# LARGE MARGIN CLASSIFIERS

David Kauchak
CS 158 – Fall 2019

## Admin

Assignment 5
- Experiments

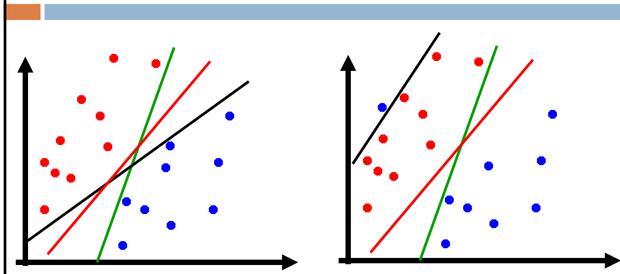
Assignment 6

**Next class: Meet in Edmunds 105**

Midterm

Course feedback
- Thanks!
- We'll go over it at the beginning of next class

## Which hyperplane?



Two main variations in linear classifiers:
- which hyperplane they choose when the data is linearly separable
- how they handle data that is not linearly separable

## Linear approaches so far

Perceptron:
- separable:
- non-separable:

Gradient descent:
- separable:
- non-separable:

## Linear approaches so far

Perceptron:

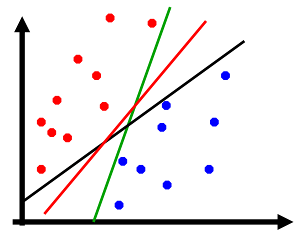
- separable:
  - finds ***some*** hyperplane that separates the data
- non-separable:
  - will continue to adjust as it iterates through the examples
  - final hyperplane will depend on which examples it saw recently

Gradient descent:

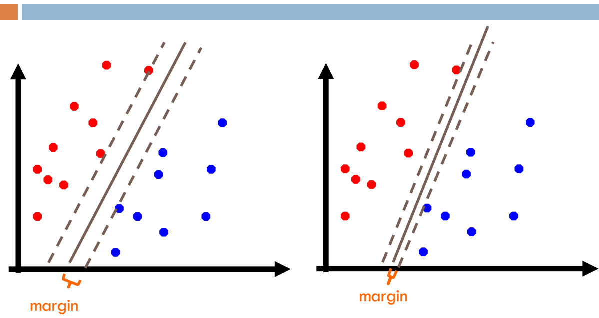
- separable and non-separable
  - finds the hyperplane that minimizes the objective function (loss + regularization)

Which hyperplane is this?

## Which hyperplane would you choose?

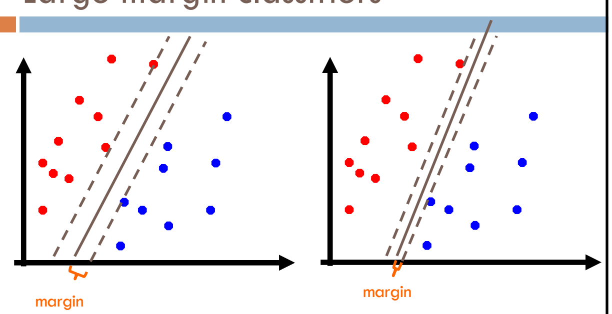


## Large margin classifiers




Choose the line where the distance to the nearest point(s) is as large as possible

## Large margin classifiers




The margin of a classifier is the distance to the closest points of either class

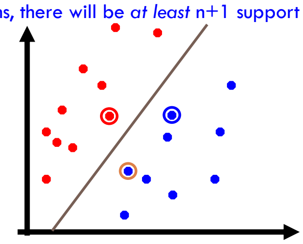
Large margin classifiers attempt to maximize this

## Support vectors

For any separating hyperplane, there exist some set of "closest points"

