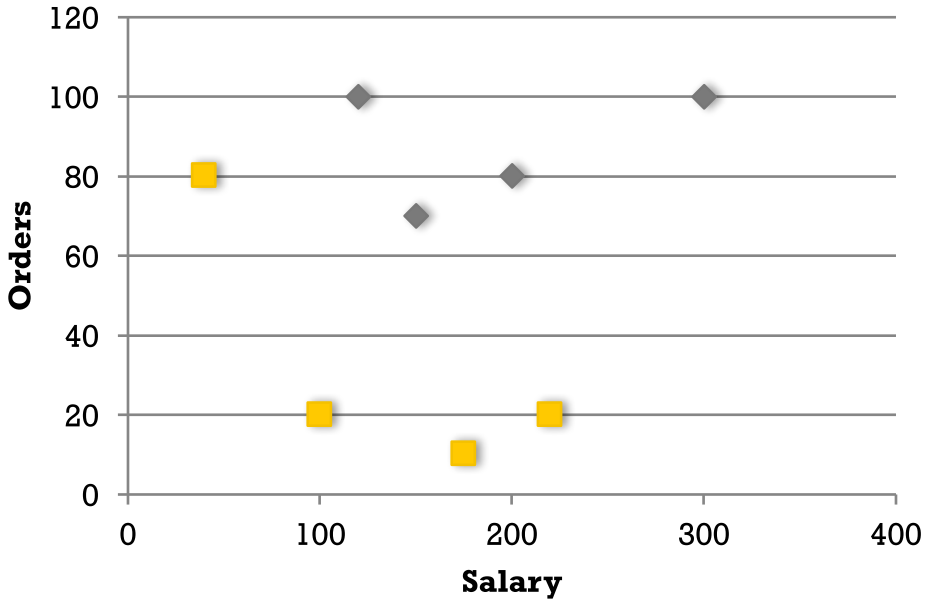
FIND A FUNCTION TO CLASSIFY HIGH VALUE CUSTOMERS

High Value Customers

Salary	Orders
150	70
300	100
200	80
120	100

Low Value Customers

Salary	Orders
40	80
220	20
100	20
175	10



Task: Find $\alpha_1, \alpha_2, \alpha_3$:

High value customer $\alpha_1 \bullet salary + \alpha_2 \bullet orders - \alpha_3 > 0$

Low value customer $\alpha_1 \bullet salary + \alpha_2 \bullet orders - \alpha_3 < 0$

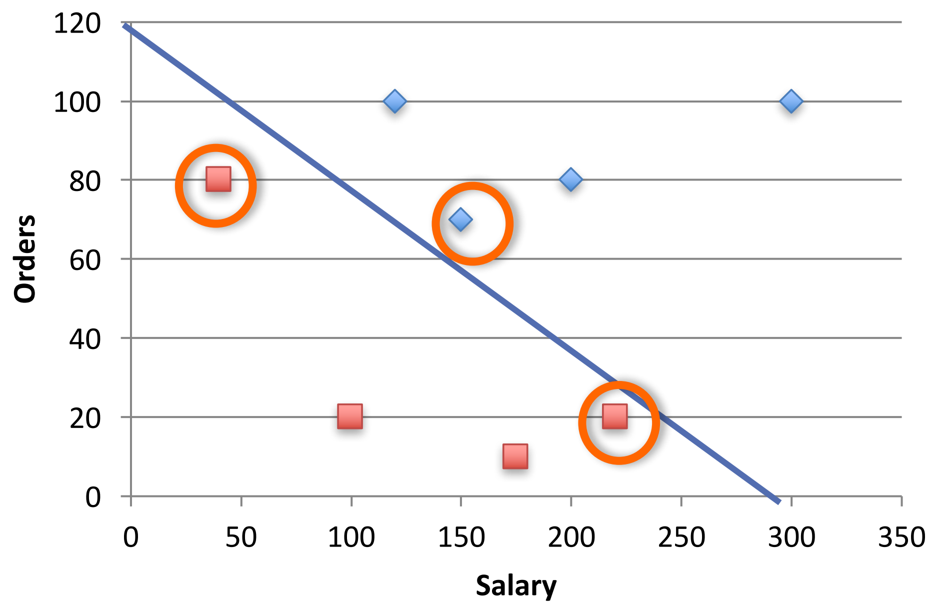
FIND A FUNCTION TO CLASSIFY HIGH VALUE CUSTOMERS

High Value Customers

Salary	Orders
150	70
300	100
200	80
120	100

Low Value Customers

Salary	Orders
40	80
220	20
100	20
175	10

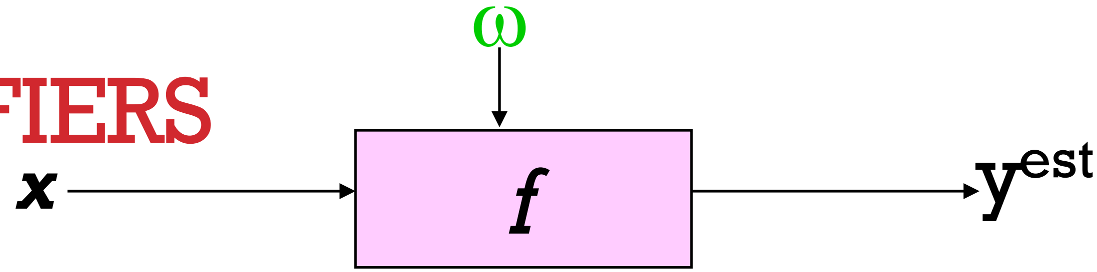


Task: Find $\alpha_1, \alpha_2, \alpha_3$:

High value customer $\alpha_1 \bullet salary + \alpha_2 \bullet orders - \alpha_3 > 0$

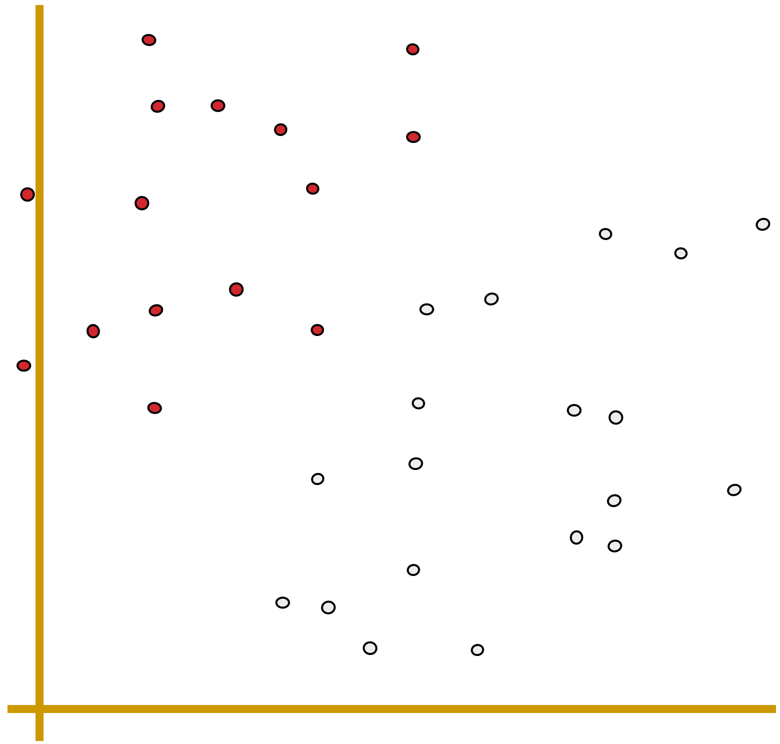
Low value customer $\alpha_1 \bullet salary + \alpha_2 \bullet orders - \alpha_3 < 0$

LINEAR CLASSIFIERS



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

- ● denotes +1
- ○ denotes -1

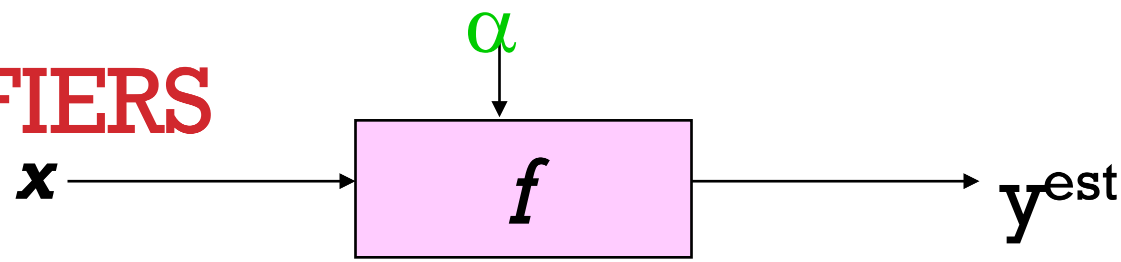


How would you classify this data?

x vector of feature values, w vector of weights. x.w=b is a line (hyperplane)

