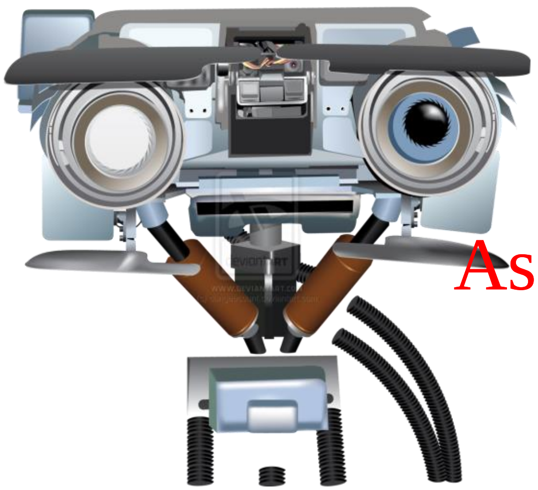
Announcements

Reminder: ps4 self-grading form out, due Friday 10/30

- pset 5 out today 10/29, due 11/5 (1 week)

- Midterm grades will go up by Monday (don't discuss it yet)



Support Vector Machines

CS542 Machine Learning

slides based on lecture by R. Urtasun

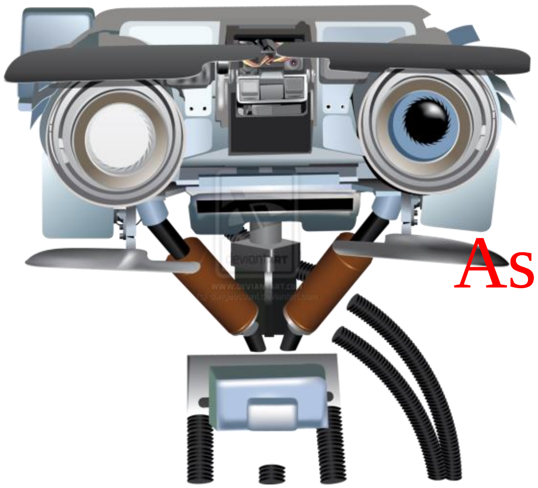
http://www.cs.toronto.edu/~urtasun/courses/CSC2515/CSC2515_Winter15.html

Support Vector Machine (SVM)

- A *maximum margin* method, can be used for classification or regression

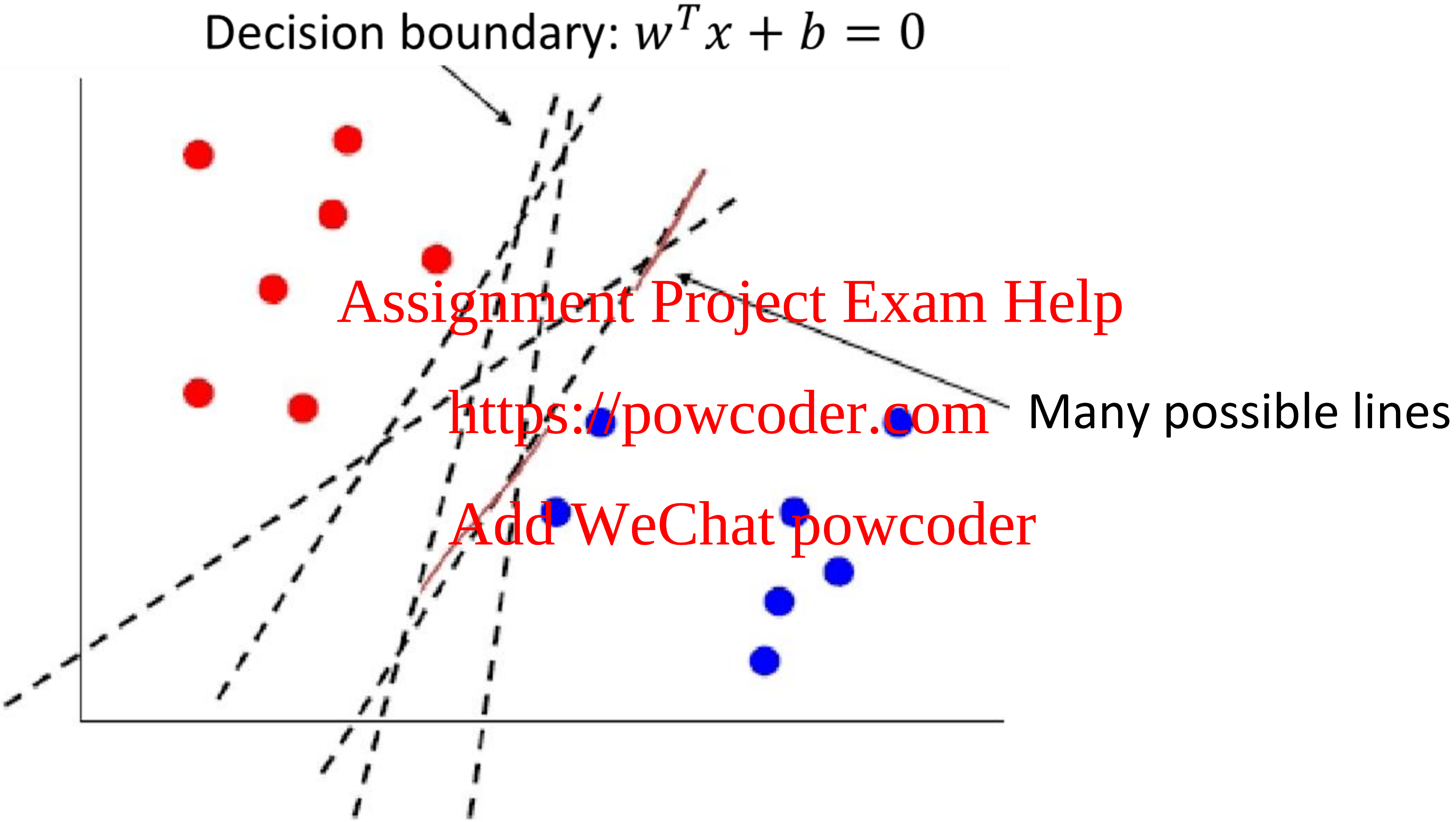
- SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces

- First, we will derive *linear, hard-margin SVM* for linearly separable data, later for non-separable (soft margin SVM), and for nonlinear boundaries (kernel SVM)



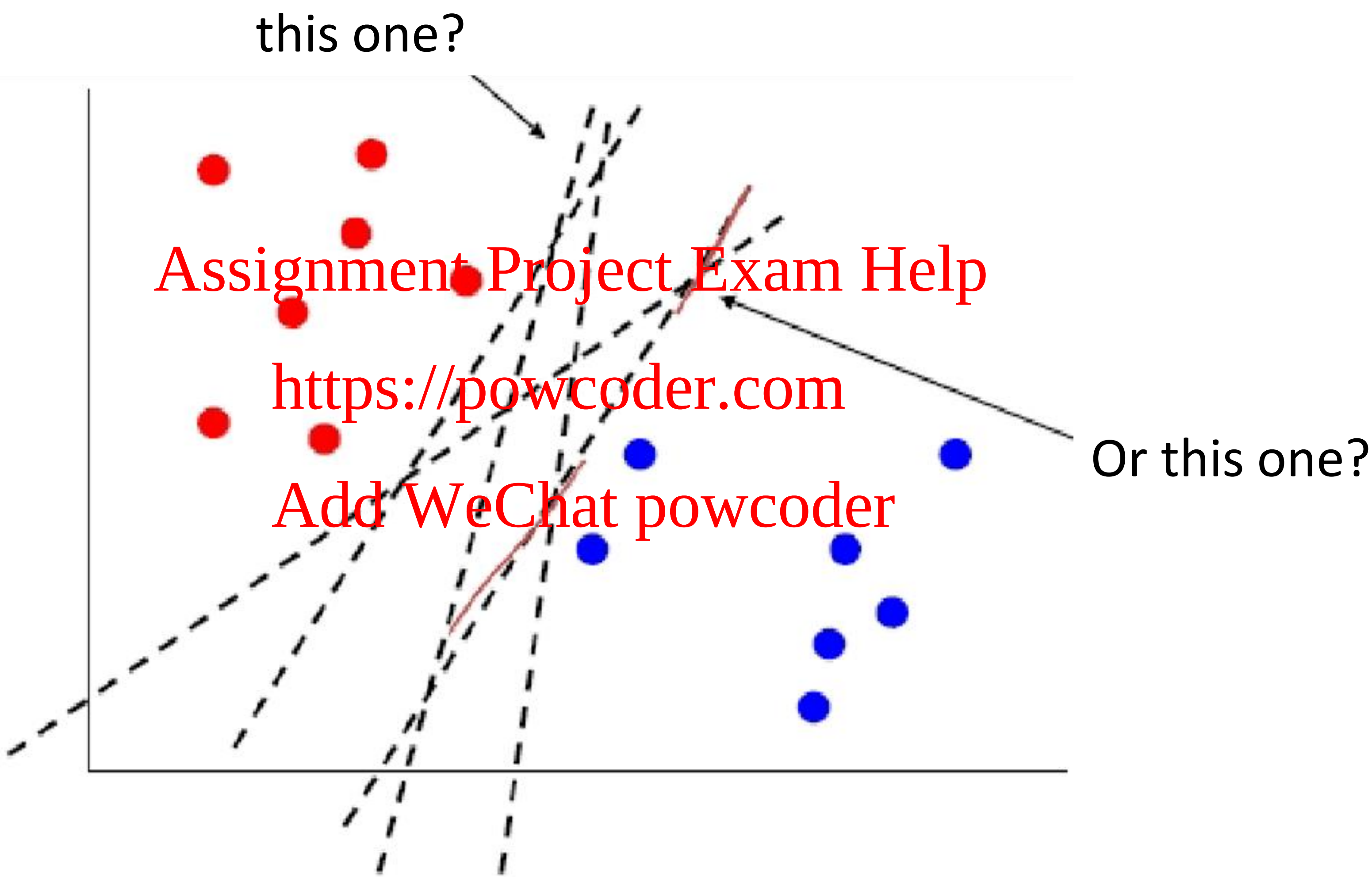
Maximum Margin

Recall: logistic regression



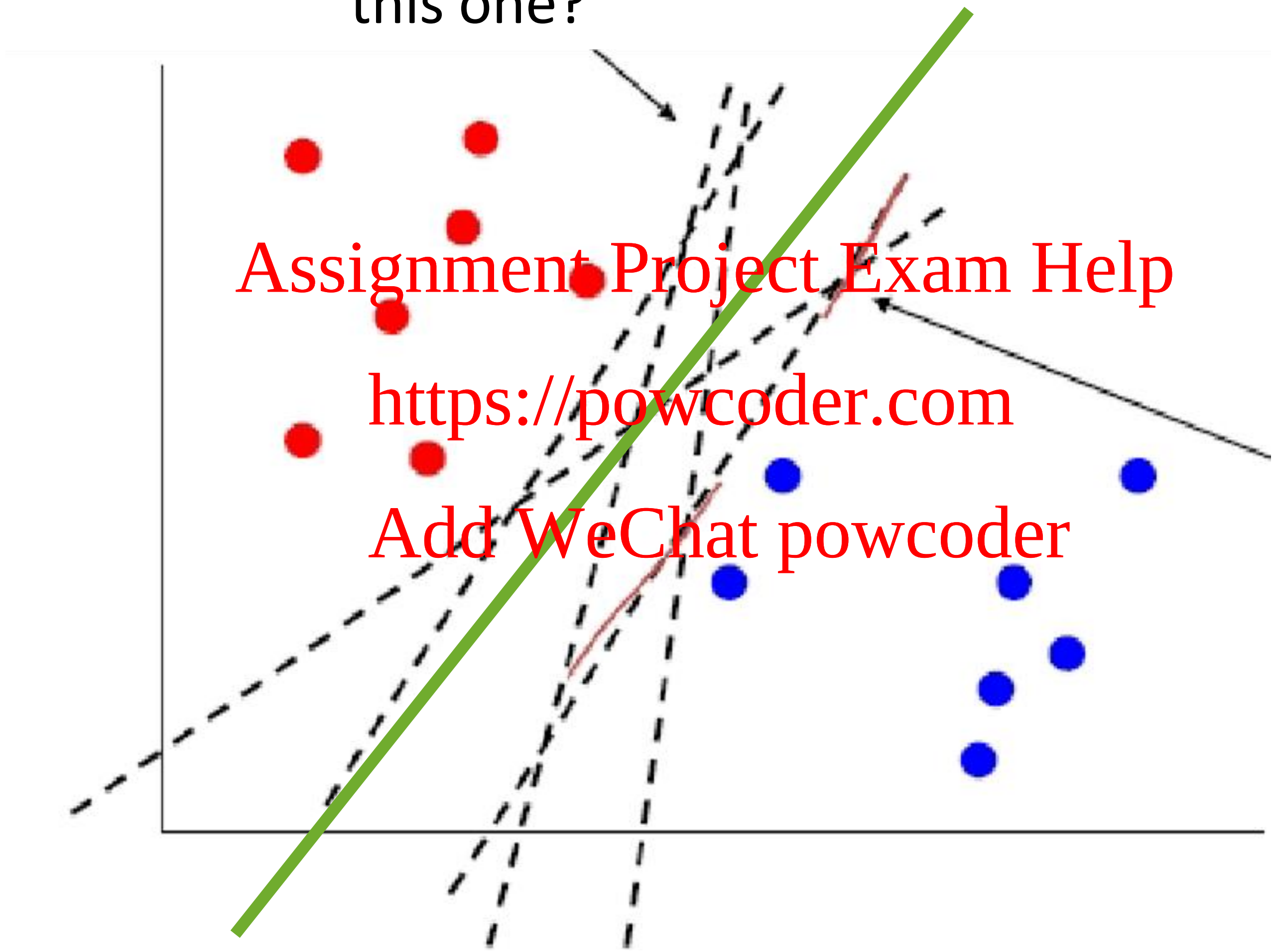
$$y = \begin{cases} +1\ [\text{red}]\ \text{if sign}(\mathbf{w}^T\mathbf{x} + b \geq 0) \\ -1\,[\text{blue}]\ \text{if sign}(\mathbf{w}^T\mathbf{x} + b < 0) \end{cases}$$

Which classifier is best?



How about the one in the middle?

this one?



Or this one?

Intuitively, this classifier avoids misclassifying new test points generated from the same distribution as the training points

