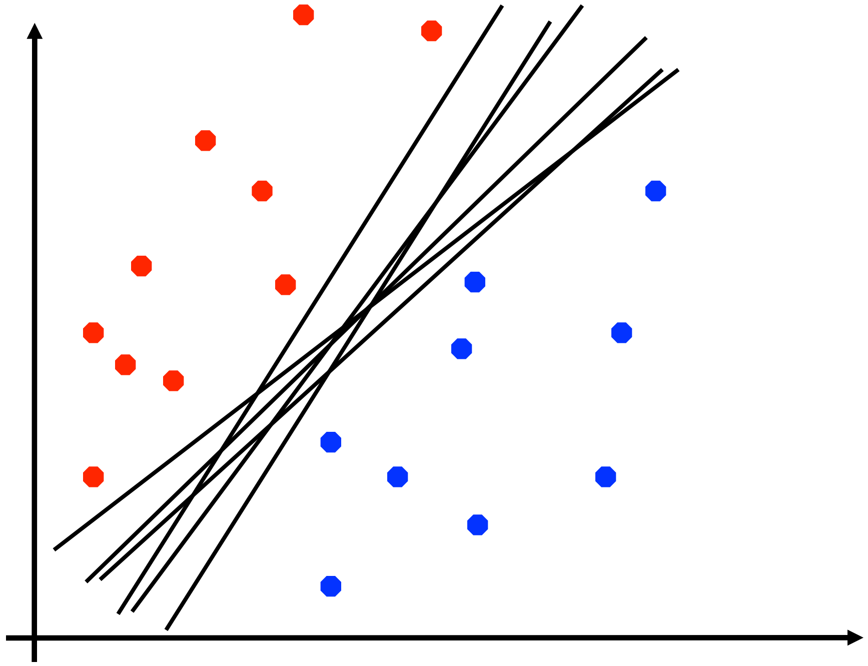
Machine Learning
Support Vector Machines (SVMs), Part 1

Slides adapted from David Sontag, who adapted from Luke Zettlemoyer, Vibhav Gogate, and Carlos Guestrin

Linear Separators

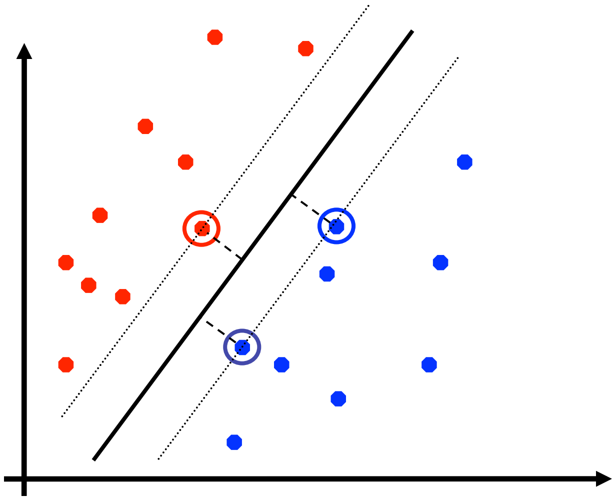
- If training data is linearly separable, perceptron is guaranteed to find *some* linear separator
- Which of these is **optimal**? How do we define optimal?



Support Vector Machine (SVM)

- SVMs (Vapnik, 1990’s) choose the linear separator with the **largest margin**

Robust to outliers!

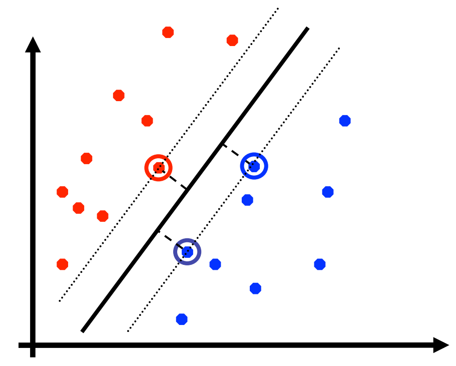


V. Vapnik

- Good according to intuition, theory, practice

- SVM became famous when, using images as input, it gave accuracy comparable to neural-network in a handwriting recognition task

Support vector machines: 3 key ideas



1. Use **optimization** to find solution with largest margin

2. Seek **large margin** separator while allowing for some test errors

3. Use **kernel trick** to make large non-linear feature spaces computationally efficient

Optimization Problem

- Assume for now data is linearly separable
- Goal: Formulate the max margin problem as a tractable optimization problem:
  - Maximize an objective function subject to constraints, giving optimal **w** and b.