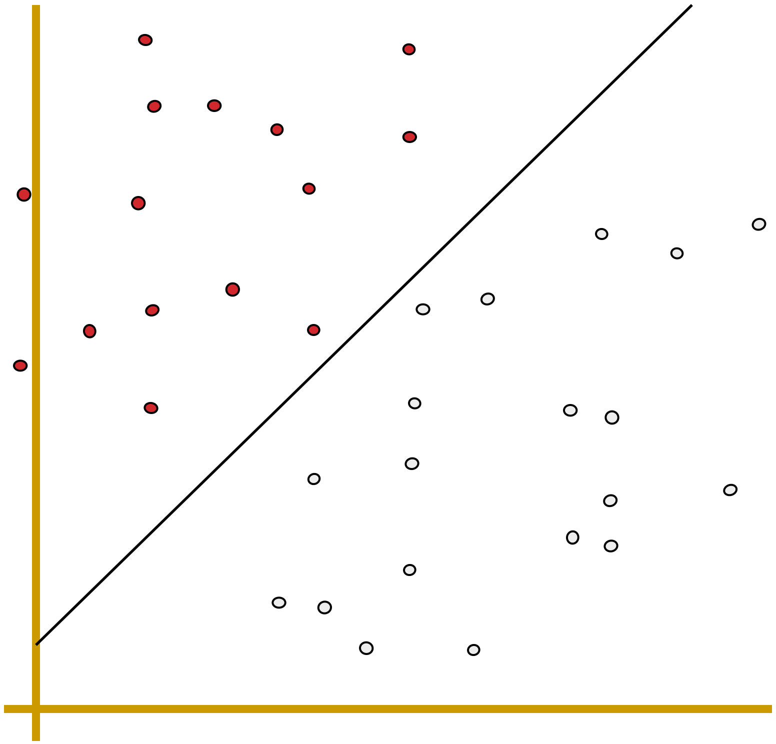
Slides from Andrew W. Moore

LINEAR CLASSIFIERS



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

- denotes +1
- denotes -1

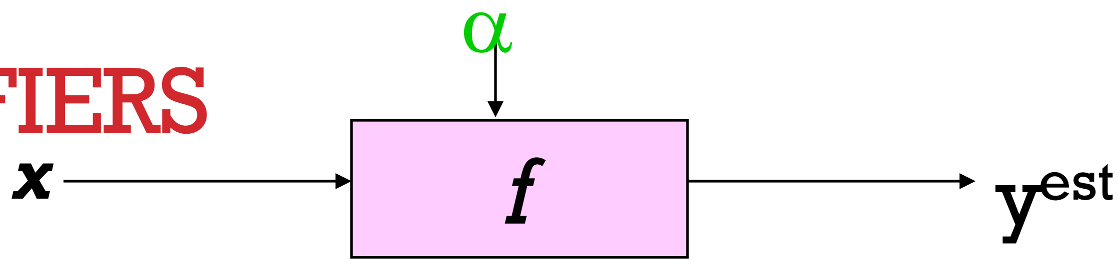


How would you classify this data?

$x.w = b$ or $w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + \ldots = b$ is an hyperplane.

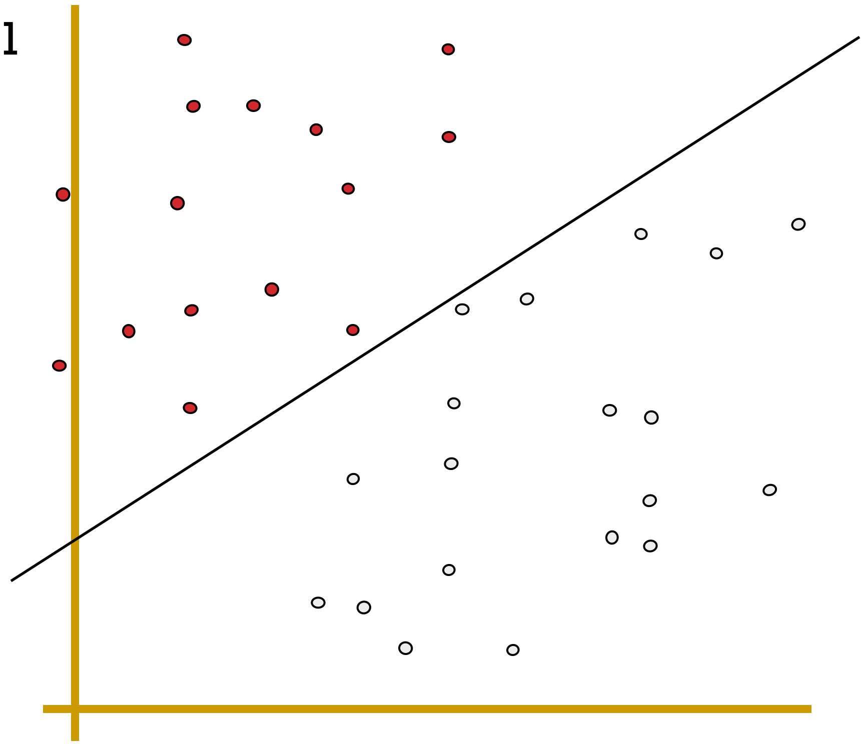
Slides from Andrew W. Moore

LINEAR CLASSIFIERS



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

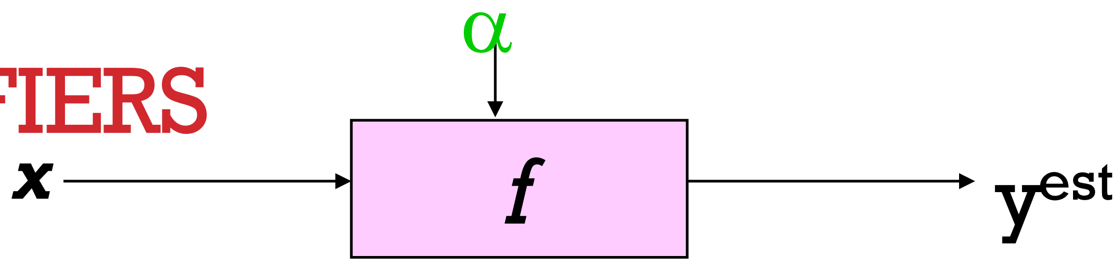
- denotes +1
- denotes -1



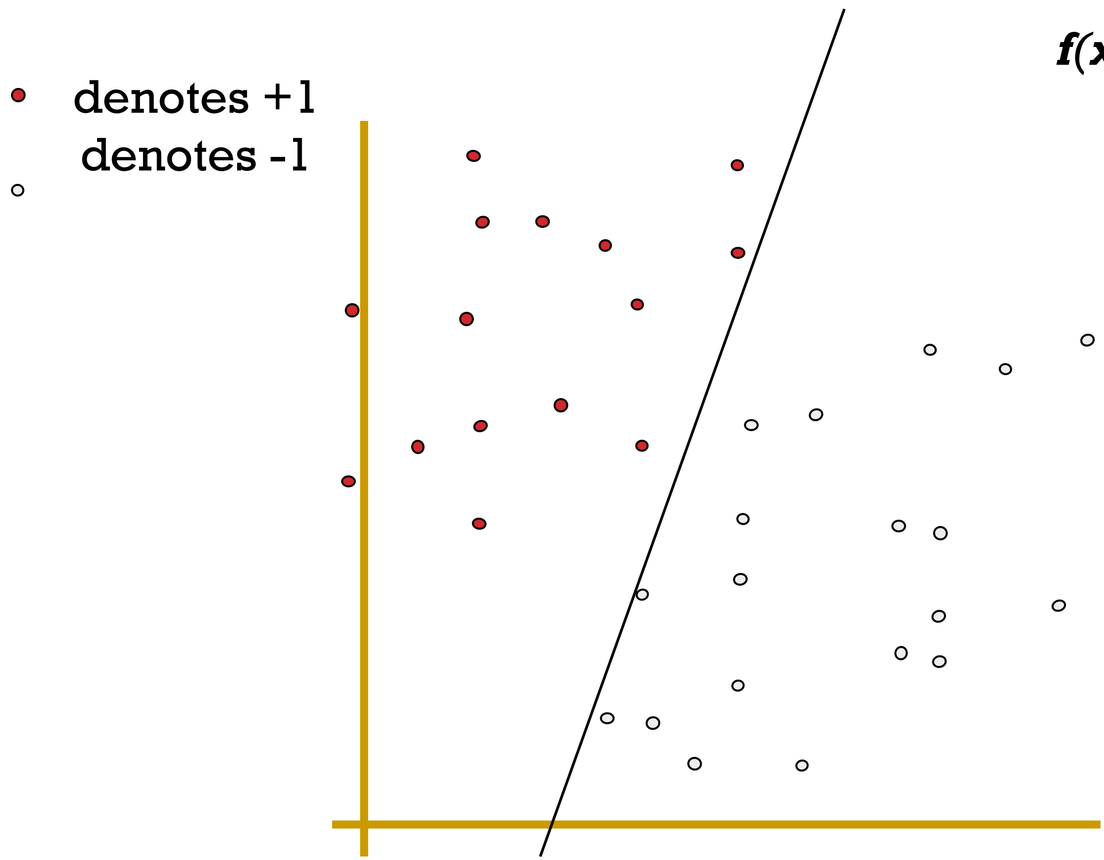
How would you classify this data?