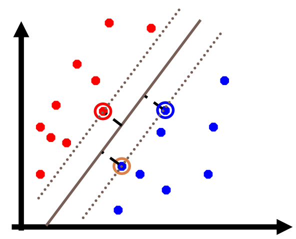
These are called the support vectors

For n dimensions, there will be *at least* n+1 support vectors

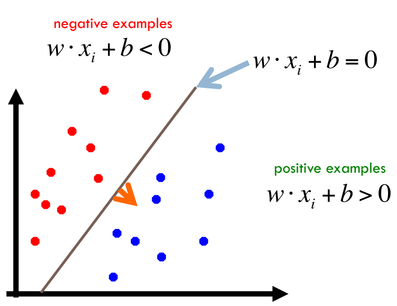


## Measuring the margin

The margin is the distance to the support vectors, i.e. the "closest points", on either side of the hyperplane

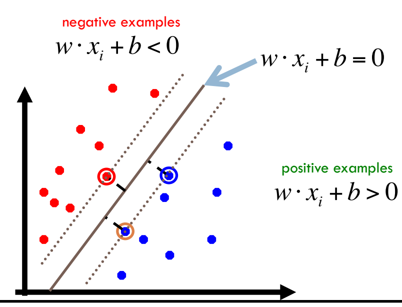


## Measuring the margin




## Measuring the margin

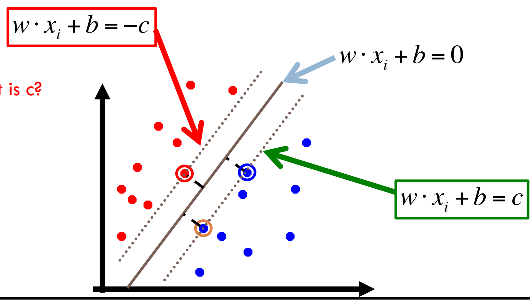
What are the equations for the margin lines?



## Measuring the margin

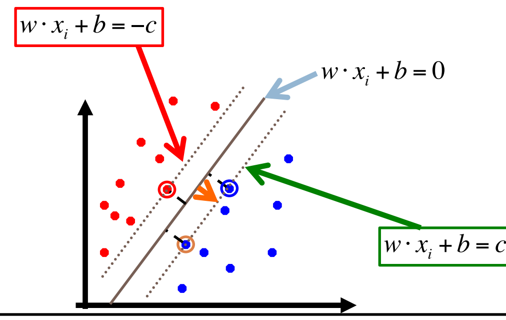
We know they're the same distance apart (otherwise, they wouldn't be support vectors!)

What is c?




## Measuring the margin

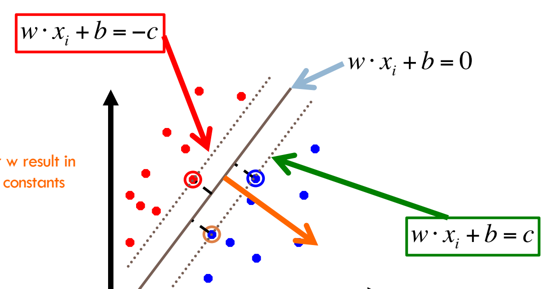
Depends! If we scale w, we vary the constant without changing the separating hyperplane




## Measuring the margin

Depends! If we scale w, we vary the constant without changing the separating hyperplane

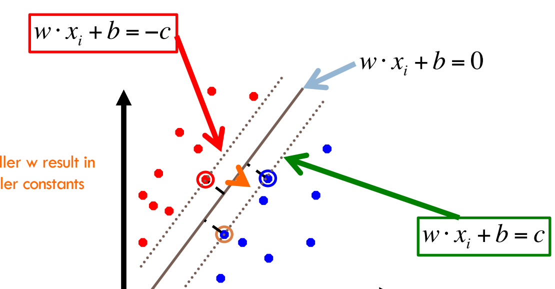
Larger w result in larger constants




## Measuring the margin

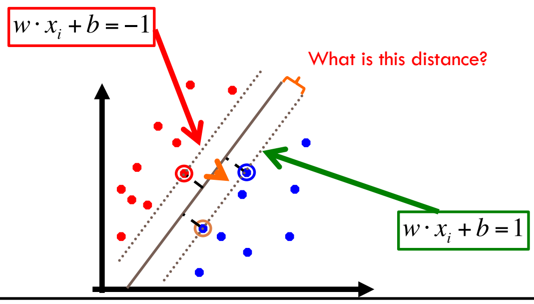
Depends! If we scale w, we vary the constant without changing the separating hyperplane

Smaller w result in smaller constants



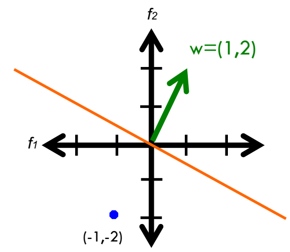
## Measuring the margin

For now, let's assume c = 1.



