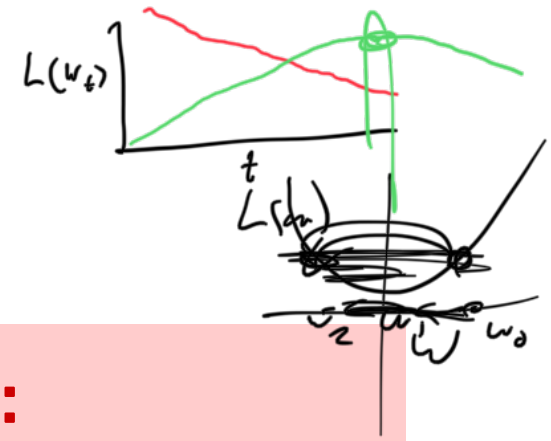
Train loss: $L(w) : \mathbb{R}^d \to \mathbb{R}$

$\nabla L(w) : \mathbb{R}^d \to \mathbb{R}^d$

$\nabla L(w) = \left(\frac{\partial}{\partial w_1} L(w), \frac{\partial}{\partial w_2} L(w), \ldots \right)^T$



Support vector machines: Maximum-margin linear classifiers

Grad Descent:

$w = (0, 0, \ldots, 0)^T$

for $t = 1 \ldots K$

$w = w - \eta_t \nabla L(w)$

Stoch. Gradient Descent:

$w = (0, 0, \ldots, 0)^T$

for $t = 1 \ldots K$

for $i = 1 \ldots N$

$w = w - \eta_t \nabla \ell_i(w)$

Stoch. Grad. Assump.

$$L(w) = \sum_{i=1}^{N} \ell_i(w)$$

Case: Logistic Regression

$$\ell_i(w) = -\log P_w(y^{(i)} \mid x^{(i)})$$

$$L(w) = \sum_{i=1}^{N} -\log P_w(y^{(i)} \mid x^{(i)})$$

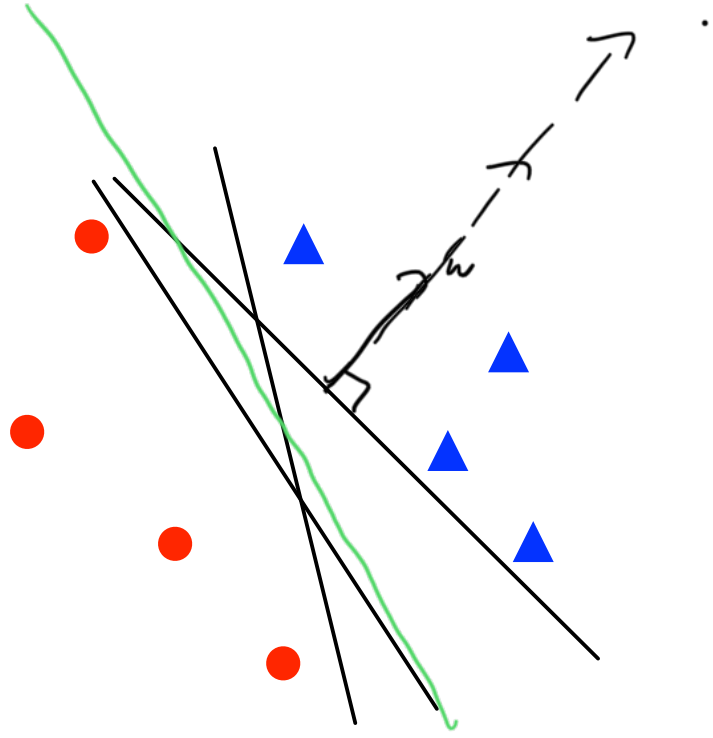
$$\nabla L(w) = \sum_{i=1}^{N} \nabla \ell_i(w)$$

Topics we'll cover

1. The margin of a linear classifier
2. Maximizing the margin
3. A convex optimization problem
4. Support vectors

Improving upon the Perceptron

For a linearly separable data set, there are in general many possible separating hyperplanes, and Perceptron is guaranteed to find one of them.



Is there a better, more systematic choice of separator?
The one with the most buffer around it, for instance?

The learning problem

Given: training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$.

Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y^{(i)}(w \cdot x^{(i)} + b) > 0$ for all i.

The learning problem

Given: training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$.

Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y^{(i)}(w \cdot x^{(i)} + b) > 0$ for all i.