Slides from Andrew W. Moore

LINEAR CLASSIFIERS

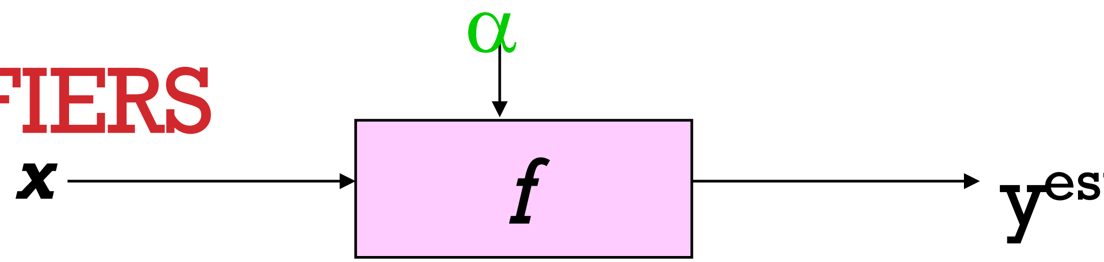


$f(\boldsymbol{x}, \boldsymbol{w}, b) = sign(\boldsymbol{w}. \boldsymbol{x} - b)$



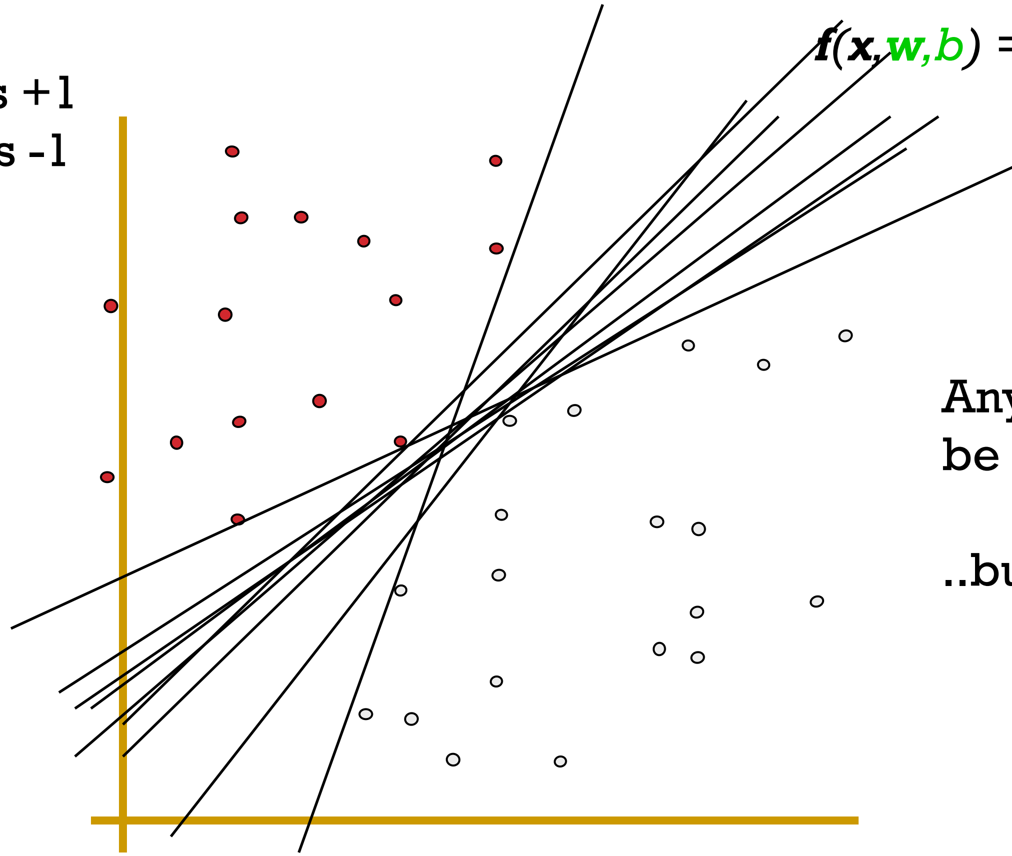
How would you classify this data?

LINEAR CLASSIFIERS



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

- denotes +1
- denotes -1

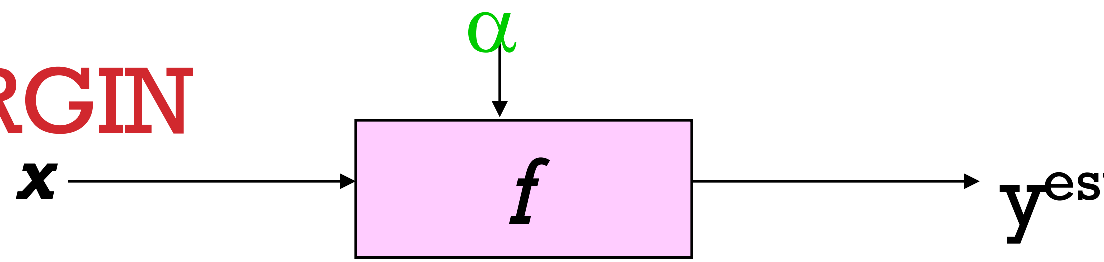


Any of these would be fine..

..but which is best?