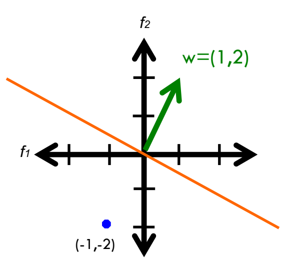

## Distance from the hyperplane

How far away is this point from the hyperplane?




## Distance from the hyperplane

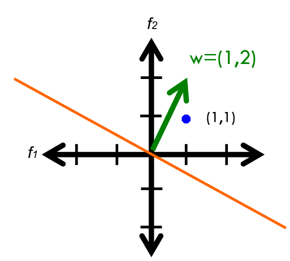
How far away is this point from the hyperplane?




$$d = \sqrt{(-1)^2 + (-2)^2} = \sqrt{5}$$

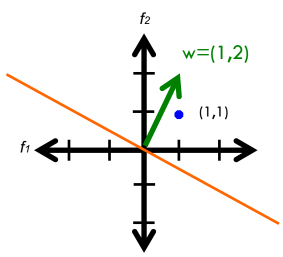
## Distance from the hyperplane

How far away is this point from the hyperplane?



## Distance from the hyperplane

How far away is this point from the hyperplane?

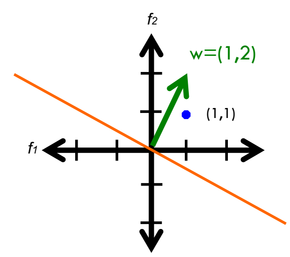



Is it?

$$d(x) = w \cdot x + b$$

## Distance from the hyperplane

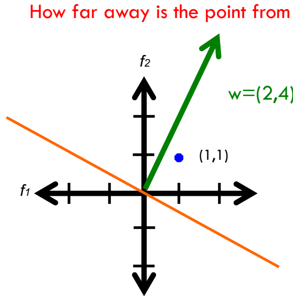
Does that seem right? What's the problem?




$$d(x) = w \cdot x + b$$
$$= w_1 x_1 + w_2 x_2 + b$$
$$= 1*1 + 1*2 + 0$$
$$\boxed{= 3?}$$

## Distance from the hyperplane

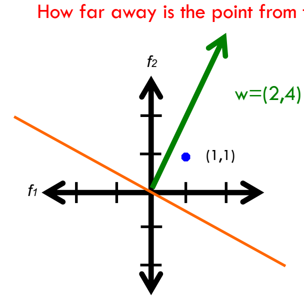
How far away is the point from the hyperplane?




$$d(x) = w \cdot x + b$$

## Distance from the hyperplane

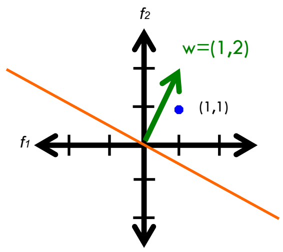
How far away is the point from the hyperplane?




$$d(x) = w \cdot x + b$$
$$= w_1 x_1 + w_2 x_2 + b$$
$$= 2*1 + 4*2 + 0$$
$$\boxed{= 10?}$$

## Distance from the hyperplane

How far away is this point from the hyperplane?