A Few Words About Linear Programming

- Class of constrained optimization problems
- Examples given on board
- Well studied, can solve problems with millions of variables and constraints
  - Many software packages, textbooks
- Quadratic programs are also tractable by similar techniques.

Going to Show for Linearly Separable Case

Theorem: Let **w**, b be an optimal solution to:

Minimize $||\mathbf{w}||^2 = w_1^2 + w_2^2 + \ldots + w_n^2$

subject to:

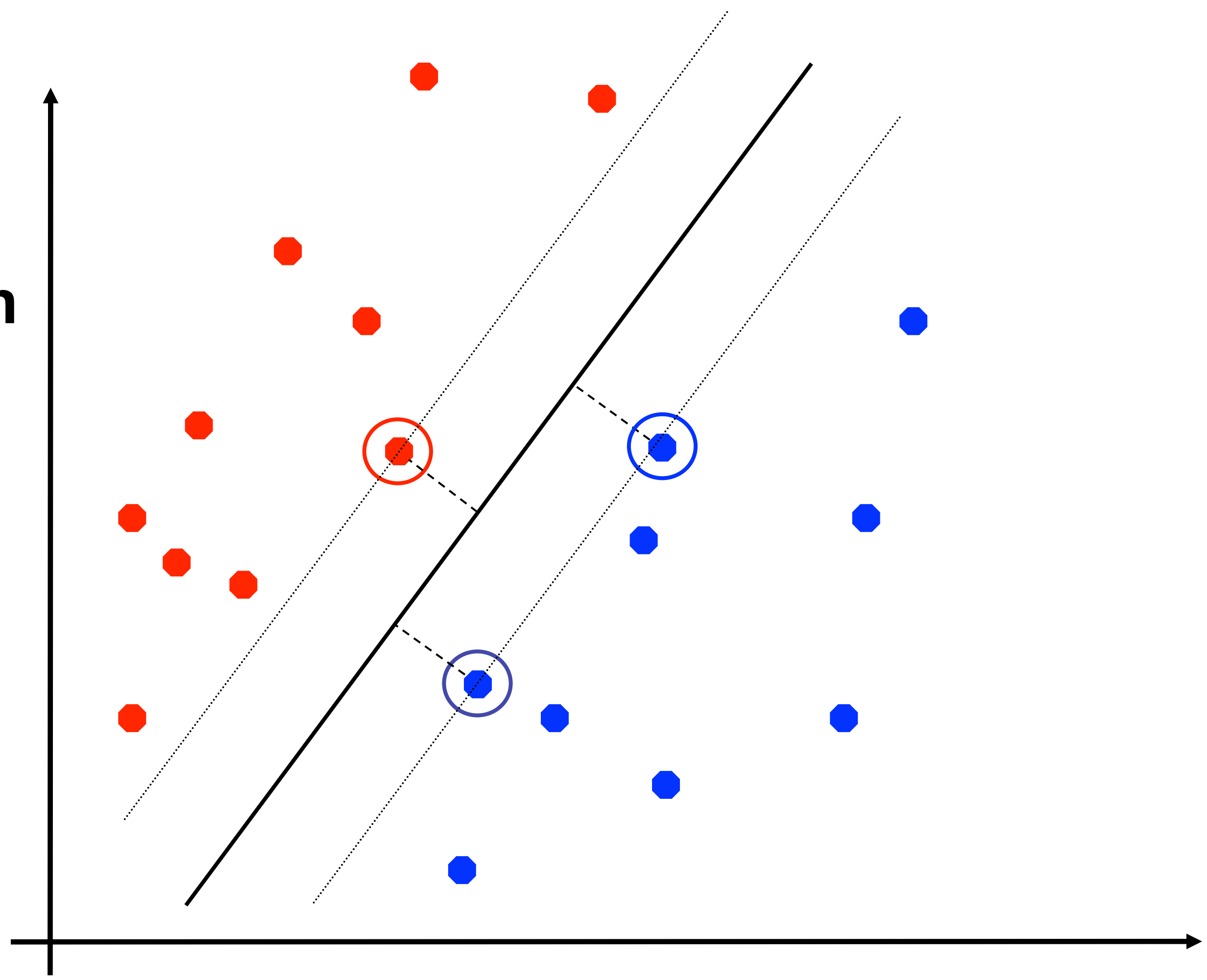
$y^{(i)}(\mathbf{w} \bullet \mathbf{x}^{(i)} + b) \geq 1$ for all $i = 1,2,\ldots,m$