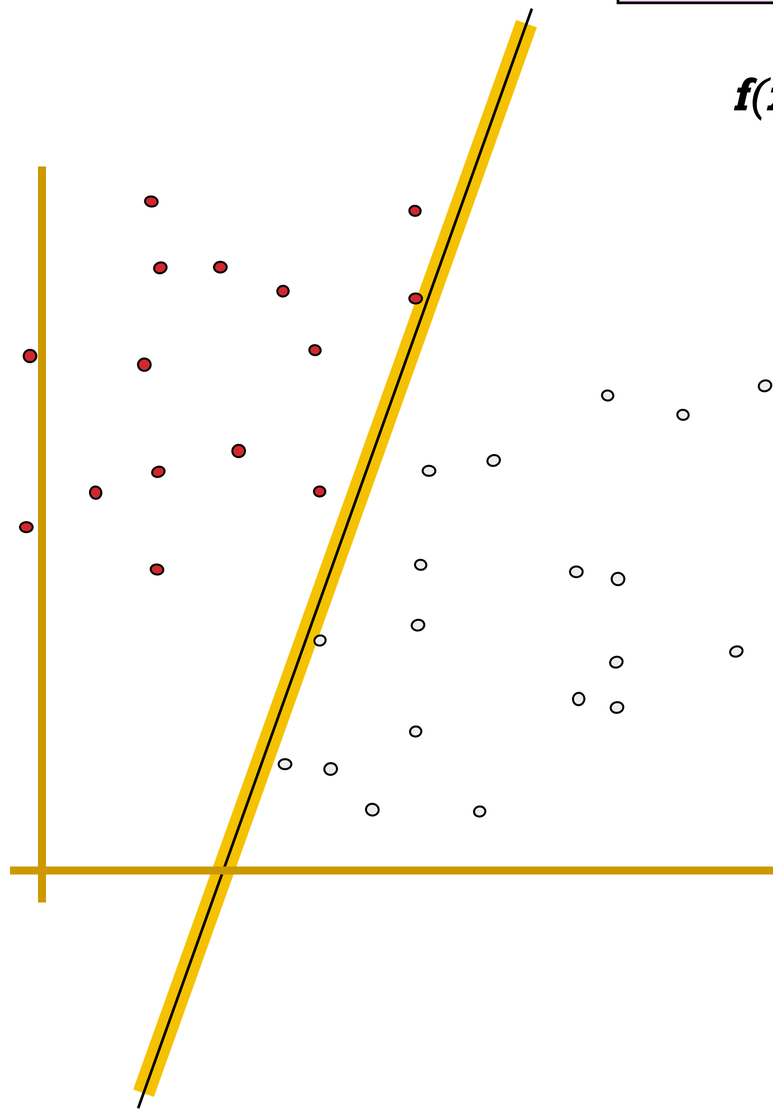
Slides from Andrew W. Moore

CLASSIFIER MARGIN



$f(\boldsymbol{x},\boldsymbol{w},b) = sign(\boldsymbol{w}.\,\boldsymbol{x} - b)$

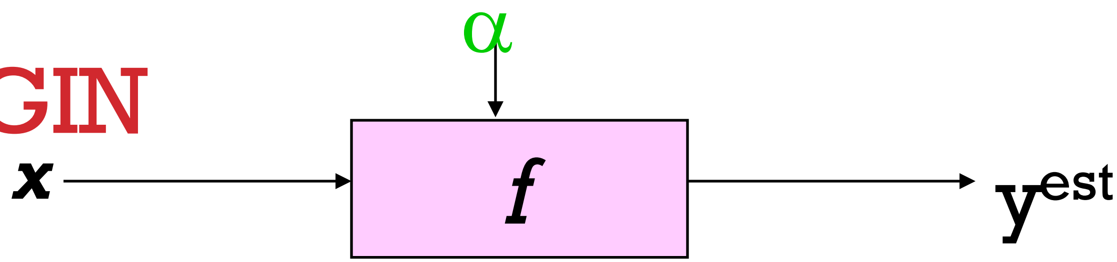
- denotes +1
- denotes -1



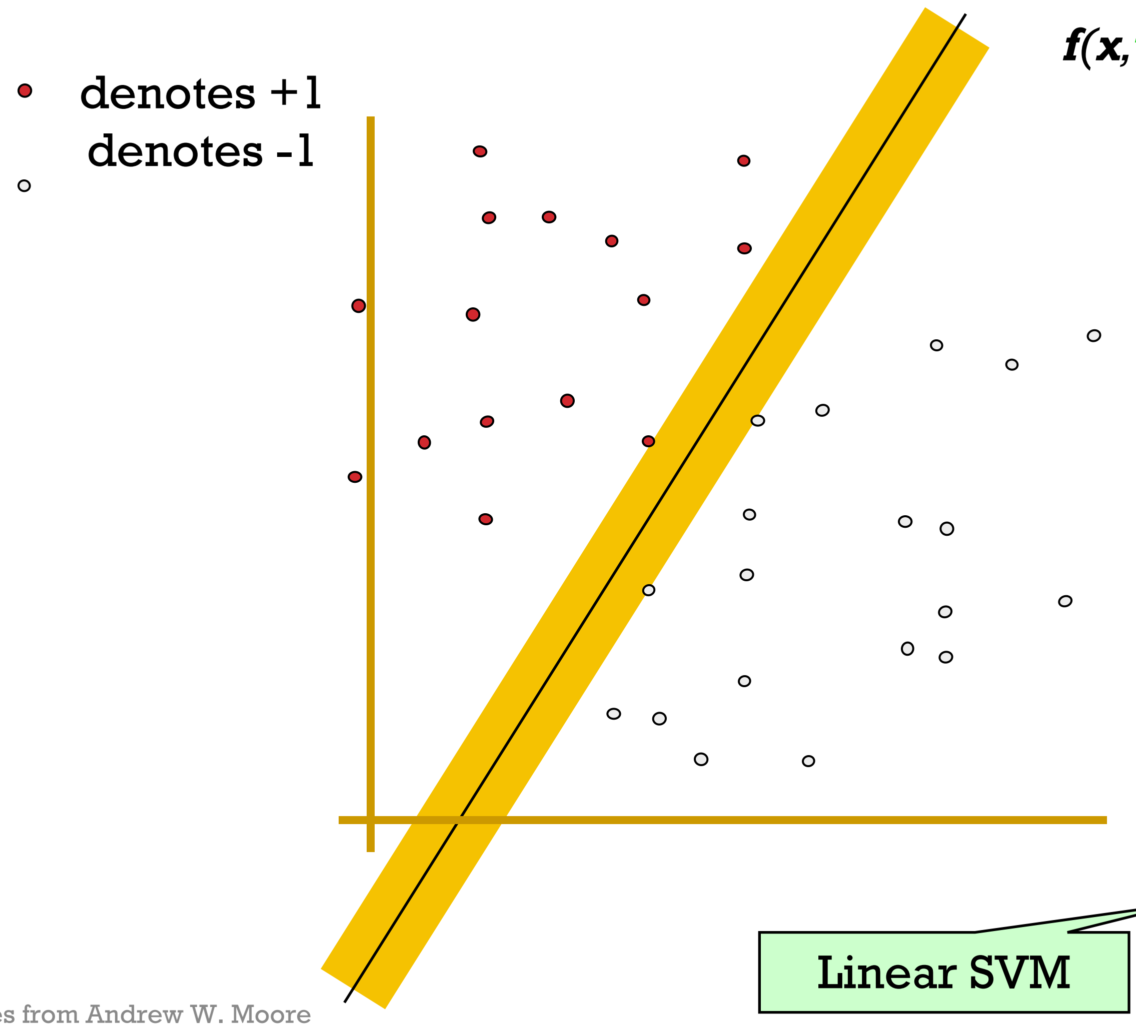
Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

Slides from Andrew W. Moore

MAXIMUM MARGIN



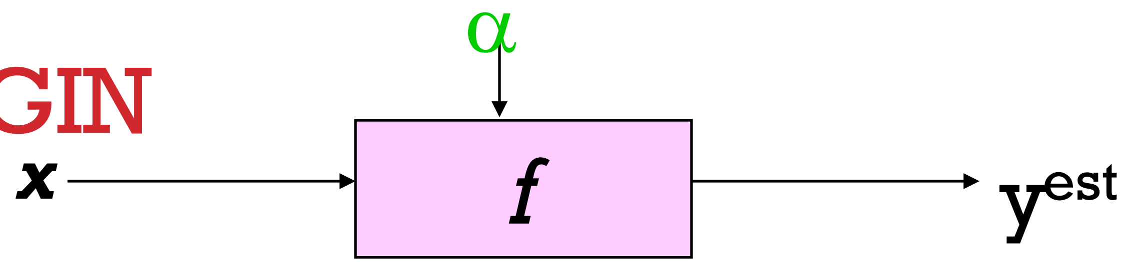
$f(\boldsymbol{x},\boldsymbol{w},b) = sign(\boldsymbol{w}.\ \boldsymbol{x} - b)$



The **maximum margin linear classifier** is the linear classifier with the, um, maximum margin. This is the simplest kind of SVM (Called an LSVM)

Slides from Andrew W. Moore

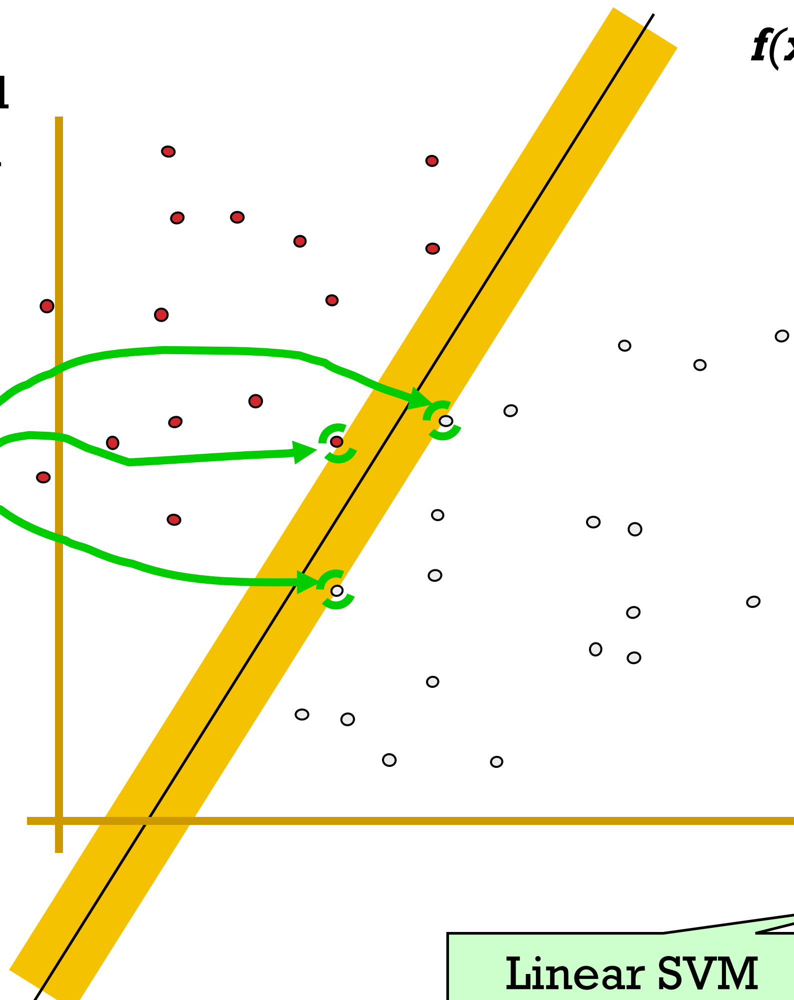
MAXIMUM MARGIN



$f(x, w, b) = sign(w. x - b)$

● denotes +1

○ denotes -1



Support Vectors are those datapoints that the margin pushes up against

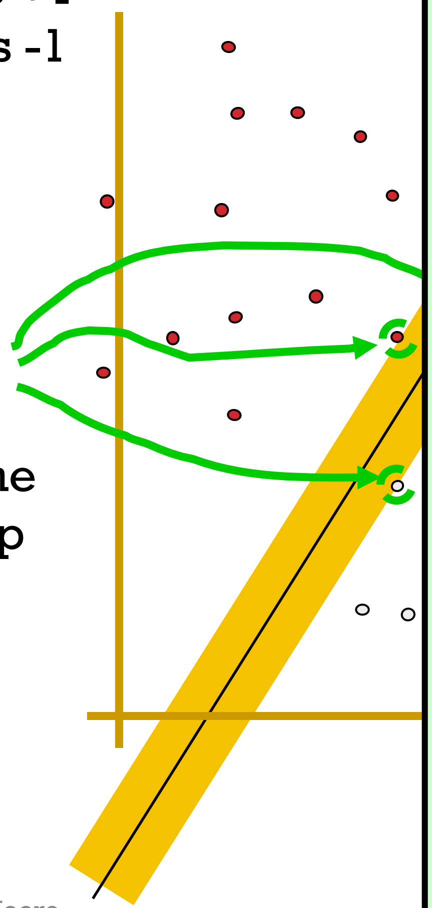
The maximum margin linear classifier is the linear classifier with the, um, maximum margin. This is the simplest kind of SVM (Called an LSVM)

Linear SVM

Slides from Andrew W. Moore

WHY MAXIMUM MARGIN?