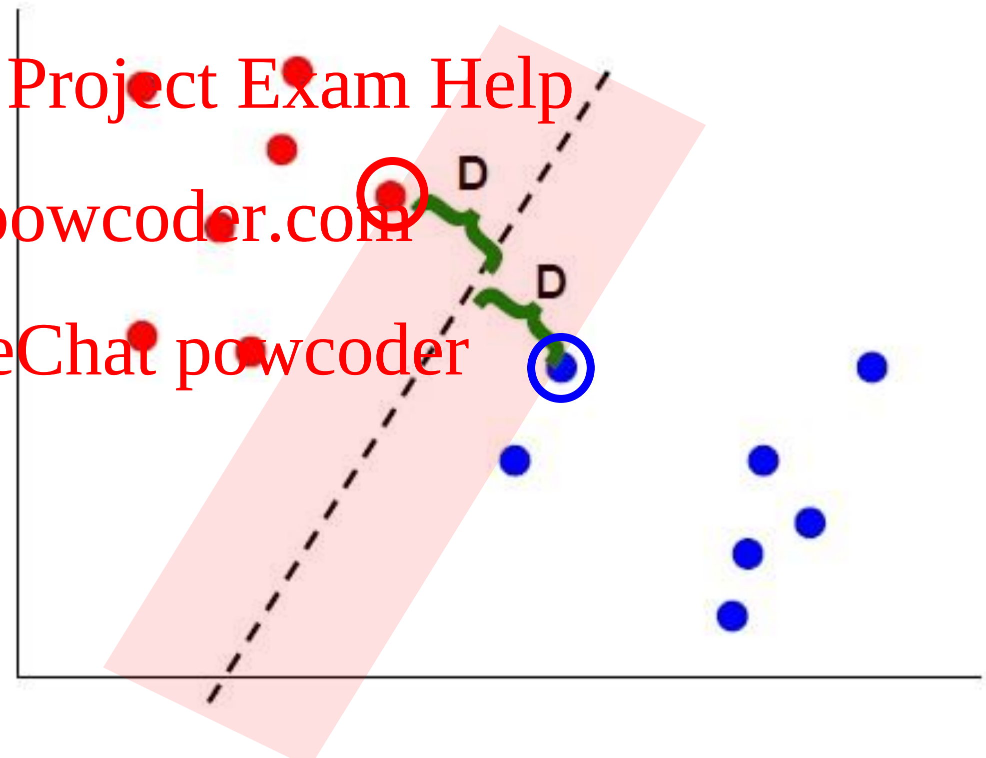
Max margin classification

Instead of fitting all the points, focus on boundary points

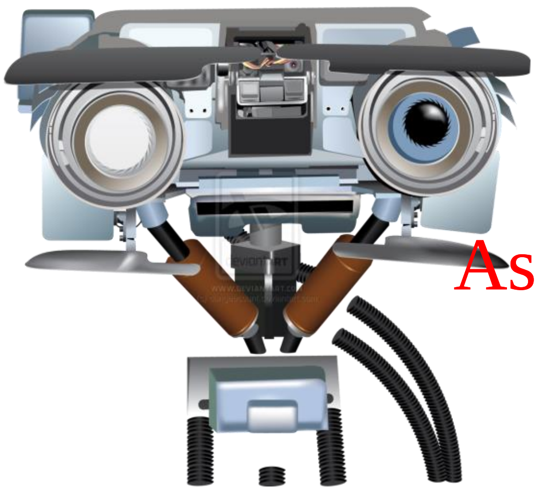
Aim: learn a boundary that leads to the largest margin (buffer) from points on both sides

Why: intuition; theoretical support: robust to small perturbations near the boundary

And works well in practice!



Subset of vectors that support (determine boundary) are called the support vectors (circled)



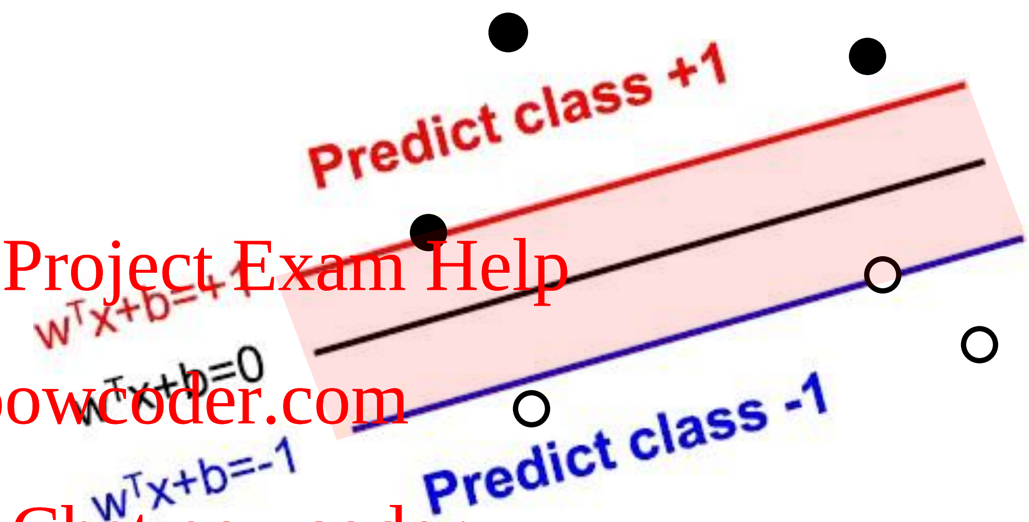
Max-Margin Classifier

Max Margin Classifier

"Expand" the decision boundary to include a margin (until we hit first point on either side)

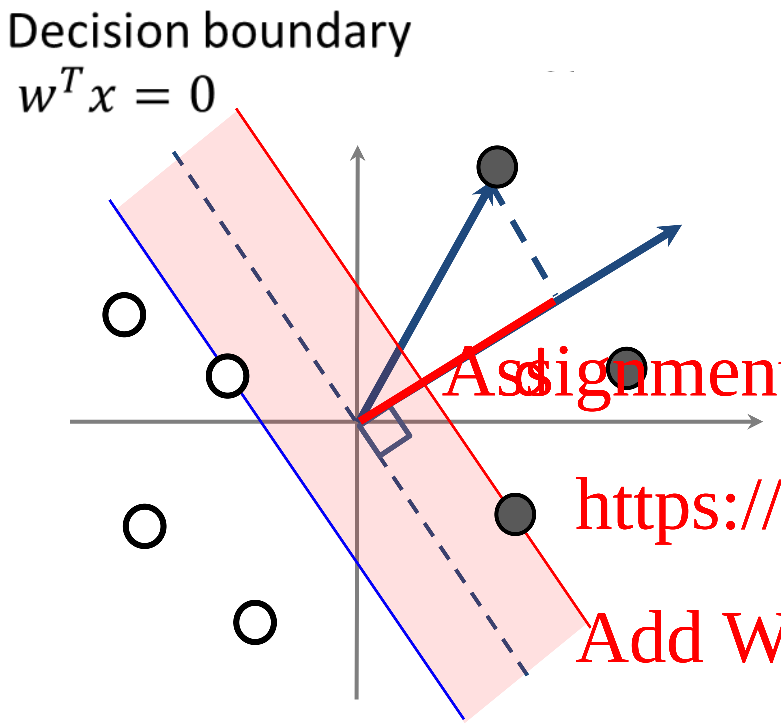
Use margin of 1

Inputs in the margins are of unknown class



Classify as +1	if	$w^Tx+b \geq 1$
Classify as -1	if	$w^Tx+b \leq -1$
Undefined	if	$-1 < w^Tx+b < 1$

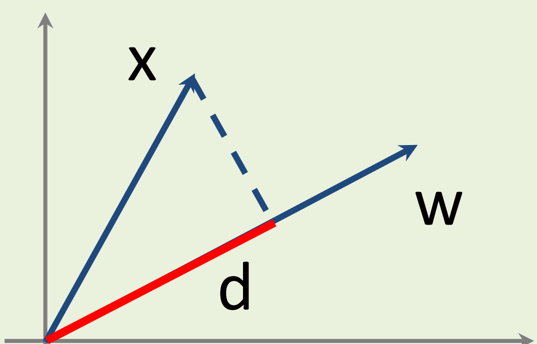
Why is the margin = 1?