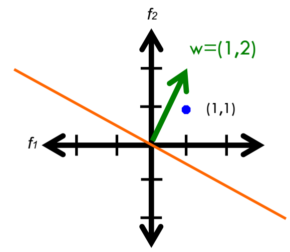



$$d(x) = \frac{w \cdot x + b}{\|w\|}$$

length normalized weight vectors

## Distance from the hyperplane

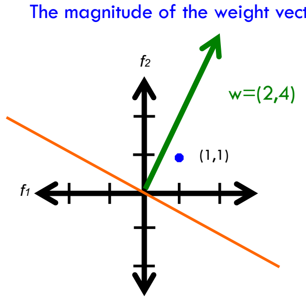
How far away is this point from the hyperplane?




$$d(x) = \frac{w \cdot x + b}{\|w\|}$$
$$= \frac{(w_1 x_1 + w_2 x_2) + b}{\sqrt{5}}$$
$$= \frac{(1*1 + 1*2) + 0}{\sqrt{5}}$$
$$= 1.34$$

## Distance from the hyperplane

The magnitude of the weight vector doesn't matter

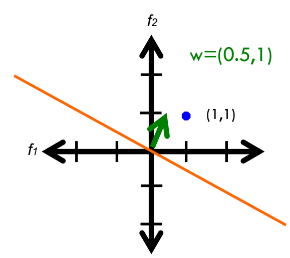



$$d(x) = \frac{w \cdot x + b}{\|w\|}$$

length normalized weight vectors

## Distance from the hyperplane

The magnitude of the weight vector doesn't matter

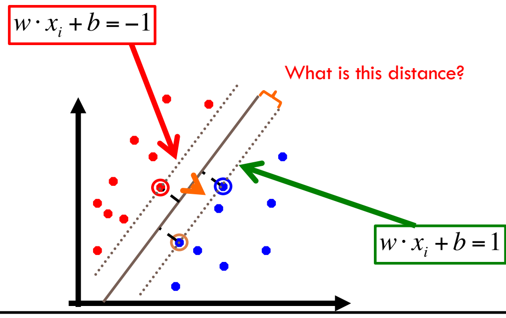



$$d(x) = \frac{w \cdot x + b}{\|w\|}$$

length normalized weight vectors

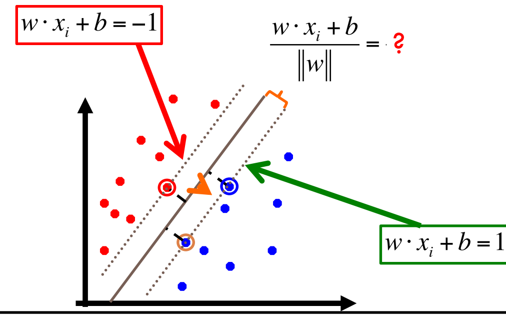
## Measuring the margin

For now, let's just assume c = 1.



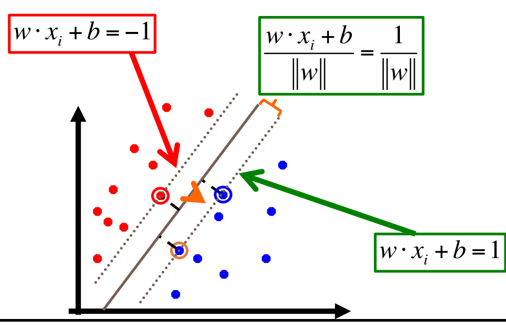

## Measuring the margin

For now, let's just assume c = 1.




## Measuring the margin

For now, let's just assume c = 1.




## Large margin classifier setup

Select the hyperplane with the largest margin where the points are classified correctly *and outside the margin!*

Setup as a constrained optimization problem:

$$\max_{w,b} \quad \text{margin}(w,b)$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 \quad \forall i$$

what does this say?