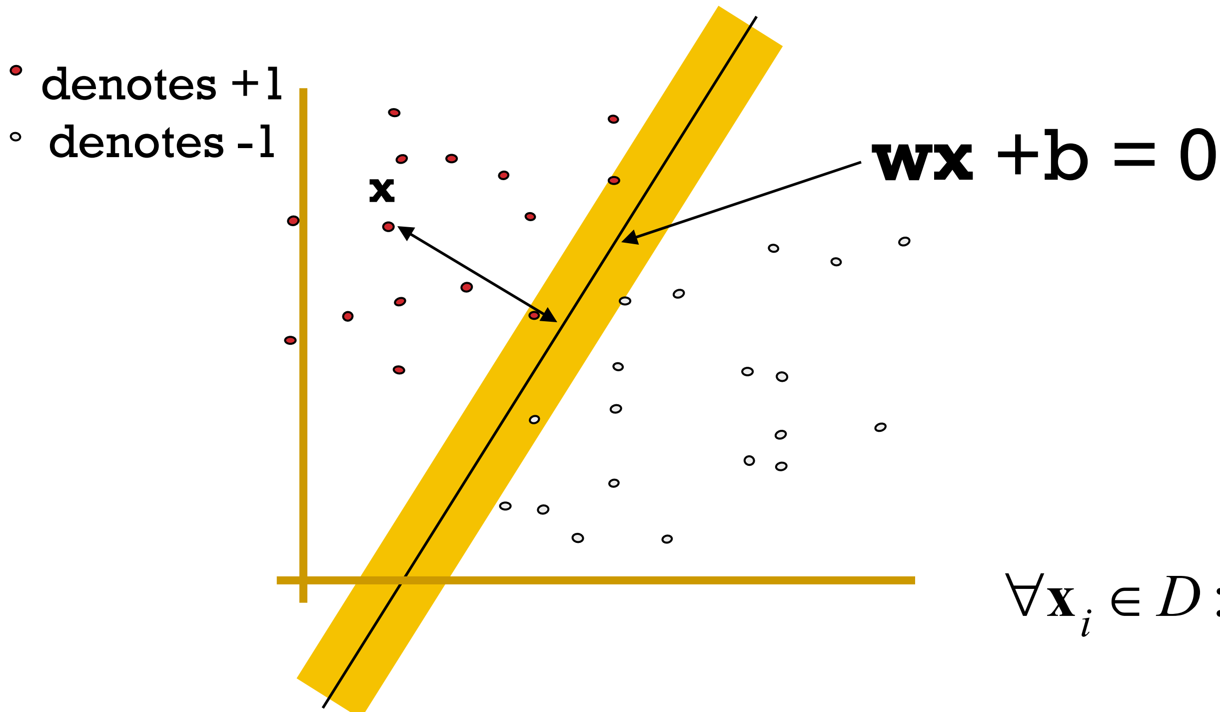
• denotes +1
° denotes -1



Support Vectors are those datapoints that the margin pushes up against

1. Intuitively this feels safest.
2. 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.
3. Leave-one-out-cross-validation (LOOCV) is easy since the model is immune to removal of any non-support-vector datapoints.
4. There's some theory (using VC dimension) that is related to (but not the same as) the proposition that this is a good thing.
5. Empirically it works very very well.

ESTIMATE THE MARGIN



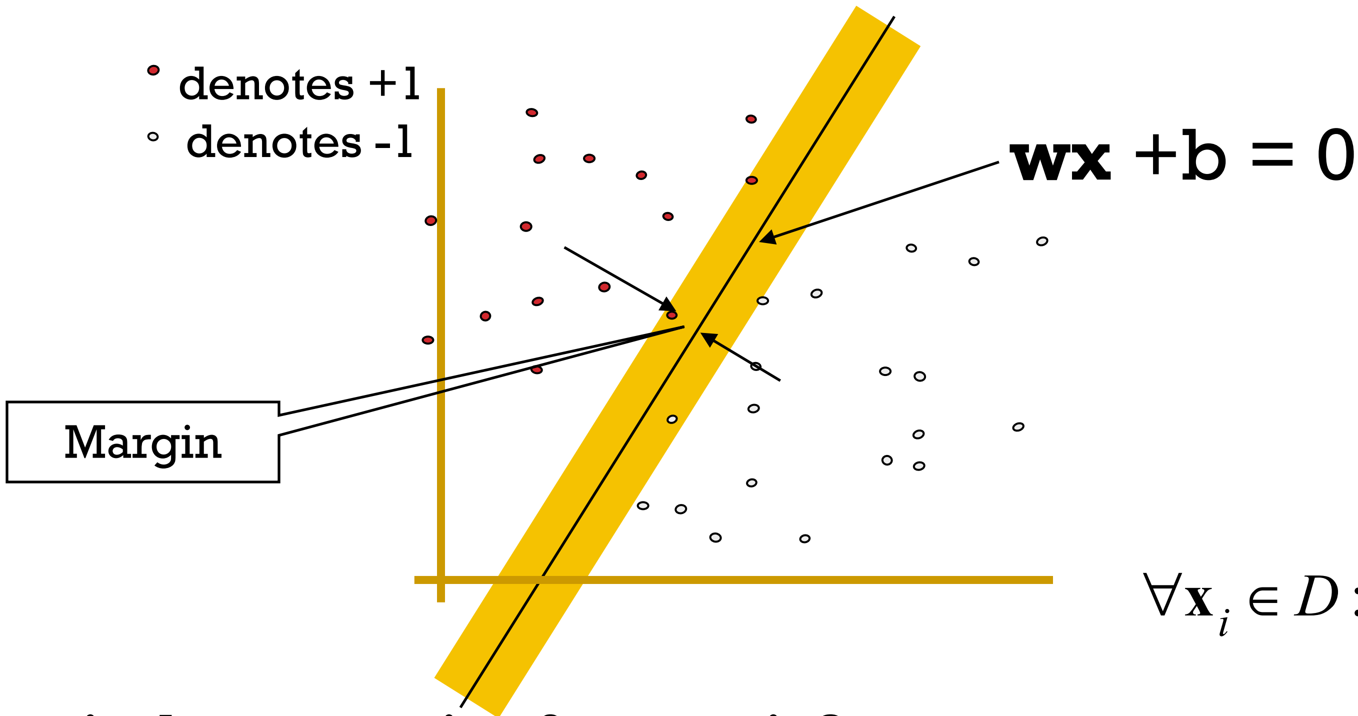
$$\forall \mathbf{x}_i \in D : y_i(\mathbf{x}_i \cdot \mathbf{w} + b) > 0$$

What is the distance expression for a point x to a line wx+b= 0?

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

Slides from Andrew W. Moore

ESTIMATE THE MARGIN



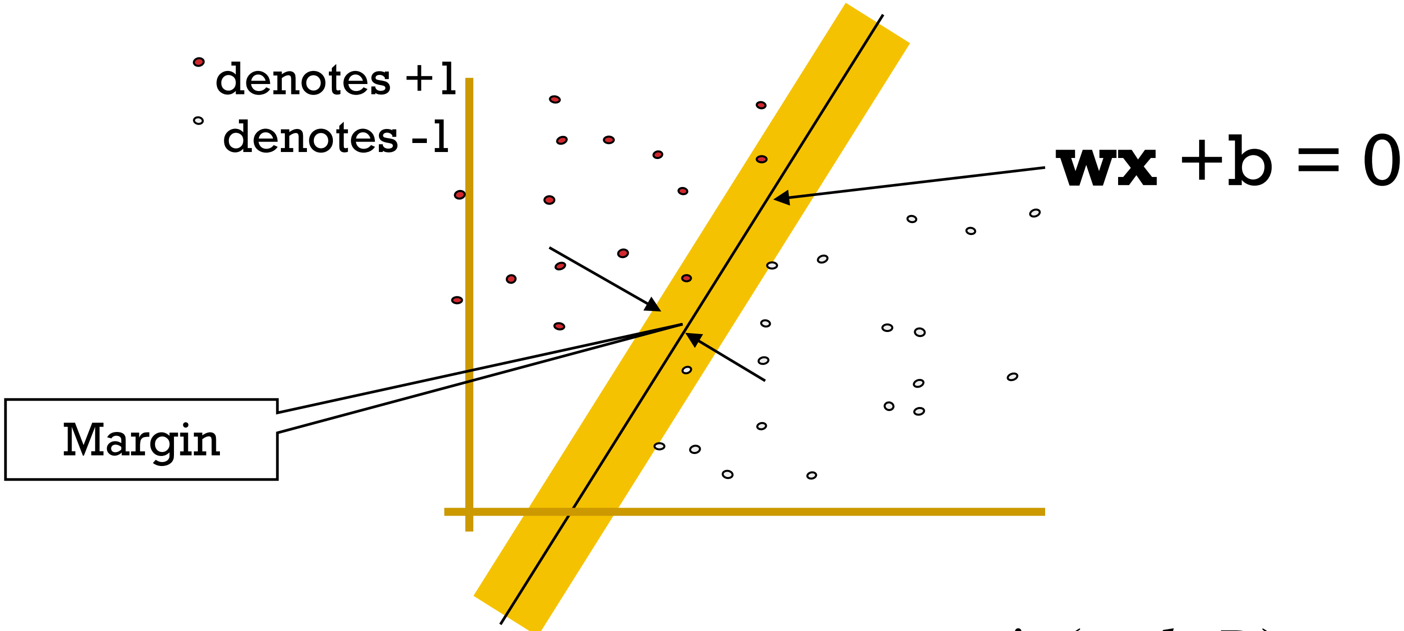
$$\forall \mathbf{x}_i \in D : y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) > 0$$

What is the expression for margin?