Then $\mathbf{w} \bullet \mathbf{x} + b = 0$ is a hyperplane that provides the maximum margin.

Maximize Margin Optimization Problem

For data point $\mathbf{x}^{(i)}$ let $\delta^{(i)}$ be the distance to the plane. The **margin** is $\delta := \min \delta^{(i)}$

We want to find a hyperplane (ie, **w** and b) that maximizes δ. To this end, let's calculate $\delta^{(i)}$



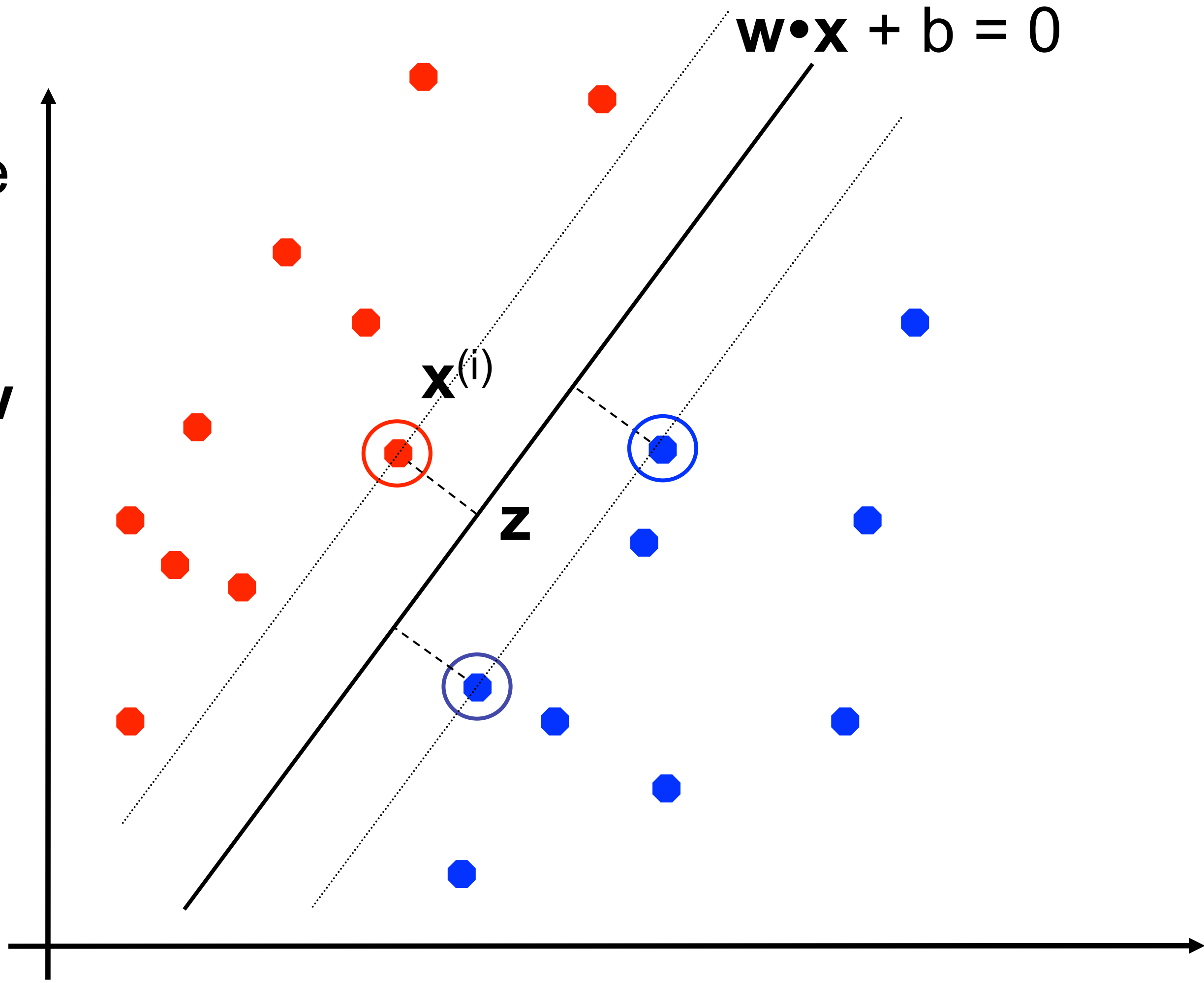
How can we calculate $\delta^{(i)}$?