By scaling w, b, can equivalently ask for

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

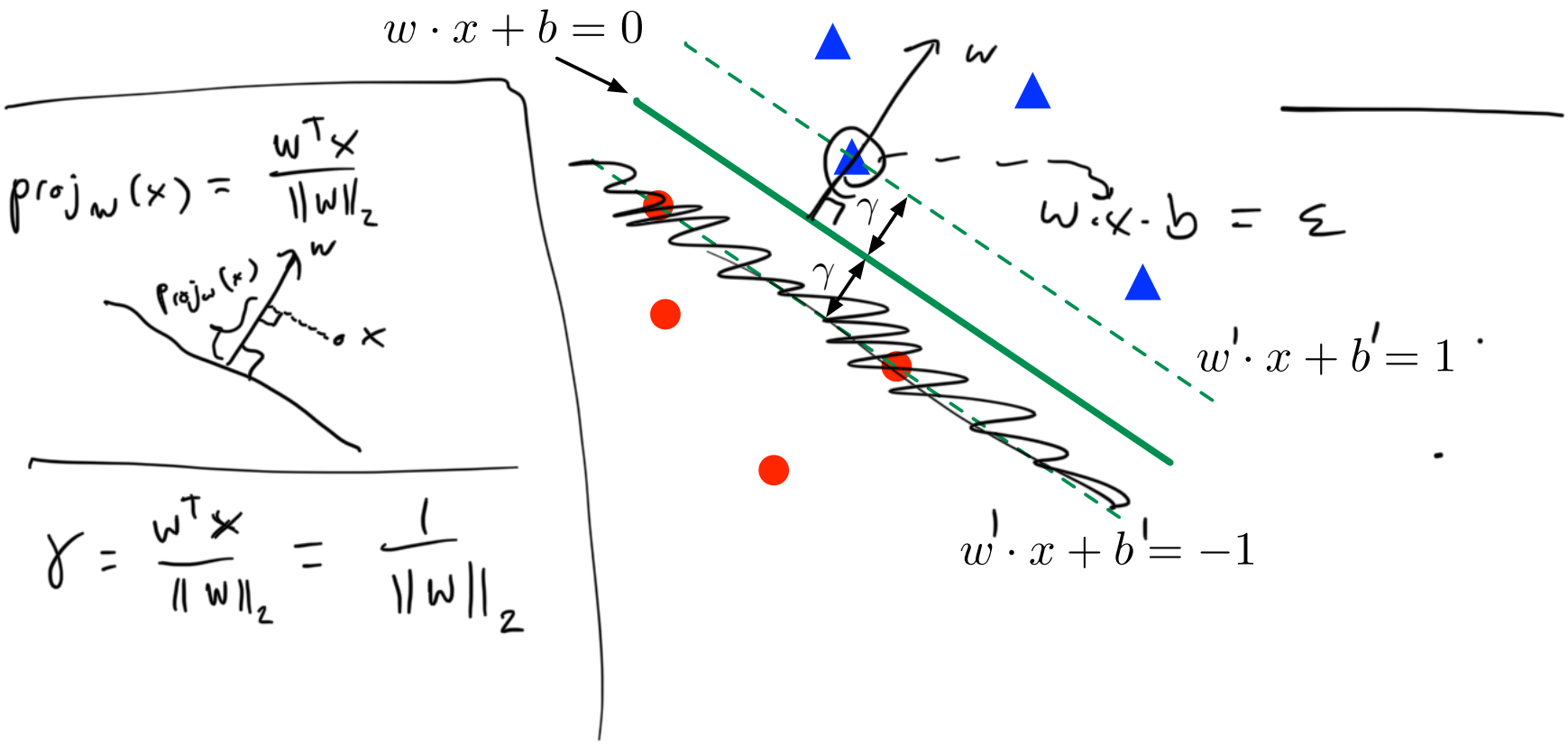
Maximizing the margin

Given: training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$.

Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y^{(i)}(w \cdot x^{(i)} + b) \geq \varepsilon \Rightarrow y^{(i)}\left(\frac{w}{\varepsilon} \cdot x^{(i)} + \frac{b}{\varepsilon}\right) \geq 1$$

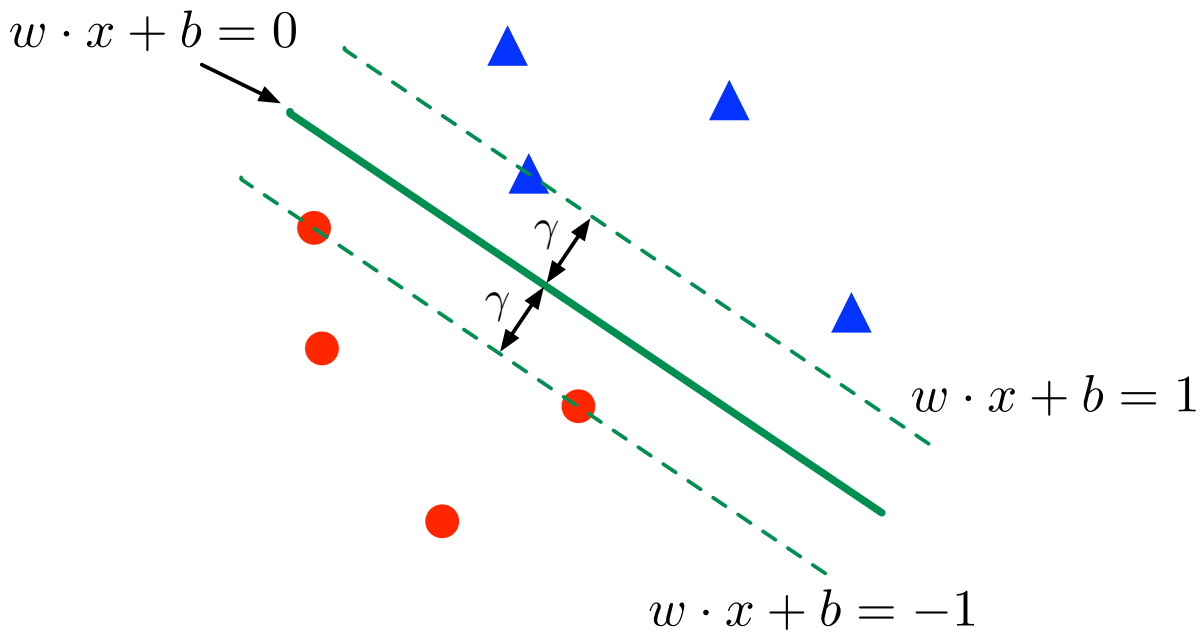
$$\Rightarrow \quad y^{(i)}(w' \cdot x^{(i)} + b') \geq 1 \quad \text{for all } i. \qquad \text{s.t. } w' = w/\varepsilon, \; b' = b/\varepsilon$$



Maximizing the margin

Given: training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$.
Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

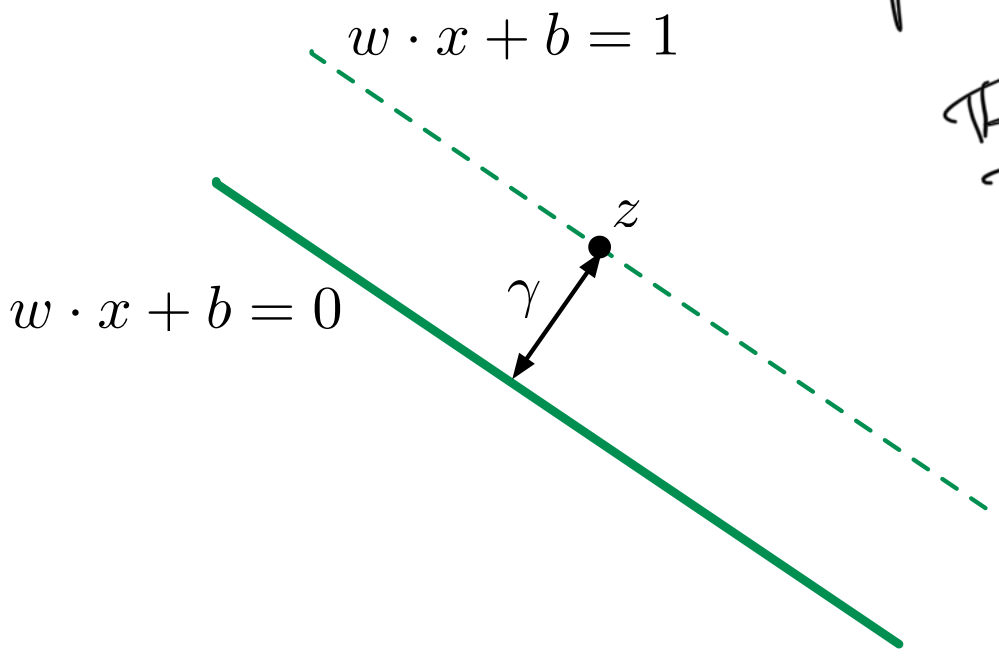
$$y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i.$$



Maximize the **margin** γ.