- Assume $b = 0$ for simplicity
- $\mathbf{w}$ is orthogonal to the decision plane
- Scaling margin and weight vector by the same constant c>0 does not change the inequality

$$\boldsymbol{w}^T \boldsymbol{x} \geq 1$$

$$c * \boldsymbol{w}^T \boldsymbol{x} \geq 1 * c$$



d is the length of projection

Aside: vector inner product

$$\boldsymbol{w}^T \boldsymbol{x} = d \|\boldsymbol{w}\|_2 =$$
$$= \mathrm{w}_1 \mathrm{x}_1 + \mathrm{w}_2 \mathrm{x}_2$$

$$d = \frac{\boldsymbol{w}^T \boldsymbol{x}}{\|\boldsymbol{w}\|_2}$$

Computing the Margin

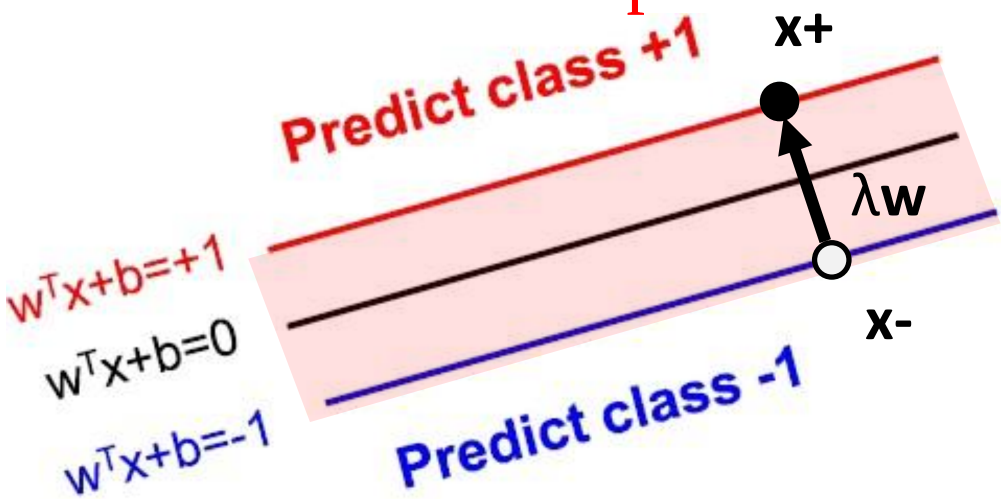
First note that the **w** vector is orthogonal to the +1 plane

If **u** and **v** are two points on that plane, then $\mathbf{w}^T(\mathbf{u}-\mathbf{v}) = 0$

Same is true for -1 plane

Also: for point $\mathbf{x}^+$ on +1 plane and $\mathbf{x}^-$ nearest point on -1 plane:

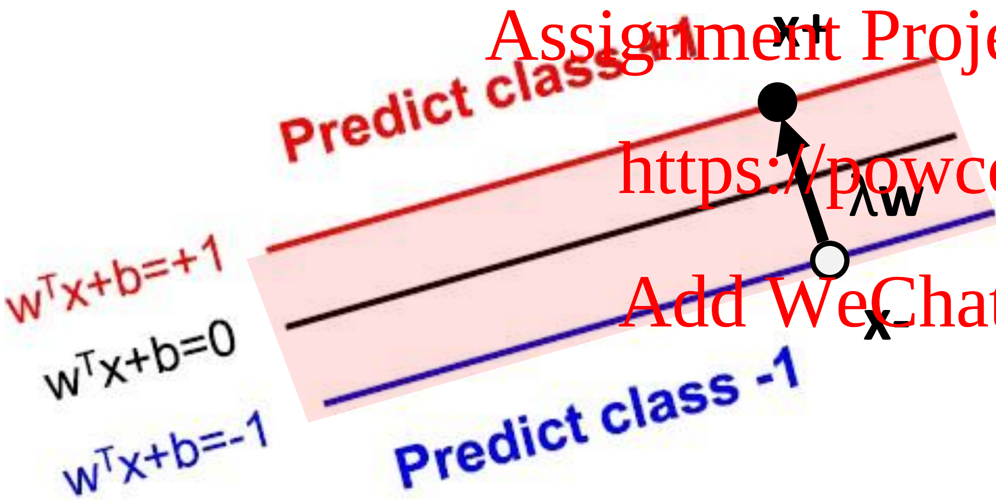
$$\mathbf{x}^+ = \lambda\mathbf{w} + \mathbf{x}^-$$



Computing the Margin

Also: for point **x+** on +1 plane and **x-**nearest point on -1 plane:

$$\mathbf{x}^{+} = \lambda \mathbf{w} + \mathbf{x}^{-}$$



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

$$\mathbf{w}^{T}(\lambda\mathbf{w} + \mathbf{x}^{-}) + b = 1$$

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

$$-1 + \lambda\mathbf{w}^{T}\mathbf{w} = 1$$

$$\lambda = 2 / \mathbf{w}^{T}\mathbf{w}$$

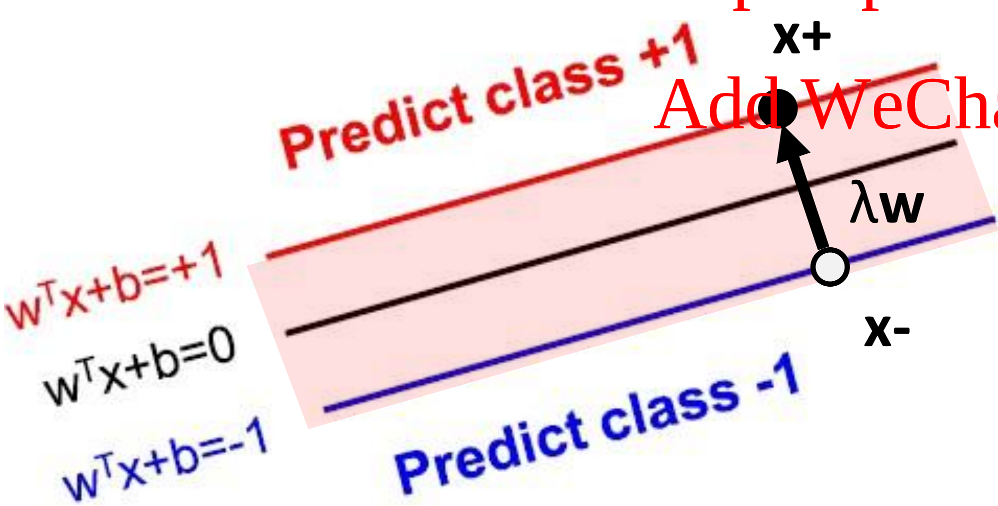
→ inversely proportional to $\mathbf{w}^{\mathrm{T}}\mathbf{w}$, the square of the length of $\mathbf{w}$

Computing the Margin

Define the margin M to be the distance between the +1 and -1 planes

We can now express this in terms of **w** □

to maximize the margin we minimize the length of **w**



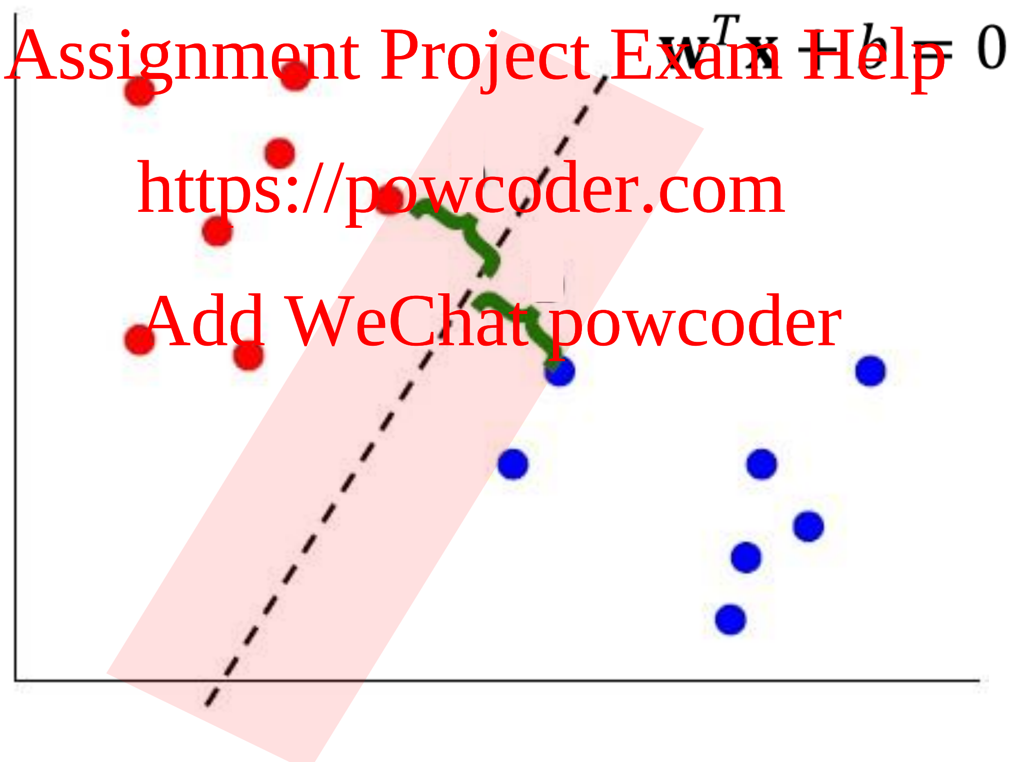
$$M = \left\| \mathbf{x}^{+} - \mathbf{x}^{-} \right\|$$

$$= \left\| \lambda \mathbf{w} \right\| = \lambda \sqrt{\mathbf{w}^{T}\mathbf{w}}$$

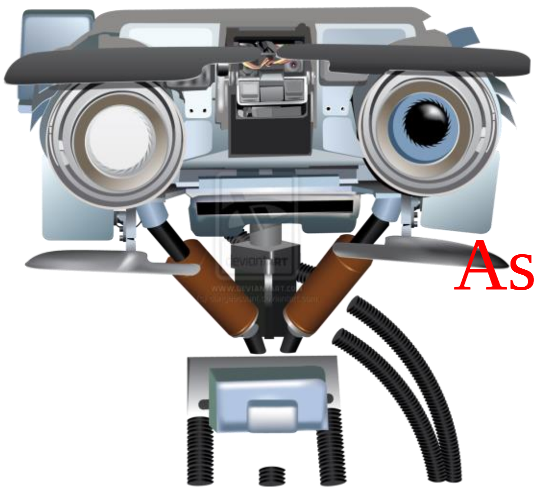
$$= 2 \frac{\sqrt{\mathbf{w}^{T}\mathbf{w}}}{\mathbf{w}^{T}\mathbf{w}} = \frac{2}{\sqrt{\mathbf{w}^{T}\mathbf{w}}}$$

Maximizing the margin is equivalent to regularization

To maximize the margin we minimize the length of $\mathbf{w}$, or $||\mathbf{w}||^2$



But not same as regularized logistic regression, the SVM loss is different! Only care about boundary points.

