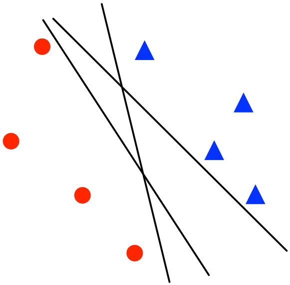
Support vector machines

CSE 250B

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.

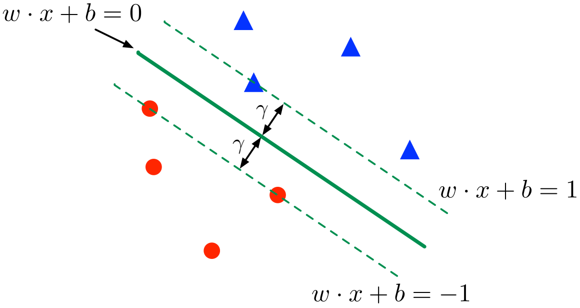
By scaling w, b, can equivalently ask for

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

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

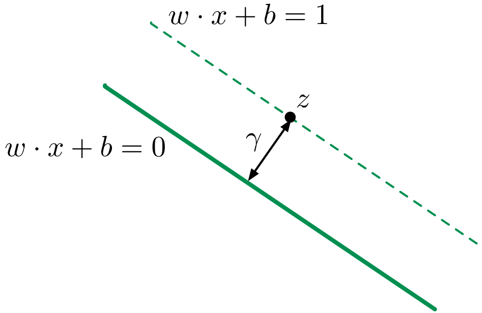
$$y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i.$$



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

$$\boxed{\begin{array}{l} \displaystyle\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \|w\|^2 \\ \text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n \end{array}}$$

Maximum-margin linear classifier

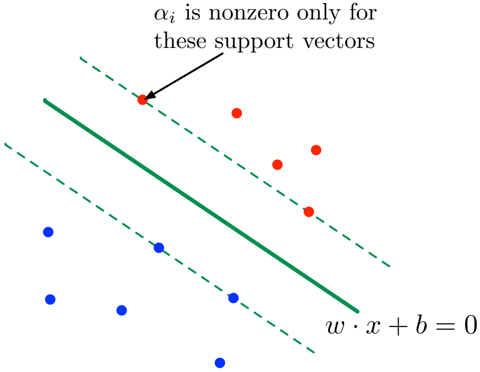
- Given $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\boxed{\begin{aligned} &\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \|w\|^2 \\ &\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \dots, n \end{aligned}}$$

- 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

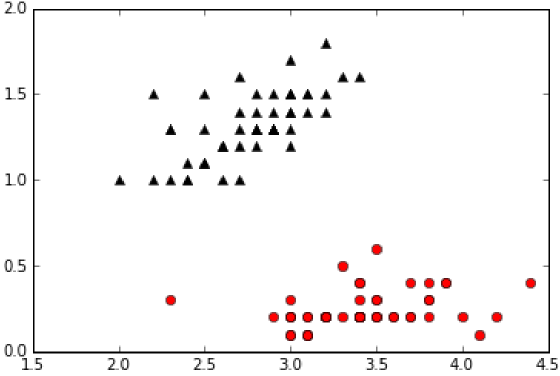
150 data points from three classes:

- iris setosa
- iris versicolor
- iris virginica

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

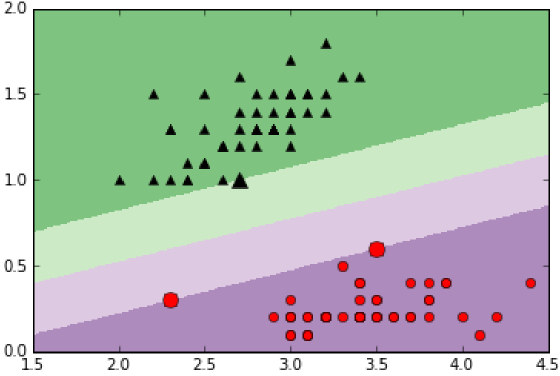
Small example: Iris data set

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



Small example: Iris data set

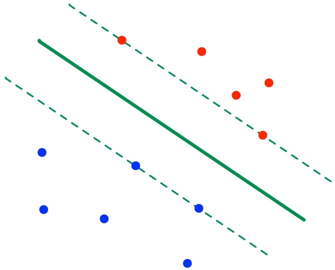
Two features: sepal width, petal width.
Two classes: setosa (red circles), versicolor (black triangles)



Recall: maximum-margin linear classifier

Given: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$.
Find: the linear separator w that perfectly classifies the data and has maximum margin.