A formula for the margin

Close-up of a point z on the positive boundary.



A quick calculation shows that $\gamma = 1/\|w\|$.

In short: to maximize the margin, minimize $\|w\|$.

Maximum-margin linear classifier

- Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \|w\|^2$$

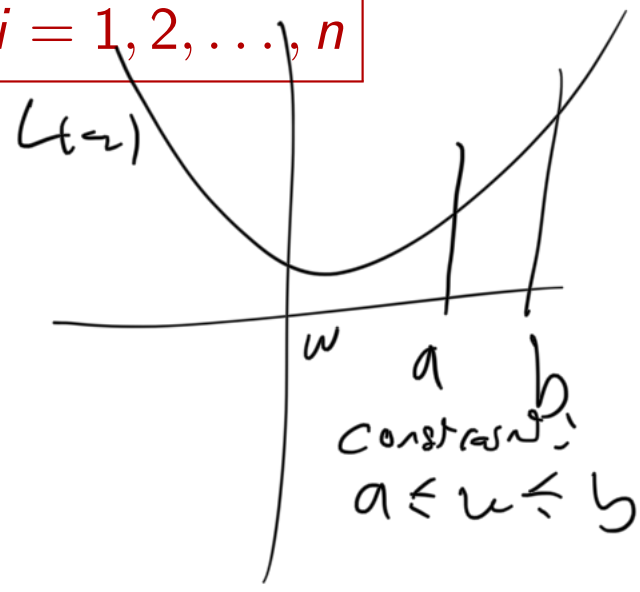
$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n$$

Maximum-margin linear classifier

- Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \|w\|^2$$

$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n$$



- This is a **convex optimization problem**:
  - Convex objective function
  - Linear constraints

Maximum-margin linear classifier

- Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \|w\|^2$$

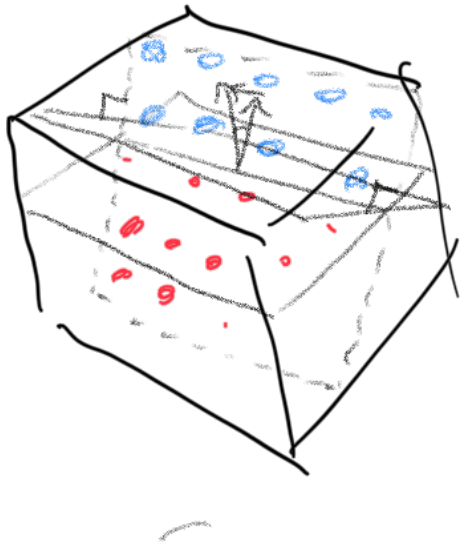
$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n$$