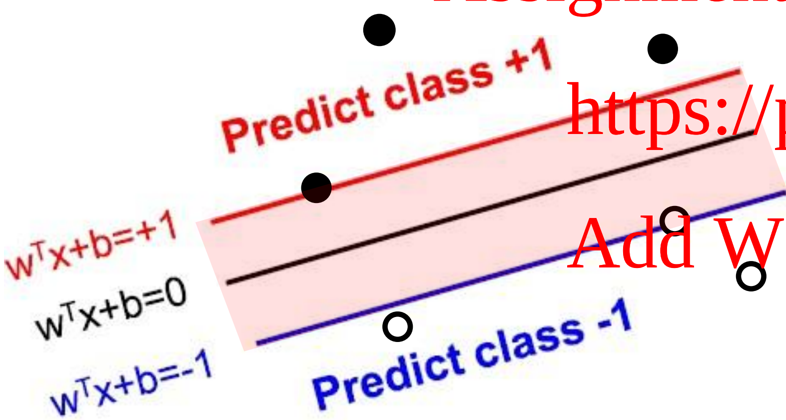


Linear SVM

Linear SVM Formulation

We can search for the optimal parameters (**w** and b) by finding a solution that:

1. Correctly classifies the training examples: $\{x_i,y_i\}$, i=1,..,n
2. Maximizes the margin (same as minimizing $\|\boldsymbol{w}\|^2$)



$$\min \ \frac{1}{2}\|\mathbf{w}\|^2$$

$$s.t. \ (\mathbf{w}^T\mathbf{x}_i + b)y_i \geq 1 \ \ \forall i$$

This is the primal formulation, can be optimized via gradient descent, EM, etc.

Apply Lagrange multipliers: formulate equivalent problem

Lagrange Multipliers

Convert the primal constrained minimization to an unconstrained optimization problem: represent constraints as penalty terms:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + \boldsymbol{penalty\_term}$$

For data $\{(x_i,y_i)\}$ use the following penalty term:

$$\begin{cases} 0 & \text{if } (\mathbf{w}^T\mathbf{x}_i + b)y_i \geq 1 \\ \infty & \text{otherwise} \end{cases} = \max_{\alpha_i \geq 0} \alpha_i[1-(\mathbf{w}^T\mathbf{x}_i + b)y_i]$$

≤ 0 if constraint satisfied

Introduced Lagrange variables $\alpha_i \geq 0$; find ones that maximize term:

- If a constraint is satisfied, large α_i ensures smaller penalty
- If a constraint is violated, large α_i ensures larger penalty

Note, we are now minimizing with respect to **w** and b, and maximizing with respect to **a** (additional parameters)

Lagrange Multipliers

Convert the primal constrained minimization to an unconstrained optimization problem: represent constraints as penalty terms:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + \boldsymbol{penalty\_term}$$

For data $\{(x_i,y_i)\}$ use the following penalty term:

$$\begin{cases} 0 & \text{if } (\mathbf{w}^T\mathbf{x}_i + b)y_i \geq 1 \\ \infty & \text{otherwise} \end{cases} = \max_{\alpha_i \geq 0} \alpha_i[1-(\mathbf{w}^T\mathbf{x}_i + b)y_i]$$

Rewrite the minimization problem:

$$\min_{\mathbf{w},b}\{\frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\max_{\alpha_i \geq 0}\alpha_i[1-(\mathbf{w}^T\mathbf{x}_i + b)y_i]\}$$

Where $\{\alpha_i\}$ are the Lagrange multipliers

$$\min_{\mathbf{w},b}\max_{\alpha_i \geq 0}\{\frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\alpha_i[1-(\mathbf{w}^T\mathbf{x}_i + b)y_i]\}$$

Solution to Linear SVM

Swap the 'max' and 'min':

$$\max_{\alpha_i \geq 0} \min_{\mathbf{w},b} \{\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [1 - (\mathbf{w}^T \mathbf{x}_i + b) y_i]\}$$

$$= \max_{\alpha_i \geq 0} \min_{\mathbf{w},b} J(\mathbf{w}, b; \alpha)$$

First minimize J() w.r.t. {w,b} for any fixed setting of the Lagrange multipliers:

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}, b; \alpha) = \mathbf{w} - \sum_{i=1}^{n} \alpha_i \mathbf{x}_i y_i = 0$$

$$\frac{\partial}{\partial b} J(\mathbf{w}, b; \alpha) = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

Then substitute back into J() and simplify to get final optimization:

$$L = \max_{\alpha_i \geq 0} \{\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)\}$$

Dual Problem

Final optimization: maximize this loss over α_i's:

only dot products of data points needed

$$L = \max_{\alpha_i \geq 0} \{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \}$$

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

Then use the obtained α_i's to solve for the weights and bias

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad b = y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_{\mathrm{i}} \quad \forall i$$

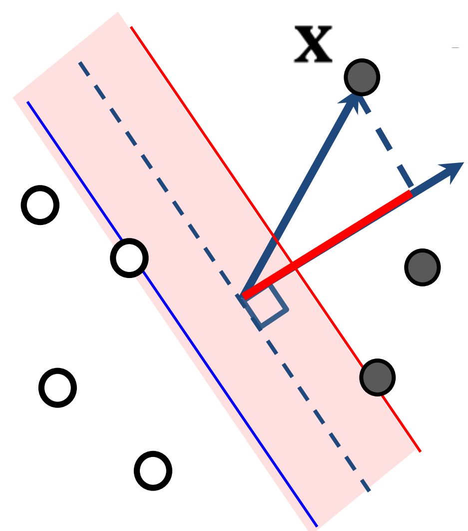
Prediction on Test Example

Now we have the solution for the weights and bias

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad\qquad b = y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathrm{i}} \quad \forall i$$

Given a new input example **x**, classify it as

$$+1 \text{ if } \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \geq 1, \text{ or}$$
$$-1 \text{ if } \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \leq -1$$