## Large margin classifier setup

Select the hyperplane with the largest margin where the points are classified correctly *and outside the margin!*

Setup as a constrained optimization problem:

$$\max_{w,b} \frac{1}{\|w\|}$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

## Maximizing the margin

$$\min_{w,b} \|w\|$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

**Maximizing the margin is equivalent to minimizing $\|w\|$!**
**(subject to the separating constraints)**

## Maximizing the margin

The minimization criterion wants w to be as small as possible

$$\min_{w,b} \|w\|$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

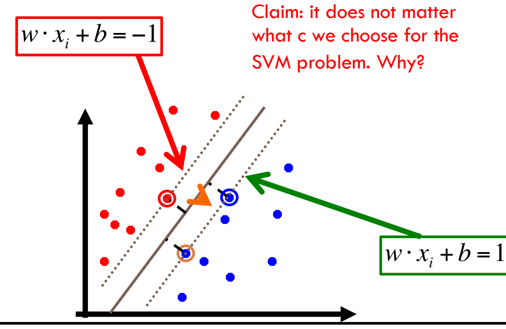
The constraints:

1. make sure the data is separable
2. encourages w to be larger (once the data is separable)

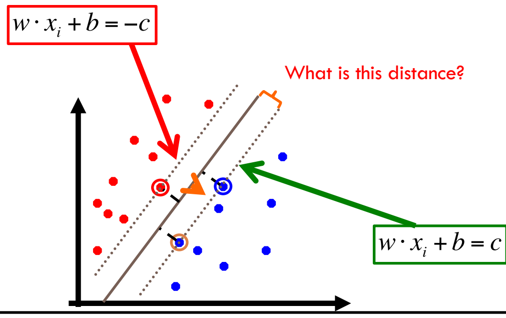
## Measuring the margin

For now, let's just assume c = 1.

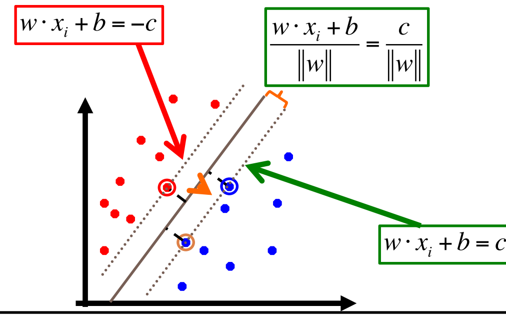
Claim: it does not matter what c we choose for the SVM problem. Why?



## Measuring the margin




## Measuring the margin




## Maximizing the margin

$$\min_{w,b} \frac{\|w\|}{c}$$

subject to:

$$y_i(w \cdot x_i + b) \geq c \quad \forall i$$

vs.

What's the difference?

$$\min_{w,b} \|w\|$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

## Maximizing the margin

$$\min_{w,b} \frac{\|w\|}{c}$$

subject to:

$$y_i(w \cdot x_i + b) \geq c \quad \forall i$$

Learn the exact same hyperplane just scaled by a constant amount

vs.

$$\min_{w,b} \|w\|$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

Because of this, often see it with c = 1

## For those that are curious…

$$\frac{\|w\|}{c} = \frac{\sqrt{w_1^2 + w_2^2 + \ldots + w_m^2 + b^2}}{c}$$
$$= \sqrt{\left(\frac{\sqrt{w_1^2 + w_2^2 + \ldots + w_m^2}}{c}\right)^2}$$
$$= \sqrt{\frac{w_1^2 + w_2^2 + \ldots + w_m^2}{c^2}}$$
$$= \sqrt{\frac{w_1^2}{c^2} + \frac{w_2^2}{c^2} + \ldots + \frac{w_m^2}{c^2}}$$
$$= \sqrt{\left(\frac{w_1}{c}\right)^2 + \left(\frac{w_2}{c}\right)^2 + \ldots + \left(\frac{w_m}{c}\right)^2}$$

scaled version of w

## Maximizing the margin: the real problem

$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

Why the squared?

## Maximizing the margin: the real problem

$$\min_{w,b} \quad \|w\| = \sqrt{\sum_i w_i^{\,2}}$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

$$\min_{w,b} \quad \|w\|^2 = \sum_i w_i^{\,2}$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

Minimizing $||\boldsymbol{w}||$ is equivalent to minimizing $||\boldsymbol{w}||^2$

The sum of the squared weights is a convex function!