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

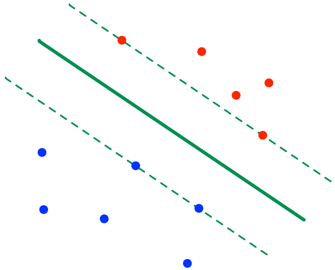


Solution $w = \sum_{i=1}^n \alpha_i y^{(i)} x^{(i)}$ is a function of just the support vectors.

Recall: maximum-margin linear classifier

Given: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$.
Find: the linear separator w that perfectly classifies the data and has maximum margin.

$$\boxed{\begin{array}{l} \displaystyle\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \|w\|^2 \\ \text{s.t.:} \;\; y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n \end{array}}$$



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

What if data is not separable?

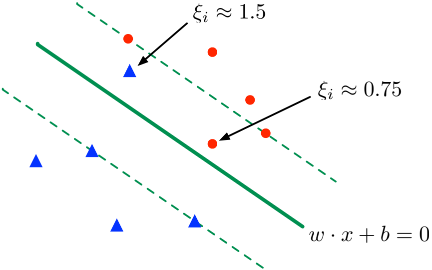
The non-separable case

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \quad \|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

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

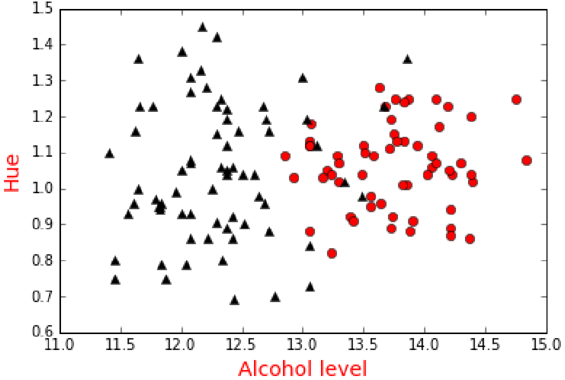
$$\xi \geq 0$$

Each data point $x^{(i)}$ is allowed some **slack** ξ_i.



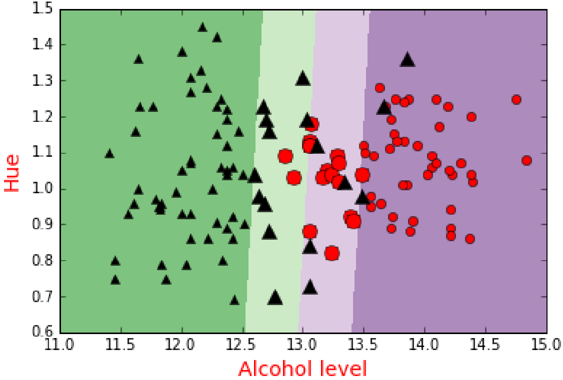
Wine data set

Here $C = 1.0$



Wine data set

Here $C = 1.0$

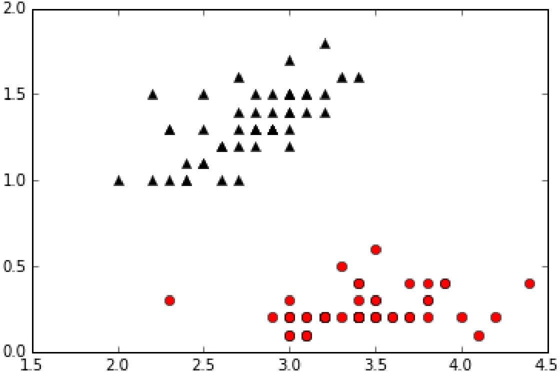


The tradeoff between margin and slack

$$\boxed{\begin{array}{c} \displaystyle\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \quad \|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.:} \quad y^{(i)}(w\cdot x^{(i)} + b) \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n \\ \xi \geq 0 \end{array}}$$