Quiz: find $\delta^{(i)}$ in terms of **w**, b, and $\mathbf{x}^{(i)}$

- Let **z** be point on plane nearest to $\mathbf{x}^{(i)}$
- $\delta^{(i)} = |\mathbf{x}^{(i)} - \mathbf{z}|$
- (Assume first $\mathbf{x}^{(i)}$ and **w** on same side of plane.)
- Note **w** and $\mathbf{x}^{(i)} - \mathbf{z}$ point in same direction
- **$(\mathbf{x}^{(i)} - \mathbf{z}) / \delta^{(i)} = \mathbf{w}/||\mathbf{w}||$**

- Since **z** is on plane:

$\mathbf{w} \bullet \mathbf{z} + b = 0$



Complete quiz!

Answer:

$$\delta^{(i)} = y^{(i)}(\mathbf{w}\bullet\mathbf{x}^{(i)} + b)/|\mathbf{w}|$$

Thus if:

$$y^{(i)}(\mathbf{w}\bullet\mathbf{x}^{(i)} + b)/|\mathbf{w}| \geq \delta \quad \text{for all } i$$

then margin will be at least δ.

Margin Optimization Problem

Therefore, to maximize the margin, want to choose **w** and b to

maximize δ

subject to:

$$y^{(i)}(\mathbf{w} \bullet \mathbf{x}^{(i)} + b)/|\mathbf{w}| \geq \delta \quad \text{for all } i = 1,\ldots,m$$

Problem: Constraints are nonlinear. Difficult to solve.

Convert to Tractable Problem

Claim: If data is linearly separable, there exists $\mathbf{w}$, b such that:

- for all positive examples ($y^{(i)} = 1$):

  $\mathbf{w} \bullet \mathbf{x}^{(i)} + b \geq 1$

- for all negative examples ($y^{(i)} = -1$):

  $\mathbf{w} \bullet \mathbf{x}^{(i)} + b \leq -1$

Prove it now !

Thus, $y^{(i)} (\mathbf{w} \bullet \mathbf{x}^{(i)} + b) \geq 1$ for all $i = 1,2,\ldots,m$

Maximum Margin: Tractable

Theorem: Let $\mathbf{w}$, b be an optimal solution to:

Minimize $\|\mathbf{w}\|^2 = w_1^2 + w_2^2 + \ldots + w_n^2$

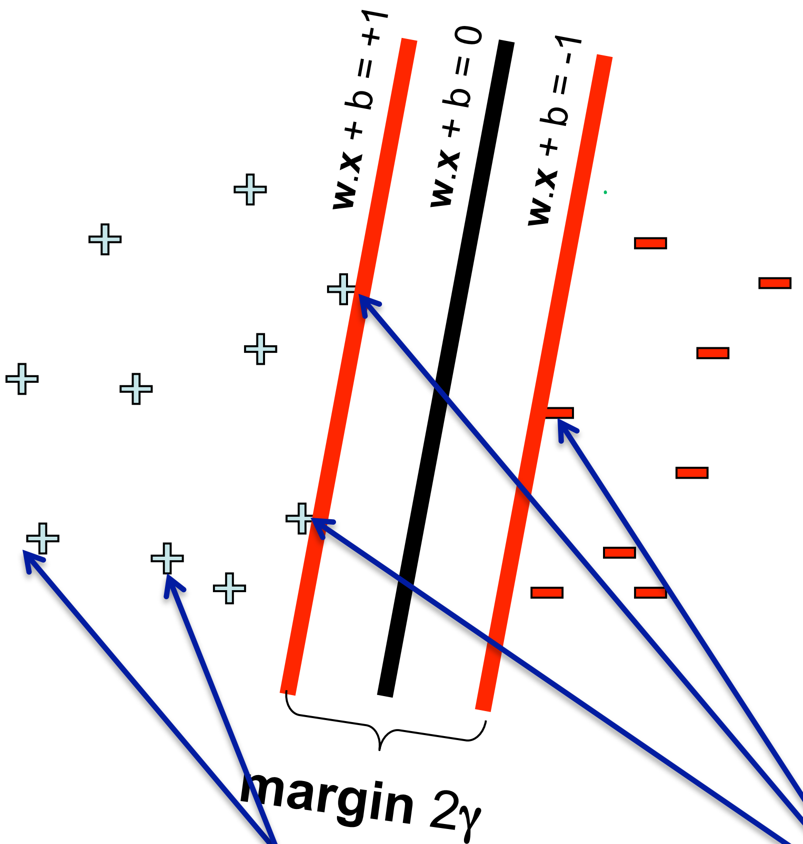
subject to:

$y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq 1$ for all $i = 1,2,\ldots,m$

Then $\mathbf{w} \cdot \mathbf{x} + b = 0$ is a hyperplane that provides the maximum margin.

Proof: You will prove this in HW.

(Hard margin) Support Vector Machines



Minimize $||\mathbf{w}||^2$
subject to:

$$y^{(i)}(\mathbf{w}\bullet\mathbf{x}^{(i)} + b) \geq 1 \text{ for all } i$$

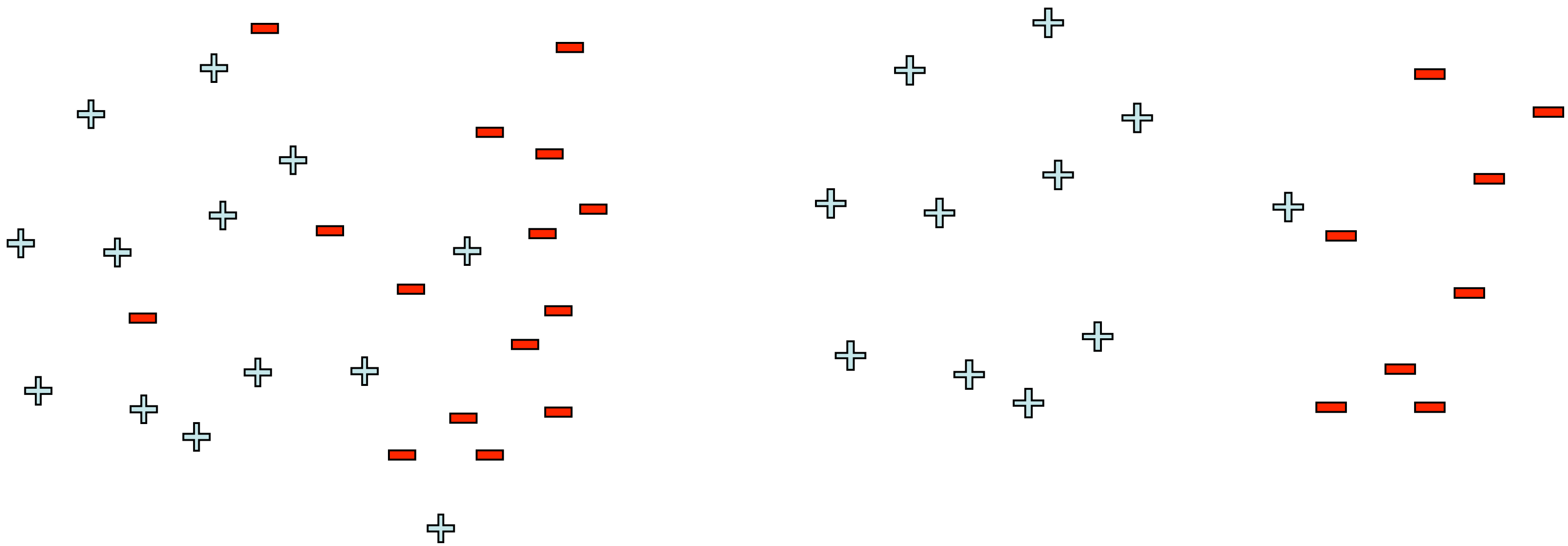
Example of a **convex optimization** problem