## Support vector machine problem

$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

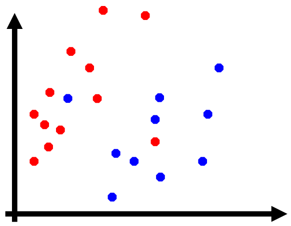
This is a version of a quadratic optimization problem

Maximize/minimize a quadratic function

Subject to a set of linear constraints

Many, many variants of solving this problem (we'll see one in a bit)

## Soft Margin Classification



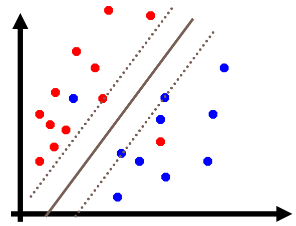
$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 \quad \forall i$$

What about this problem?

## Soft Margin Classification



$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 \quad \forall i$$

We'd like to learn something like this, but our constraints won't allow it ☹

## Slack variables

$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 \quad \forall i$$

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

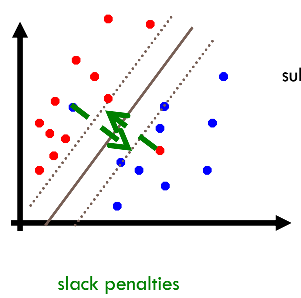
$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \ge 0$$

slack variables (one for each example)

What effect does this have?

## Slack variables




$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \ge 0$$

## Slack variables

margin

trade-off between margin maximization and penalization

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

penalized by how far from "correct"

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

allowed to make a mistake

$$\varsigma_i \ge 0$$

## Soft margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

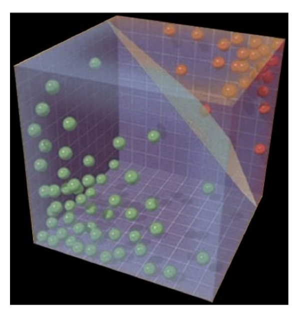
$$\varsigma_i \ge 0$$

Still a quadratic optimization problem!

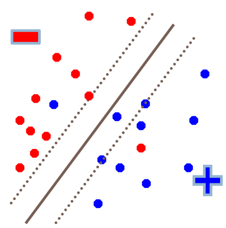
## Demo

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

## Solving the SVM problem



## Understanding the Soft Margin SVM



$$\min_{w,b} \|w\|^2 + C\sum_i \varsigma_i$$

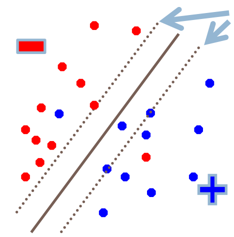
subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

Given the optimal solution, w, b:

Can we figure out what the slack penalties are for each point?

## Understanding the Soft Margin SVM



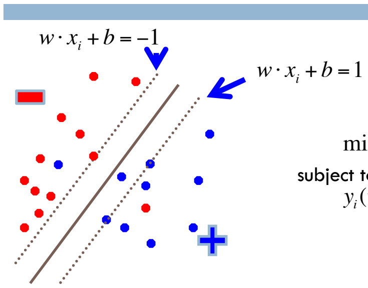
What do the margin lines represent wrt w,b?

$$\min_{w,b} \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

## Understanding the Soft Margin SVM



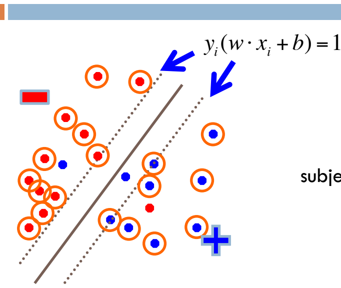

$$\min_{w,b} \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

Or: $y_i(w \cdot x_i + b) = 1$

## Understanding the Soft Margin SVM



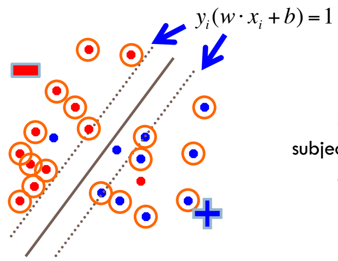

$$\min_{w,b} \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

What are the slack values for points outside (or on) the margin AND correctly classified?