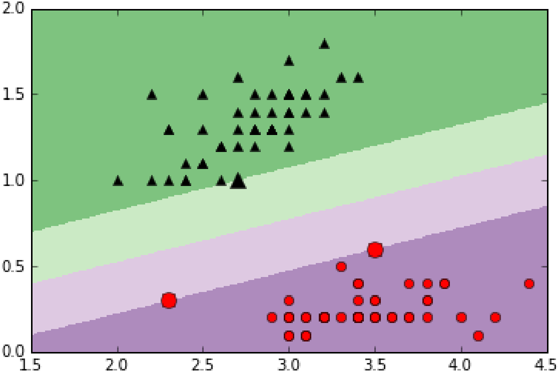
Back to Iris

$C = 10$



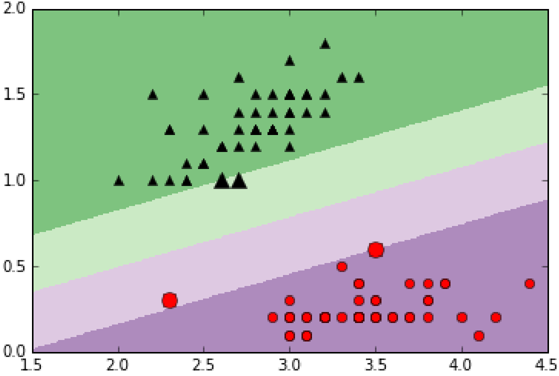
Back to Iris

$C = 10$



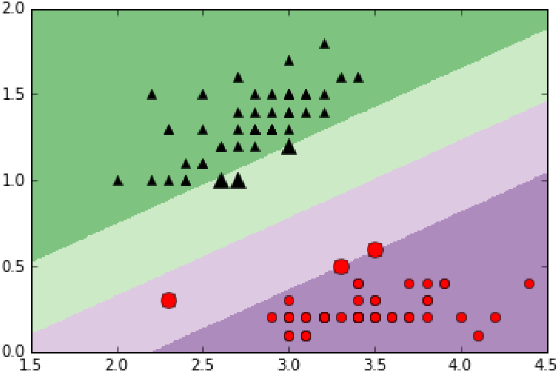
Back to Iris

$C = 3$



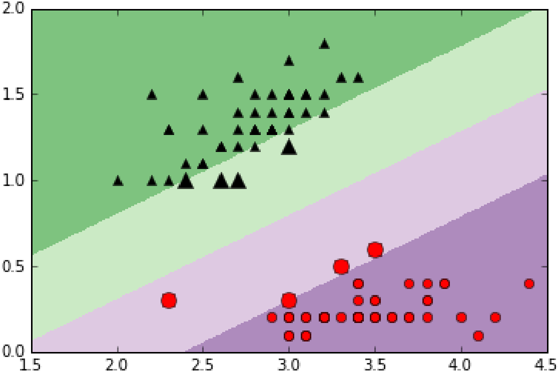
Back to Iris

$C = 2$



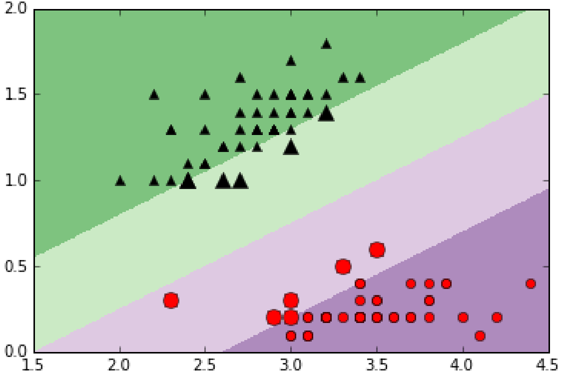
Back to Iris

$C = 1$



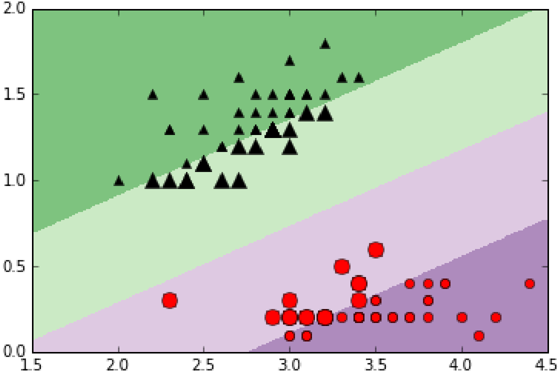
Back to Iris

$C = 0.5$



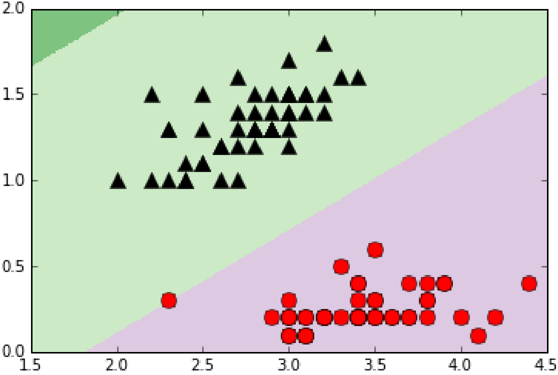
Back to Iris

$C = 0.1$



Back to Iris

$C = 0.01$



Sentiment data

Sentences from reviews on Amazon, Yelp, IMDB, each labeled as positive or negative.

- `Needless to say, I wasted my money.`
- `He was very impressed when going from the original battery to the extended battery.`