- A quadratic program
- Polynomial-time algorithms to solve!

Non-support Vectors:
- everything else
- moving them will not change **w**

Support Vectors:
- data points on the margin lines

But what if you have:

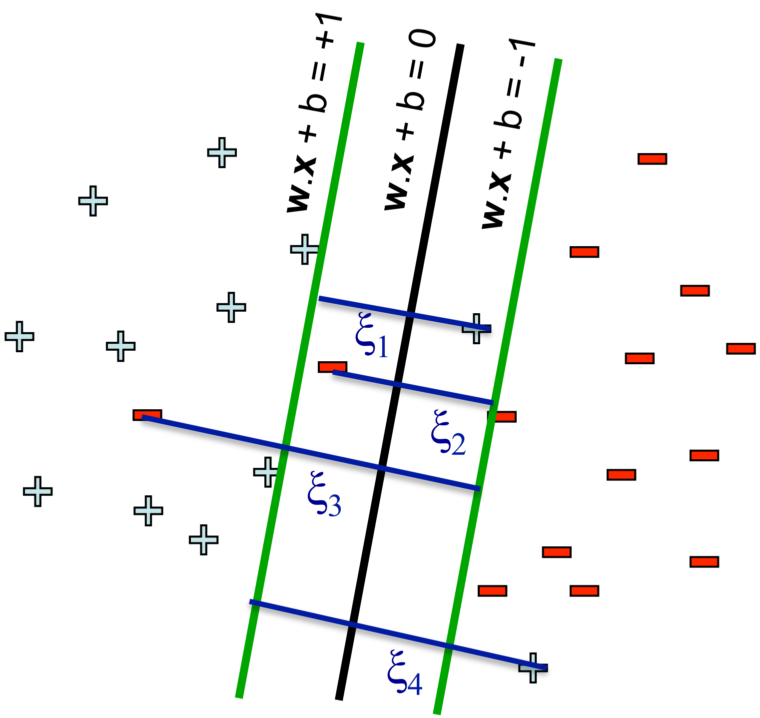


non-separable or separable

Natural Approach

- Allow for some of the data points to be within the margin.
- Tradeoff: As margin gets bigger, more data points inside
- Can penalize for the # of points inside: maximize: margin – C × (# points inside)
- But difficult to solve. Instead use distance from margin for violating points.