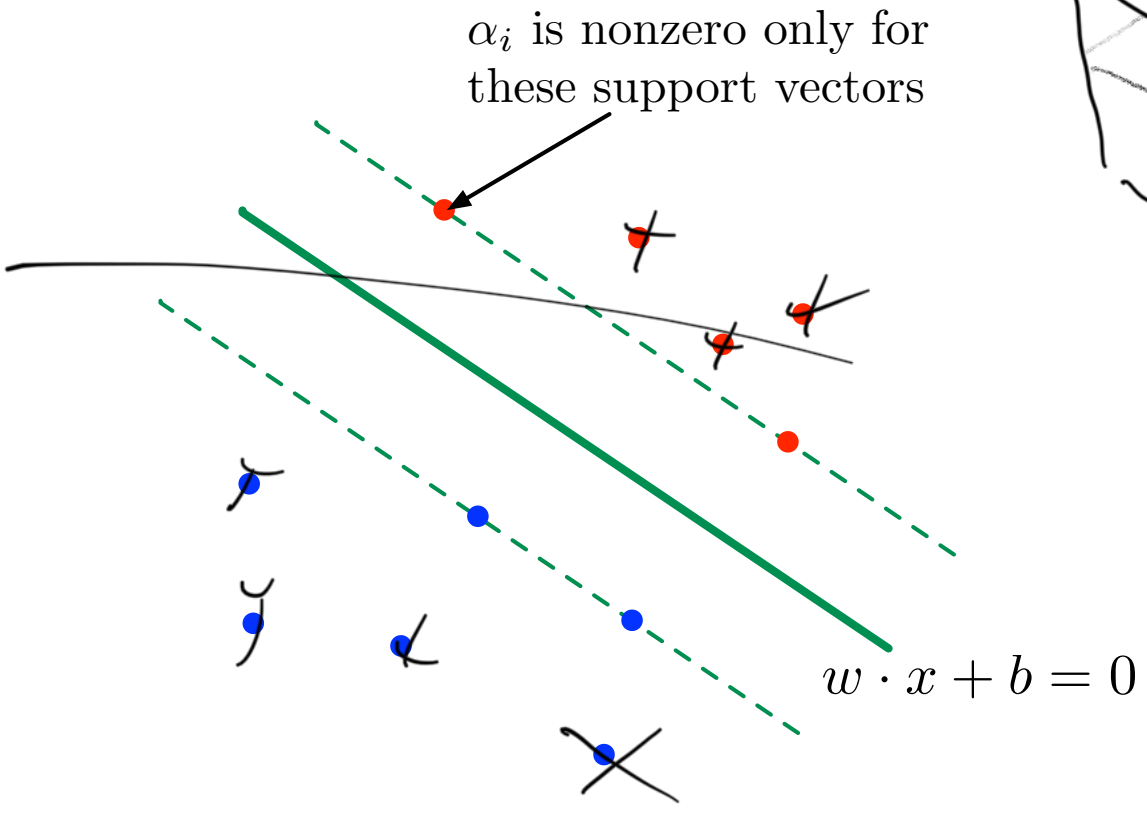
- This is a **convex optimization problem**:
  - Convex objective function
  - Linear constraints

- This means that:
  - the optimal solution can be found efficiently
  - **duality** gives us information about the solution

Support vectors

Support vectors: training points right on the margin, i.e. $y^{(i)}(w \cdot x^{(i)} + b) = 1$.





$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$ is a function of just the support vectors.

Small example: Iris data set

Fisher's **iris** data





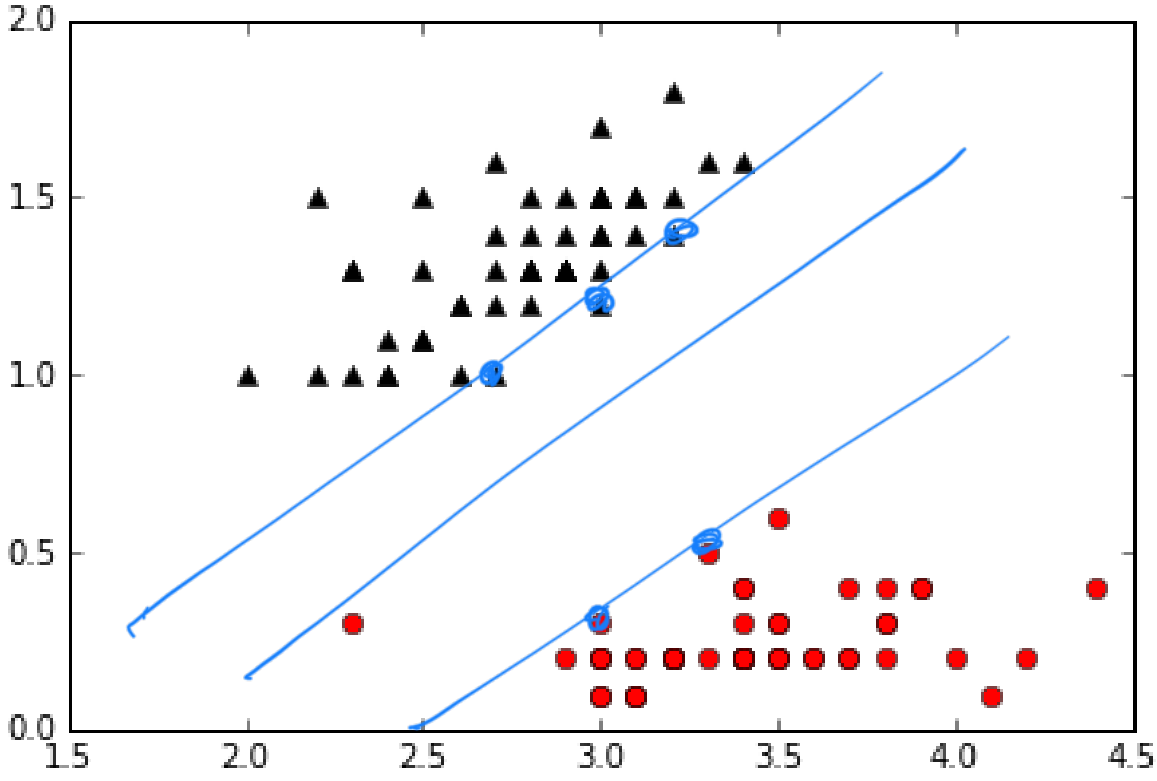
150 data points from three classes:

- iris setosa
- iris versicolor
- iris virginica

Four measurements: petal width/length, sepal width/length

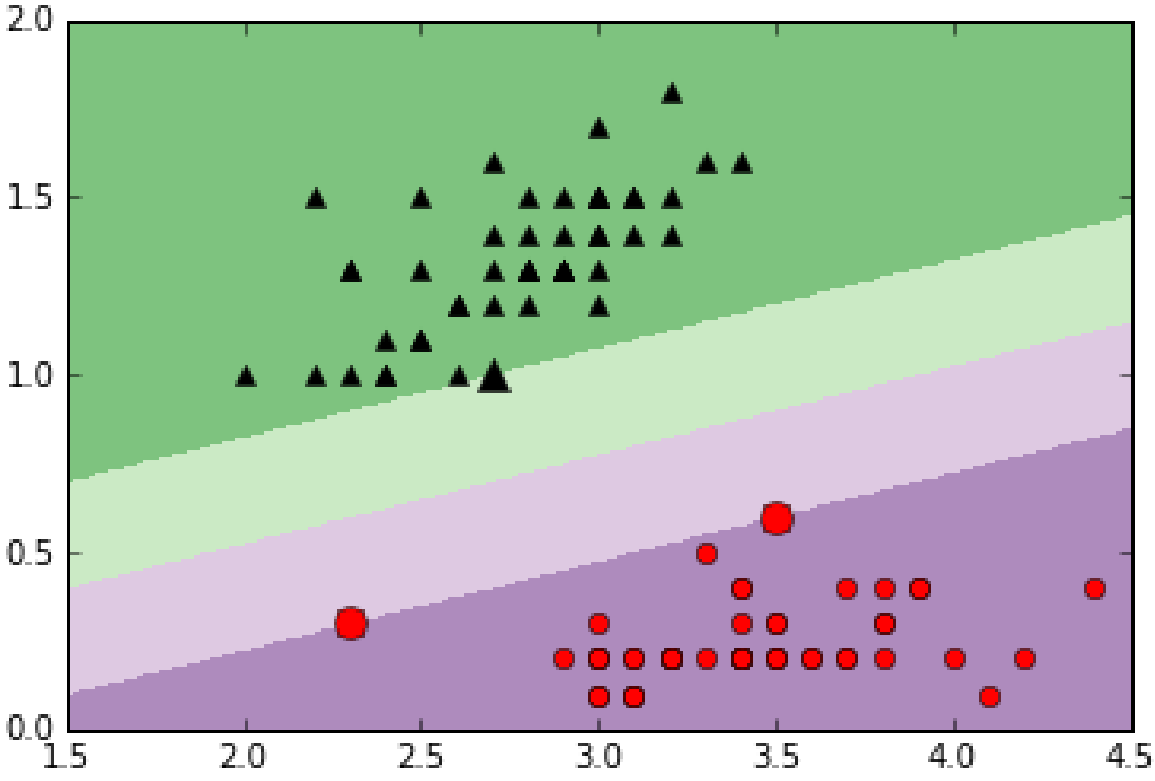
Small example: Iris data set

Two features: sepal width, petal width.
Two classes: setosa (red circles), versicolor (black triangles)



Small example: Iris data set

Two features: sepal width, petal width.
Two classes: setosa (red circles), versicolor (black triangles)



Soft Maximum-margin linear classifier

- Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \quad \|w\|^2 + C \sum_i \xi_i$$

$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n$$

$$\xi \geq 0$$

correct label

model score

Soft Maximum-margin linear classifier

- Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \quad \|w\|^2 + C \sum_i \xi_i$$

$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n$$

$$\xi \geq 0$$

- Allows for violation of constraints:
  - Model pays for violation via slack variables ξ_i
  - Works with non-separable data!