In practice, predict $y = \mathrm{sign}[\mathbf{w}^{\mathrm{T}}\mathbf{x} + b]$

Dual vs Primal SVM

n is the number of training points, d is dimension of $\mathbf{x}$, $\mathbf{w}$

Primal problem: for $\mathbf{w} \in \mathbb{R}^{\boldsymbol{d}}$, hyperparameter C, the unconstrained

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\|^2 + C\sum_{i}^{n} \max(0, 1 - y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)$$

Dual problem: for $\boldsymbol{\alpha} \in \mathbb{R}^{\boldsymbol{n}}$

$$L = \max_{\alpha_i \geq 0}\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)\} \quad \text{s.t } \alpha_i \geq 0; \ \sum_{i=1}^{n}\alpha_i y_i = 0$$

- Efficiency: need to learn d parameters for primal, n for dual

Dual vs Primal SVM

- Dual: quadratic programming problem in which we optimize a quadratic function of **a** subject to a set of inequality constraints

- The solution to a quadratic programming problem in d variables in general has computational complexity that is $O(d^3)$

- For a fixed set of basis functions whose number *d* is smaller than the number *n* of data points, the move to the dual problem appears disadvantageous.

- However, it allows the model to be reformulated using kernels which allow *infinite* feature spaces (more on this later)

Dual vs Primal SVM

- Most of the SVM literature and software solves the Lagrange dual problem formulation

- Why prefer solving the dual problem over the primal?
  - provides a way to deal with constraints
  - expresses solution in terms of dot products of data points, allowing kernels
  - historical reasons

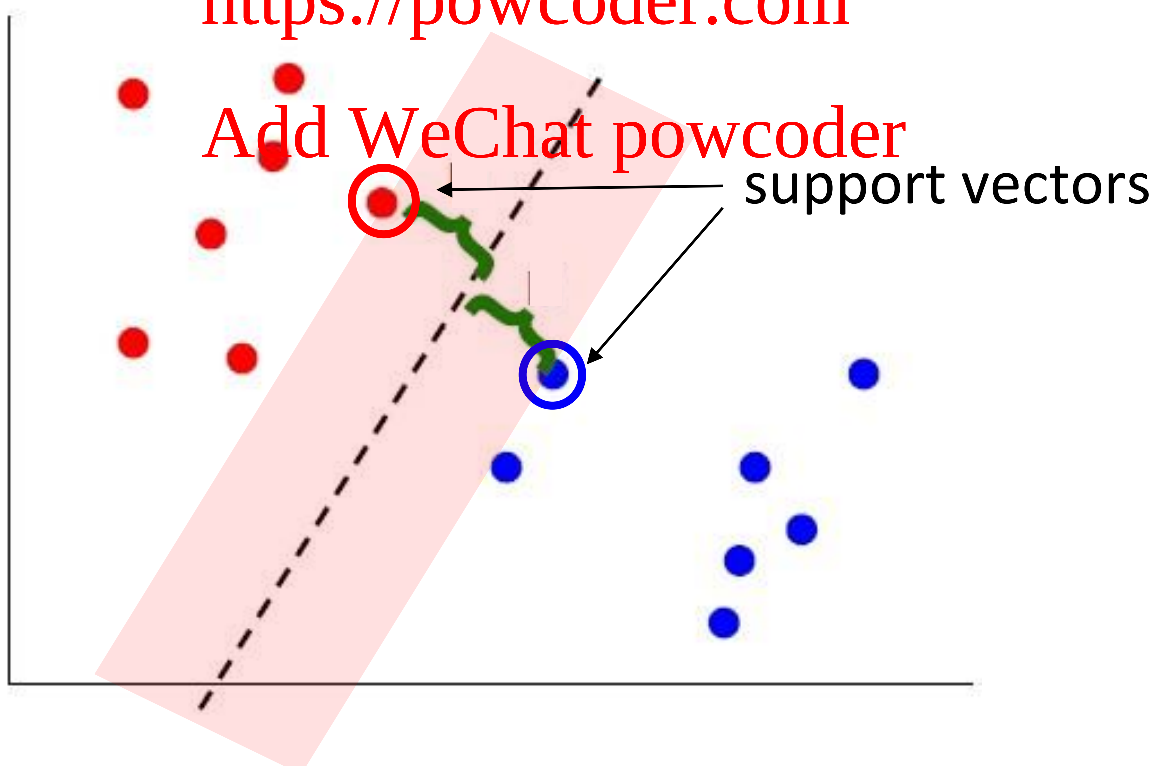
For an in-depth discussion refer to

http://olivier.chapelle.cc/pub/neco07.pdf (optional reading)

Support Vectors

Only a small subset of α_i's will be nonzero, and the corresponding x_i's are the support vectors **S**

$$y = \operatorname{sign}[b + \mathbf{x}\cdot(\sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i)] = \operatorname{sign}[b + \mathbf{x}\cdot(\sum_{i \in \mathbf{S}} y_i \alpha_i \mathbf{x}_i)]$$



Summary of Linear SVM

- Binary and linear separable classification (regression possible too)
- Linear classifier with maximal margin
- Training SVM by maximizing

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

- Subject to $\alpha_i \geq 0; \sum_{i=1}^{n}\alpha_i y_i = 0$

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

- Only a small subset of α_i's will be nonzero, and the corresponding x_i's are the support vectors **S**
- Prediction on a new example:

$$y = \mathrm{sign}[b + \mathbf{x}\cdot(\sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i)] = \mathrm{sign}[b + \mathbf{x}\cdot(\sum_{i\in \mathbf{S}} y_i \alpha_i \mathbf{x}_i)]$$

Next Class

Support Vector Machines II

non-separable data; slack variables;kernels; multiclass SVM

Reading: Bishop Ch 6.1-6.2, Ch 7.1.3