“Soft margin SVM”



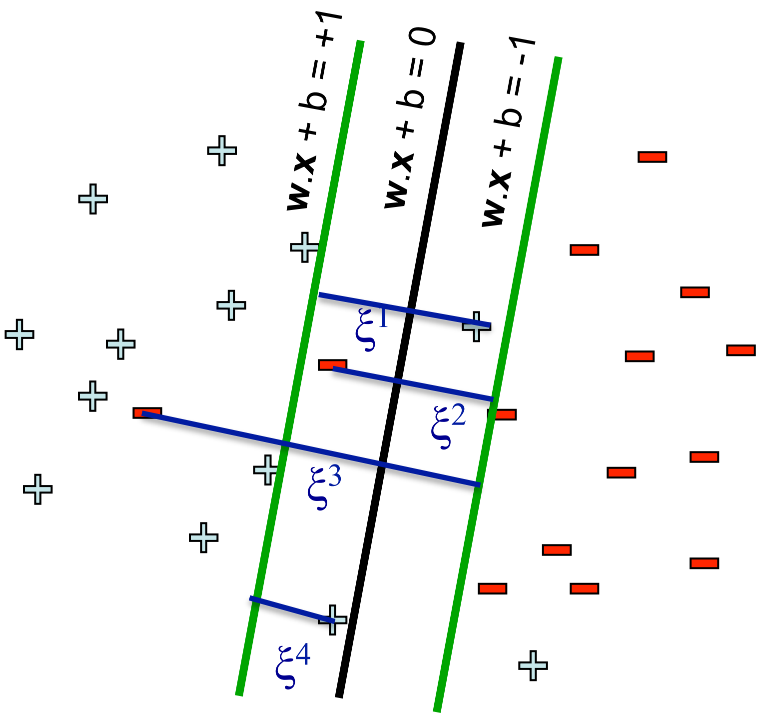
Minimize$_{\mathbf{w},b,\xi}$ $\mathbf{w} \bullet \mathbf{w} + C \sum_i \xi^{(i)}$
subject to
$(\mathbf{w} \bullet \mathbf{x}^{(i)} + b) y^{(i)} \geq 1 - \xi^{(i)}$ for all i

$\xi^{(i)} \geq 0$ for all i

Slack penalty $C > 0$:

- Want to find **w**, b so that the the margin is large and the # of errors is small.
- Want large margin to prevent overfitting.
- Solve optimization problem for different values of C. Choose the C that gives the smallest validation error.

Can Remove the Constraints



Minimize$_{\mathbf{w},b,\xi}$ $\mathbf{w \bullet w} + C \sum_i \xi^{(i)}$
subject to
$(\mathbf{w \bullet x}^{(i)}+b)y^{(i)} \geq 1 - \xi^{(i)}$ for all i
$\xi^{(i)} \geq 0$ for all i

At optimal solution, for each data point:

- If $(\mathbf{w \bullet x}^{(i)}+b)y^{(i)} \geq 1$, $\xi^{(i)} = 0$
- If $(\mathbf{w \bullet x}^{(i)}+b)y^{(i)} < 1$, $\xi^{(i)} = 1 - (\mathbf{w \bullet x}^{(i)}+b)y^{(i)}$ (constraint binding)

So $\xi^{(i)} = \max(0, 1 - (\mathbf{w \bullet x}^{(i)}+b)y^{(i)})$

Summary: SVM Soft-Margin Optimization Problem

$\text{Minimize}_{\mathbf{w},b,\xi}\ \mathbf{w}\bullet\mathbf{w} + C\ \Sigma_i\ \xi^{(i)}$
subject to
$(\mathbf{w}\bullet\mathbf{x}^{(i)}+b)y^{(i)} \geq 1 - \xi^{(i)}$ for all i, $\xi^{(i)} \geq 0$ for all i

Or equivalently, the non-linear unconstrained problem:

$$\text{Minimize}_{\mathbf{w},b}\ \mathbf{w}\bullet\mathbf{w} + C\ \Sigma_i \max(\ 0, 1 - y^{(i)}\,(\mathbf{w}\bullet\mathbf{x}^{(i)}+b))$$

regularization

Cross validation

- Divide labeled examples into 10 parts.
  - For each part
    - Use that part for validation, other 9 parts for training (optimization problem). Obtain validation error.
  - Calculate the average validation error over the 10 parts.
- Do for each C.
- Choose the C that has the lowest average validation error.