Support vector machines II: Soft-margin SVM

Sanjoy Dasgupta

University of California, San Diego

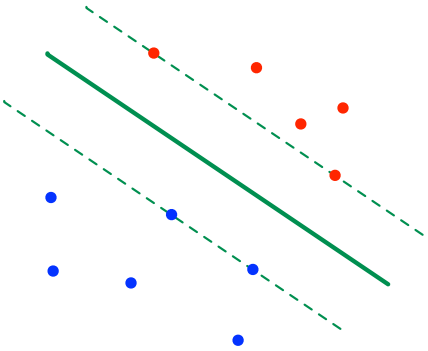
Topics we'll cover

1. Data that isn't linearly separable
2. Adding slack variables for each point
3. Revised convex optimization problem
4. Setting the slack parameter

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.

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



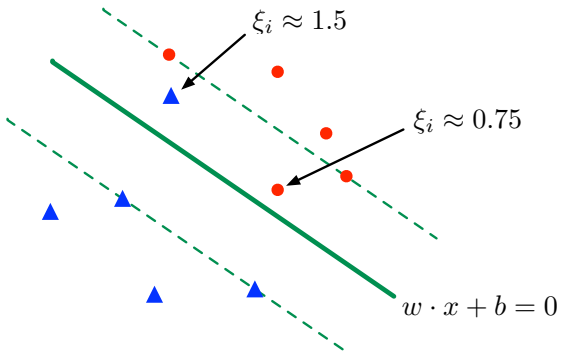
Solution $w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$ is a function of just the support vectors.

What if data is not separable?

The non-separable case

Idea: allow each data point $x^{(i)}$ some **slack** ξ_i.

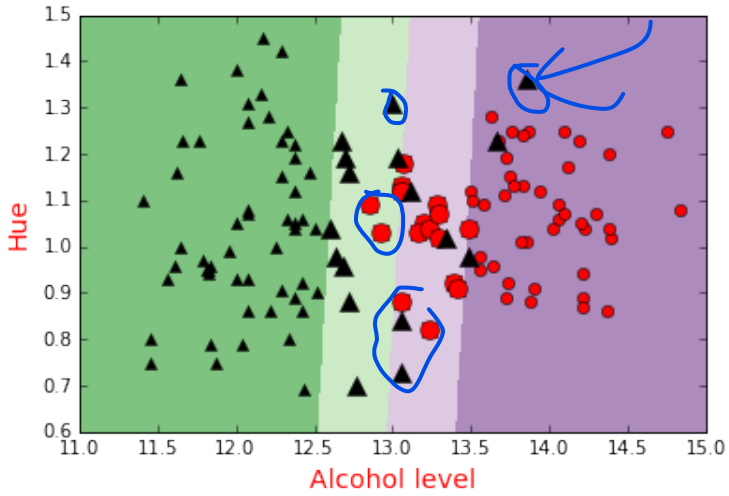
$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \quad \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$
$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n$$
$$\xi \geq 0$$



Wine data set

Here $C = 1.0$

Support vector



The tradeoff between margin and slack

C is used to manage the tradeoff between W and e

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

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

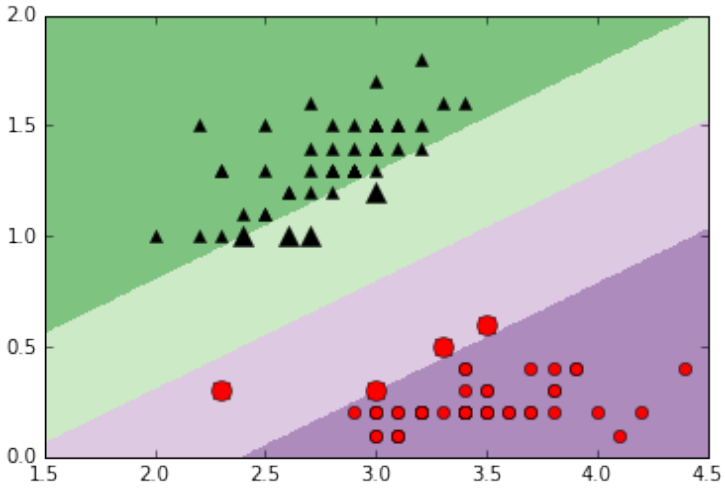
$$\xi \geq 0$$

C=0
slack is free
w=0

C=inf
slack is inf
hard-margin svm

Back to Iris

$C = 1$



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:

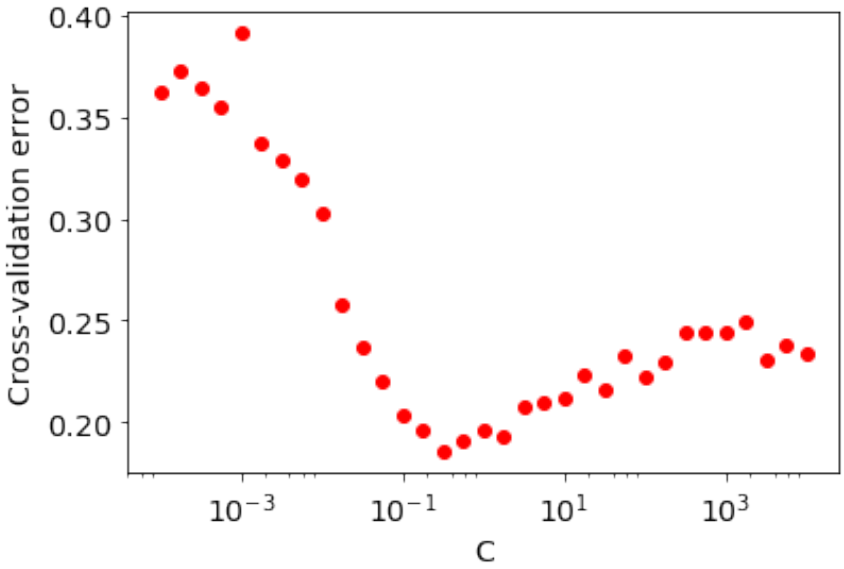
- Bag-of-words representation using a vocabulary of size 4500
- 2500 training sentences, 500 test sentences

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:



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