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

Slides from Andrew W. Moore

MAXIMIZE MARGIN



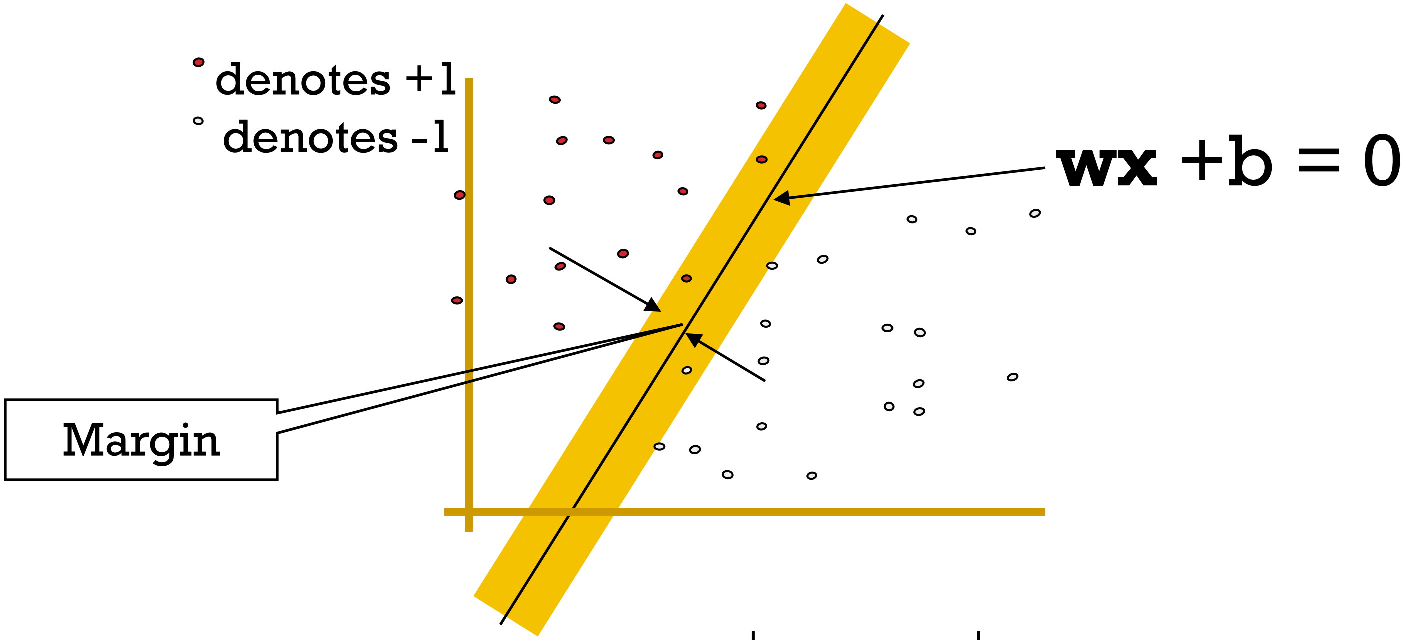
$$\underset{\mathbf{w},b}{\operatorname{argmax}} \ \operatorname{margin}(\mathbf{w}, b, D)$$

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

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

Slides from Andrew W. Moore

MAXIMIZE MARGIN

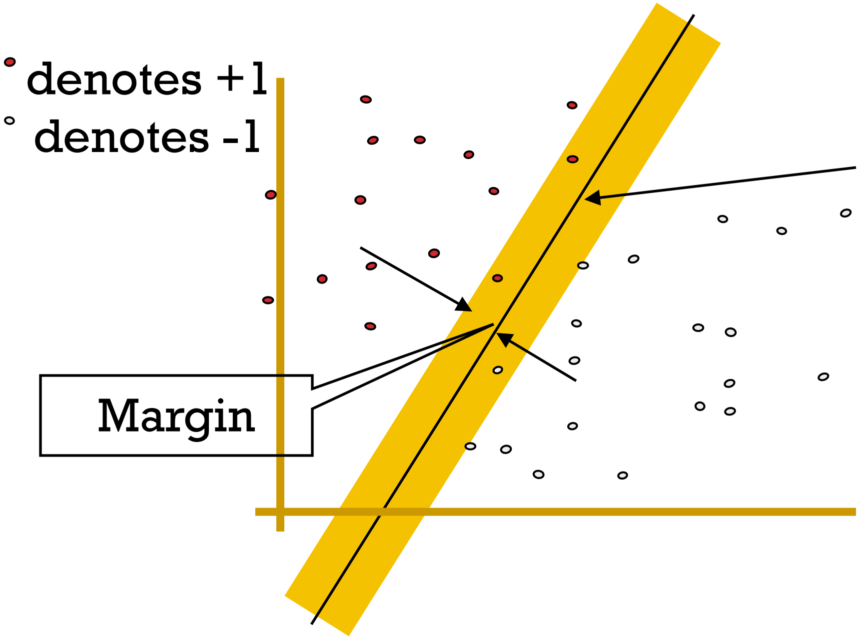


$$\underset{\mathbf{w},b}{\operatorname{argmax}} \min_{\mathbf{x}_i \in D} \frac{\left| b + \mathbf{x}_i \cdot \mathbf{w} \right|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

$$\text{subject to } \forall \mathbf{x}_i \in D : y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) > 0$$

Min-max problem → game problem

MAXIMIZE MARGIN



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

$$\text{subject to } \forall \mathbf{x}_i \in D : y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \geq 0$$

Strategy:

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

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

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

If you want to learn why this holds I highly recommend the following video
https://www.youtube.com/watch?v=_PwhiWxHK8o (not necessary to know for the

MAXIMUM MARGIN LINEAR CLASSIFIER

$$\{\vec{w}^*, b^*\} = \underset{\vec{w},b}{\operatorname{argmin}} \sum_{k=1}^{d} w_k^2$$

subject to

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

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

....

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

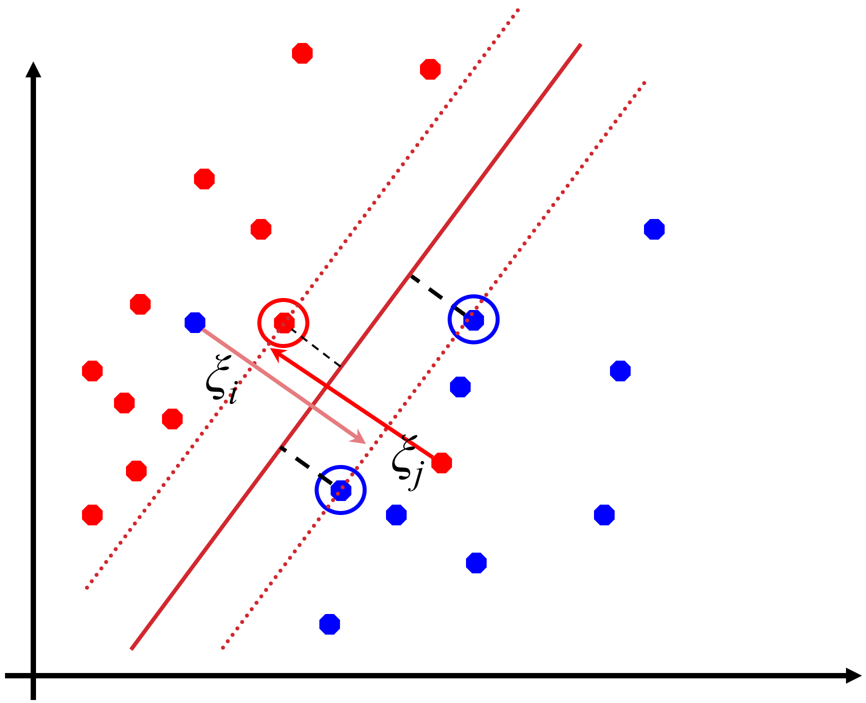
How to solve it?

LEARNING VIA QUADRATIC PROGRAMMING

QP is a well-studied cla[illegible]s of [illegible]ti[illegible]za[illegible]n a[illegible]gorit[illegible]ms to maximize a quadr[illegible] [illegible]ed [illegible]riables subject to line[illegible]

There exist algorithms for finding such constrained quadratic optima much more efficiently and reliably than gradient ascent.
(But they are very fiddly…you probably don't want to write one yourself)

SOFT MARGIN CLASSIFICATION



If the training data is not linearly separable, *slack variables* ξ_i can be added to allow misclassification of difficult or noisy examples.