- `I have to jiggle the plug to get it to line up right to get decent volume.`
- `Will order from them again!`

Data details:

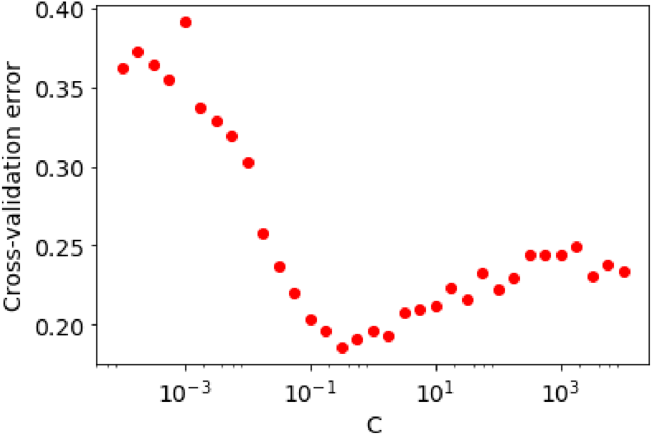
- Bag-of-words representation using a vocabulary of size 4500
- 2500 training sentences, 500 test sentences

What C to use?

C	training error (%)	test error (%)	# support vectors
0.01	23.72	28.4	2294
0.1	7.88	18.4	1766
1	1.12	16.8	1306
10	0.16	19.4	1105
100	0.08	19.4	1035
1000	0.08	19.4	950

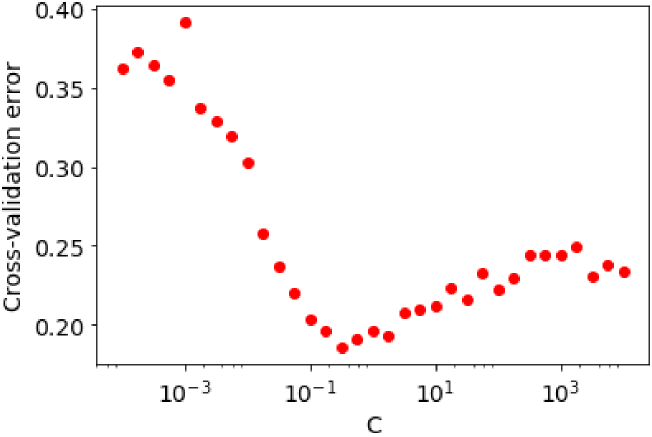
Cross-validation

Results of 5-fold cross-validation:



Cross-validation

Results of 5-fold cross-validation:



Chose $C = 0.32$. Test error: 15.6%