## Understanding the Soft Margin SVM



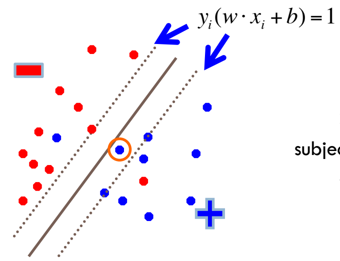

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

0! The slack variables have to be greater than or equal to zero and if they're on or beyond the margin then $y_i(wx_i+b) \ge 1$ already

## Understanding the Soft Margin SVM



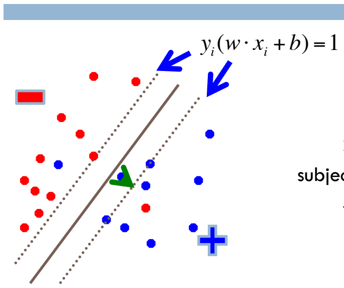

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

What are the slack values for points inside the margin AND classified correctly?

## Understanding the Soft Margin SVM




$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

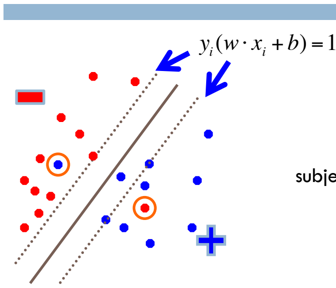
subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

Difference from the point to the margin. Which is?

$$\varsigma_i = 1 - y_i(w \cdot x_i + b)$$

## Understanding the Soft Margin SVM



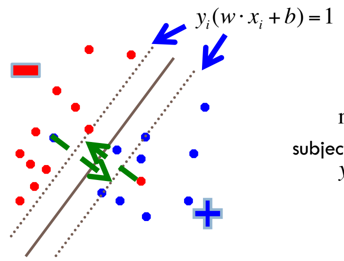

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

What are the slack values for points that are incorrectly classified?

## Understanding the Soft Margin SVM




$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

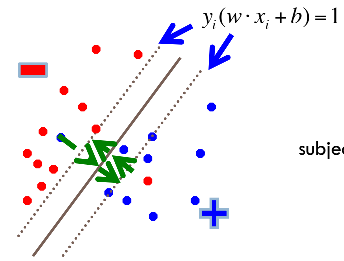
subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \ge 0$$

Which is?

## Understanding the Soft Margin SVM




$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

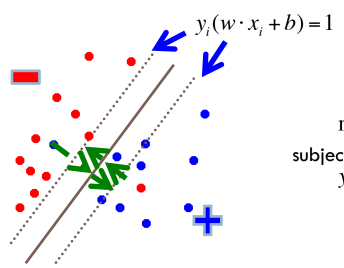
$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \ge 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

?

## Understanding the Soft Margin SVM




$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

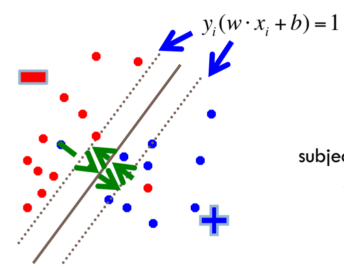
$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \ge 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$-y_i(w \cdot x_i + b)$ Why -?

## Understanding the Soft Margin SVM




$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

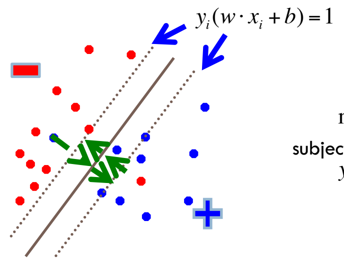
$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \ge 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$-y_i(w \cdot x_i + b)$ ?