Allow some errors

- Let some points be moved to where they belong, at a cost

Still, try to minimize training set errors, and to place hyperplane "far" from each class (large margin)

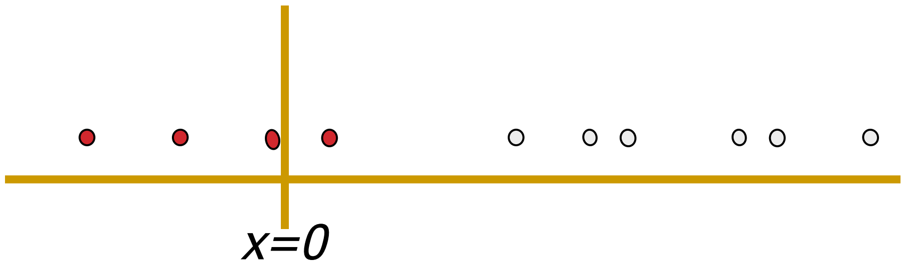
REGULARIZATION
SOFT MARGIN CLASSIFICATION MATHEMATICALLY

Find $\mathbf{w}$ and b such that
$\mathbf{\Phi}(\mathbf{w}) = \frac{1}{2}\, \mathbf{w}^{\mathrm{T}}\mathbf{w}$ is minimized and for all $\{(\mathbf{x_i}, y_i)\}$
$y_i\, (\mathbf{w^T x_i} + \mathrm{b}) \geq 1$

Find $\mathbf{w}$ and b such that
$\mathbf{\Phi}(\mathbf{w}) = \frac{1}{2}\, \mathbf{w}^{\mathrm{T}}\mathbf{w} + C \sum \xi_i$ is minimized and for all $\{(\mathbf{x_i}, y_i)\}$
$y_i\, (\mathbf{w^T x_i} + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for all i

SUPPOSE WE'RE IN 1-DIMENSION

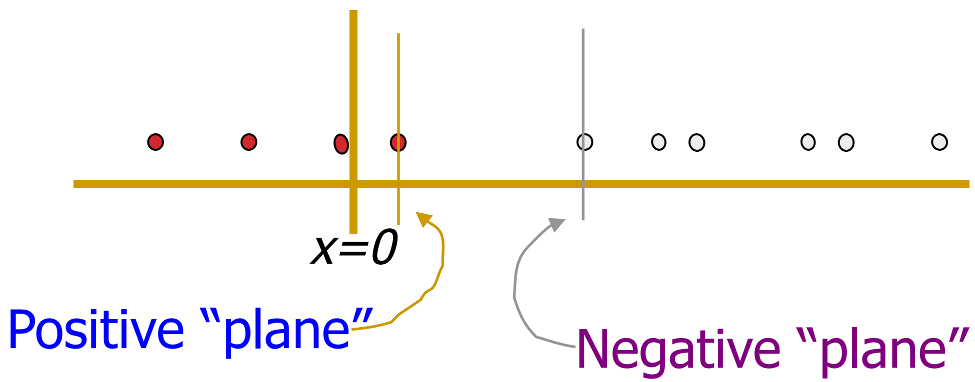
What would SVMs do with this data?



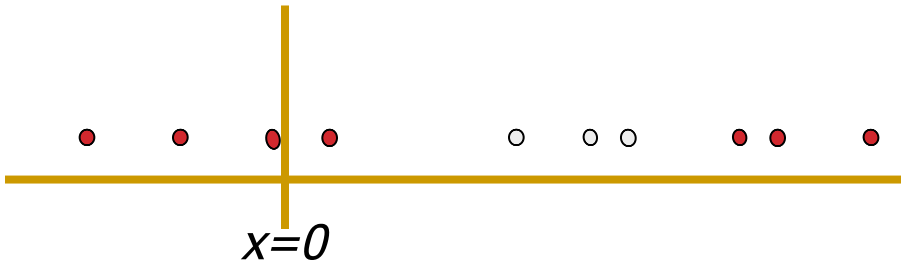
Slides from Andrew W. Moore

SUPPOSE WE'RE IN 1-DIMENSION

Not a big surprise



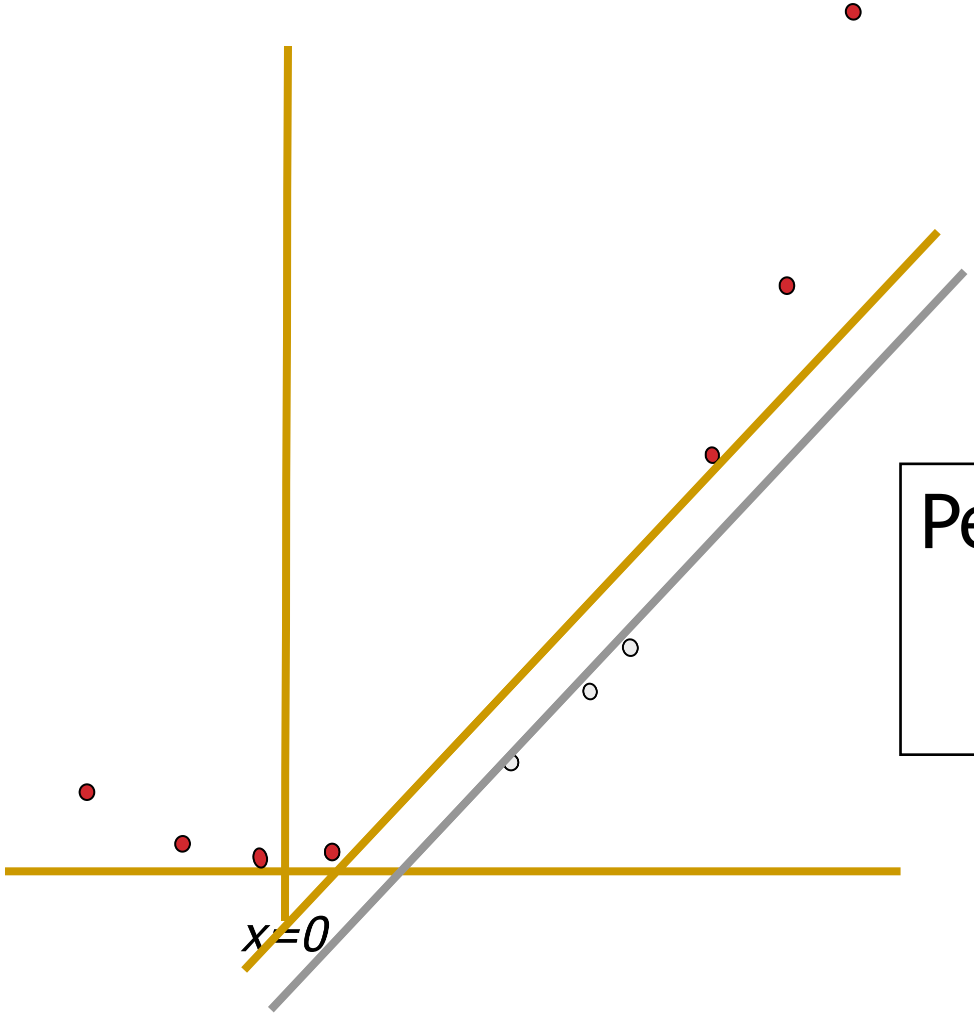
HARDER 1-DIMENSIONAL DATASET



That's wiped the smirk off SVM's face.

What can be done about this?

HARDER 1-DIMENSIONAL DATASET



Permitting non-linear basis functions

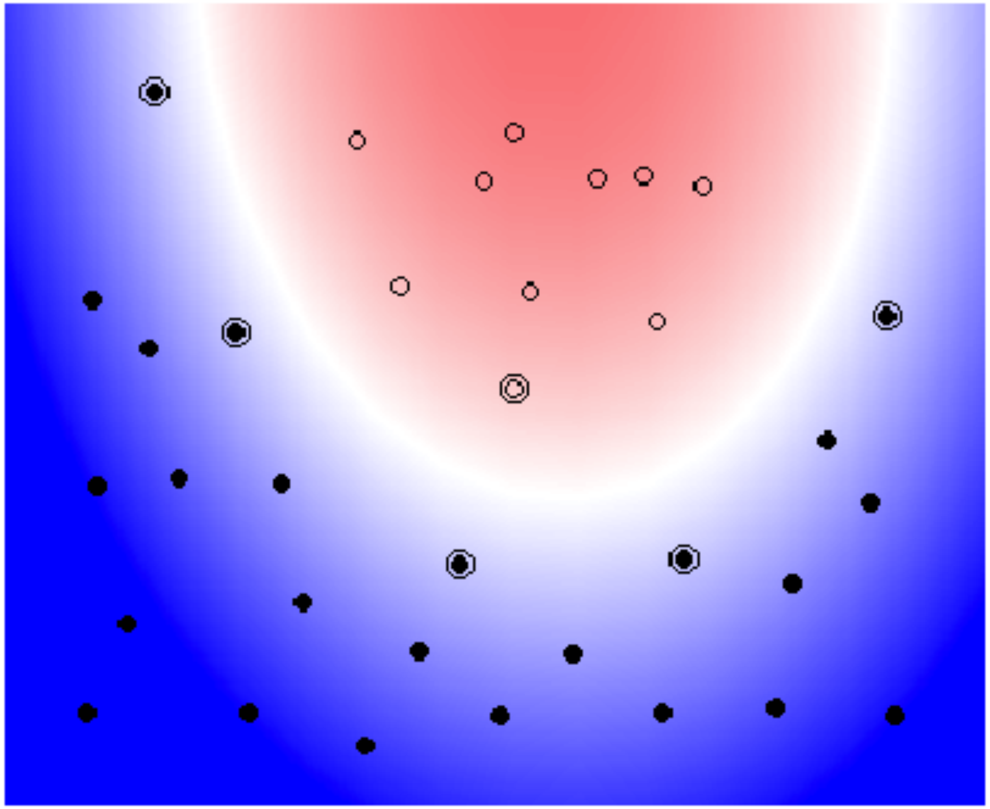
$$\mathbf{z}_k = (x_k, x_k^2)$$

NONLINEAR KERNEL (I)

Example: SVM with Polynomial of Degree 2

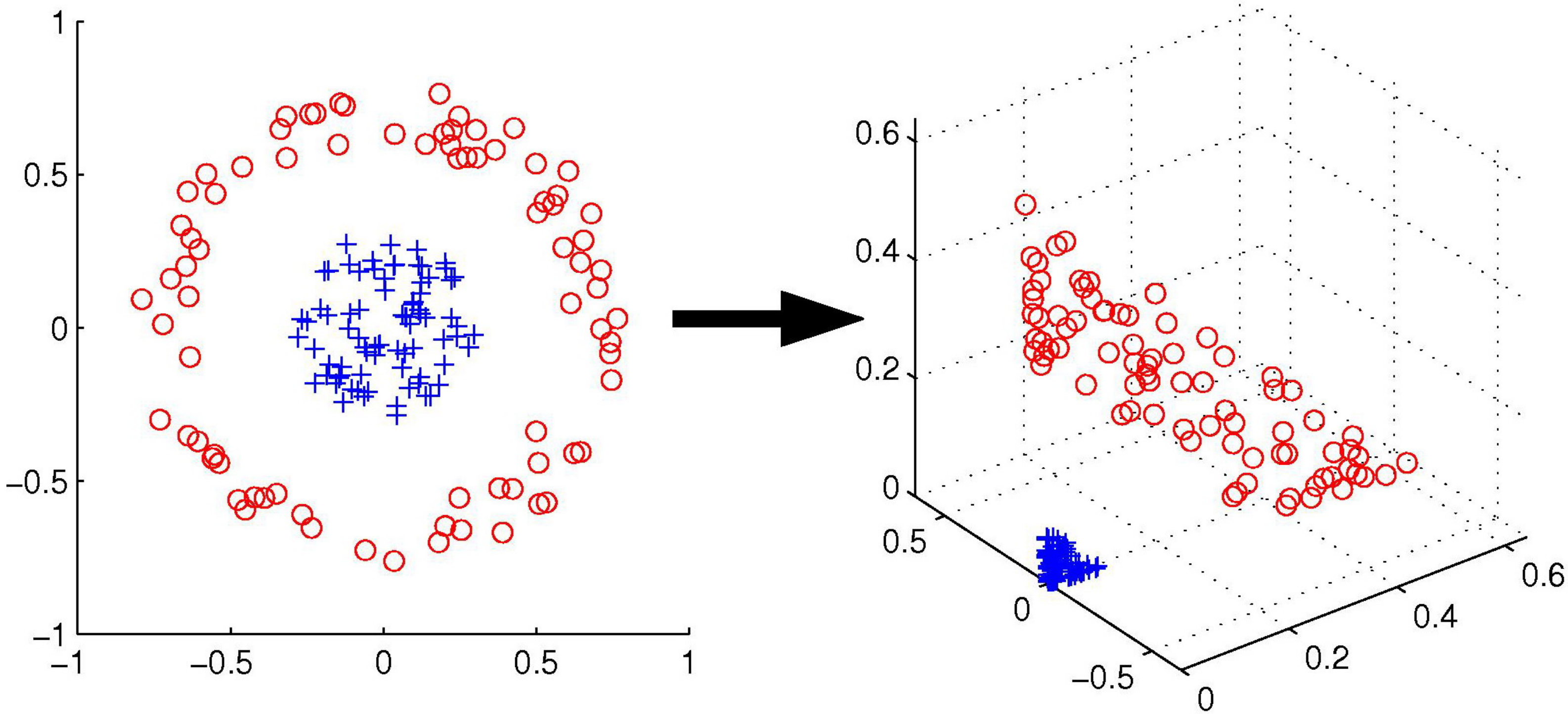
Kernel: $K(\vec{x}_i, \vec{x}_j) = [\vec{x}_i \cdot \vec{x}_j + 1]^2$

plot by Bell SVM applet



Slides from Andrew W. Moore

NON-LINEAR KERNEL



$$\begin{array}{rcccl} \phi : & \Re^2 & \longrightarrow & \Re^3 & \\ & (x_1, x_2) & \longmapsto & (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \end{array}$$

[http://www.cs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf]

SVM with a polynomial Kernel visualization

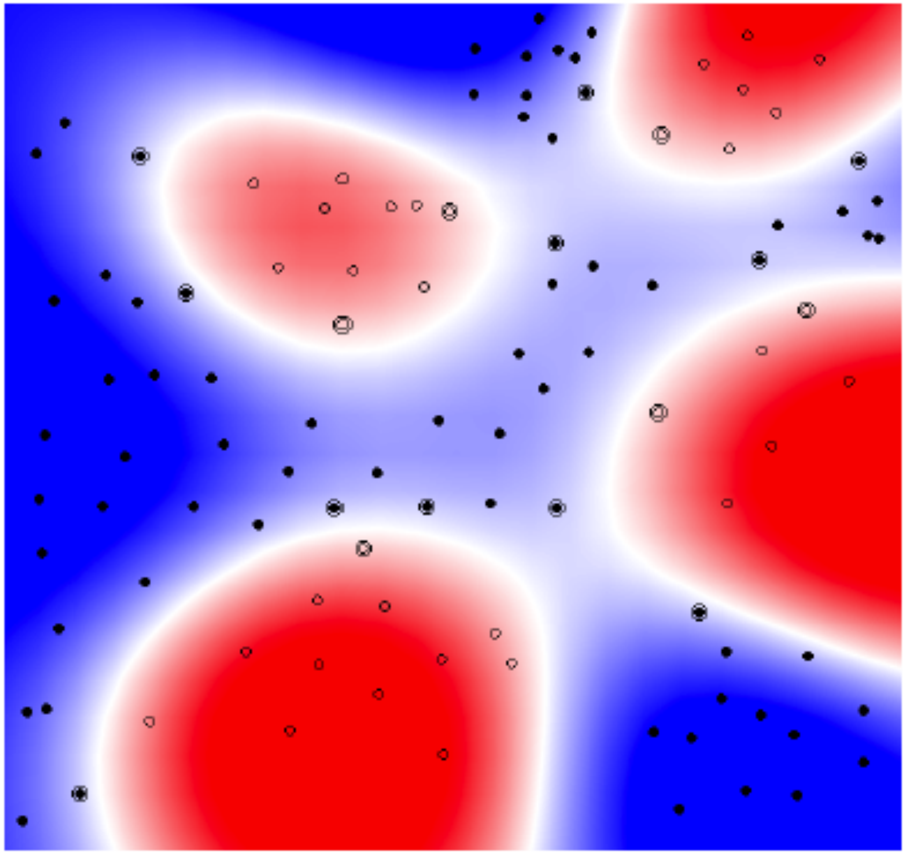
Created by:
Udi Aharoni

NONLINEAR KERNEL (II)

Example: SVM with RBF-Kernel

Kernel: $K(\vec{x}_i, \vec{x}_j) = \exp(-|\vec{x}_i - \vec{x}_j|^2 / \sigma^2)$

plot by Bell SVM applet



Nice video lecture from CalTech on RBF kernels:
https://www.youtube.com/watch?v=O8CfrnOPtLc

Slides from Andrew W. Moore

DEMO

http://cs.stanford.edu/people/karpathy/svmjs/demo/

KERNEL TRICKS

Pro

- Introducing nonlinearity into the model
- Computational cheap

Con

- Still have potential overfitting problems

Slides from Andrew W. Moore

SVM PERFORMANCE

Anecdotally they work very very well.

Example: They are currently the best-known classifier on a well-studied hand-written-character recognition benchmark

Anecdotally reliable people doing practical real-world work who claim that SVMs have saved them when their other favorite classifiers did poorly.

There is a lot of excitement and religious fervor about SVMs

Despite this, some practitioners are a little skeptical.

REFERENCES

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

Software: SVM-light, http://svmlight.joachims.org/, free download