## Understanding the Soft Margin SVM




$$\min_{w,b} \ \|w\|^2 + C \sum_i \varsigma_i$$

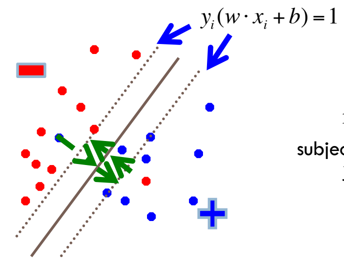
subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$-y_i(w \cdot x_i + b)$ 1

## Understanding the Soft Margin SVM




$$\min_{w,b} \ \|w\|^2 + C \sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$$\varsigma_i = 1 - y_i(w \cdot x_i + b)$$

## Understanding the Soft Margin SVM

$$\min_{w,b} \ \|w\|^2 + C \sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

---

$$\varsigma_i = \begin{cases} 0 & \text{if } y_i(w \cdot x_i + b) \ge 1 \\ 1 - y_i(w \cdot x_i + b) & \text{otherwise} \end{cases}$$

## Understanding the Soft Margin SVM

$$\varsigma_i = \begin{cases} 0 & \text{if } y_i(w \cdot x_i + b) \ge 1 \\ 1 - y_i(w \cdot x_i + b) & \text{otherwise} \end{cases}$$

$$\varsigma_i = \max(0, 1 - y_i(w \cdot x_i + b))$$
$$= \max(0, 1 - yy')$$

Does this look familiar?

## Hinge loss!

0/1 loss: $l(y,y') = 1[yy' \le 0]$

Hinge: $l(y,y') = \max(0, 1 - yy')$

Exponential: $l(y,y') = \exp(-yy')$

Squared loss: $l(y,y') = (y - y')^2$

## Understanding the Soft Margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

$$\varsigma_i = \max(0, 1 - y_i(w \cdot x_i + b))$$

Do we need the constraints still?

## Understanding the Soft Margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

$$\varsigma_i = \max(0, 1 - y_i(w \cdot x_i + b))$$

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \max(0, 1 - y_i(w \cdot x_i + b))$$

Unconstrained problem!

## Understanding the Soft Margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i loss_{hinge}(y_i, y_i')$$

Does this look like something we've seen before?

$$\operatorname{argmin}_{w,b} \sum_{i=1}^{n} loss(yy') + \lambda \; regularizer(w,b)$$

Gradient descent problem!

## Soft margin SVM as gradient descent

$$\min_{w,b} \quad \|w\|^2 + C\sum_i loss_{hinge}(y_i, y_i')$$

multiply through by 1/C and rearrange

$$\min_{w,b} \quad \sum_i loss_{hinge}(y_i, y_i') + \frac{1}{C}\|w\|^2$$

let λ=1/C

$$\min_{w,b} \quad \sum_i loss_{hinge}(y_i, y_i') + \lambda\|w\|^2$$

What type of gradient descent problem?

$$\text{argmin}_{w,b} \sum_{i=1}^{n} loss(yy') + \lambda \; regularizer(w,b)$$

## Soft margin SVM as gradient descent

One way to solve the soft margin SVM problem is using gradient descent

$$\min_{w,b} \quad \sum_i loss_{hinge}(y_i, y_i') + \lambda\|w\|^2$$

hinge loss

L2 regularization

## Gradient descent SVM solver

- pick a starting point (w)
- repeat until loss doesn't decrease in all dimensions:
  - pick a dimension
  - move a small amount in that dimension towards decreasing loss (using the derivative)

$$w_i = w_i - \eta \frac{d}{dw_i}(loss(w) + regularizer(w,b))$$

$$w_j = w_j + \eta \sum_{i=1}^{n} y_i x_i 1[y_i(w \cdot x + b) < 1] - \eta\lambda w_j$$

hinge loss      L2 regularization

Finds the largest margin hyperplane while allowing for a soft margin

## Support vector machines: 2013

One of the most successful (if not the most successful) classification approach:

| | 2013 | 2016 | 2019 |
|---|---|---|---|
| decision tree | About 2,160,000 results | About 2,480,000 | About 3,000,000 r |
| Support vector machine | About 1,960,000 results | About 2,430,000 | About 3,020,000 |
| k nearest neighbor | About 746,000 results | About 979,000 | About 1,380,000 |
| perceptron algorithm | About 84,300 results | About 104,000 | About 153,000 r |

Google scholar

## Trends over time