Lecture 18. Support Vector Machine (SVM)

CS 109A/AC 209A/STAT 121A Data Science:

Harvard University

Fall 2016

Instructors: P. Protopapas, K. Rader, W. Pan

Announcements

Pavlos office hours: Wed 4-5 will be covered by Hari.

HW7: Due tomorrow midnight

HW8: Q2 is optional. In-class competition hosted at Kaggle, top 20% will be awarded +1 point applicable to any HW

Milestone #4-5 deadline: Nov 28

Quiz

- Code quiz: bernie2020

Elections

- What have we learned? Why the predictions were so off?
  - Modeling errors?
  - Methodology?
  - Data?
  - People?

Support Vector Machine (SVM)

- Support vector machines (SVM) is a classification method from the 90s that is very popular.
- SVMs performs well in a many different applications
- One of the best “out of the box” classifiers (as RF or AdaBoost).

Support Vector Machine (SVM)

1. Maximal margin classifier, a simple and intuitive model that is not useful since it requires that the classes are separable.

2. Support vector classifier is more general and useful classifier, still only allows for linear class boundaries

3. Support vector machine, which is a further extension of the support vector classifier in order to accommodate non-linear class boundaries.

Support Vector Machine (SVM)

1. Maximal margin classifier, a simple and intuitive model that is not useful since it requires that the classes are separable.

2. Support vector classifier is more general and useful classifier, still only allows for linear class boundaries

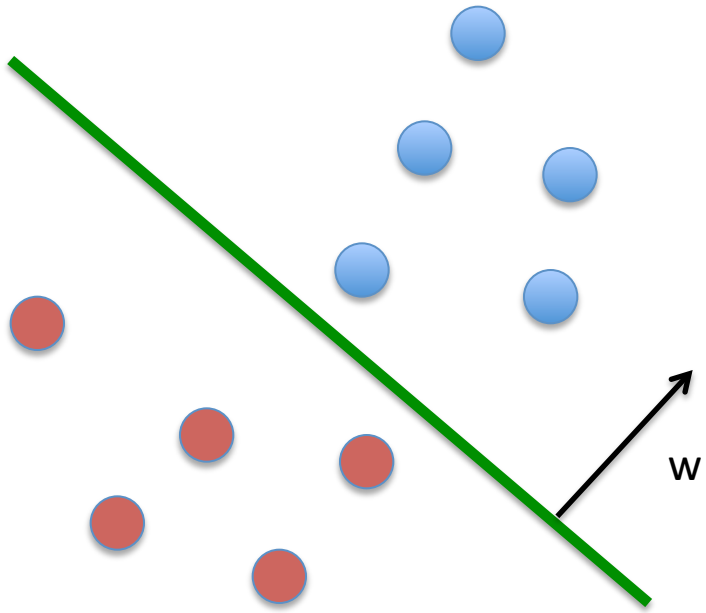
3. Support vector machine, which is a further extension of the support vector classifier in order to accommodate non-linear class boundaries.

Hyperplane

x: data point

y: label {-1, 1}

w: weight vector



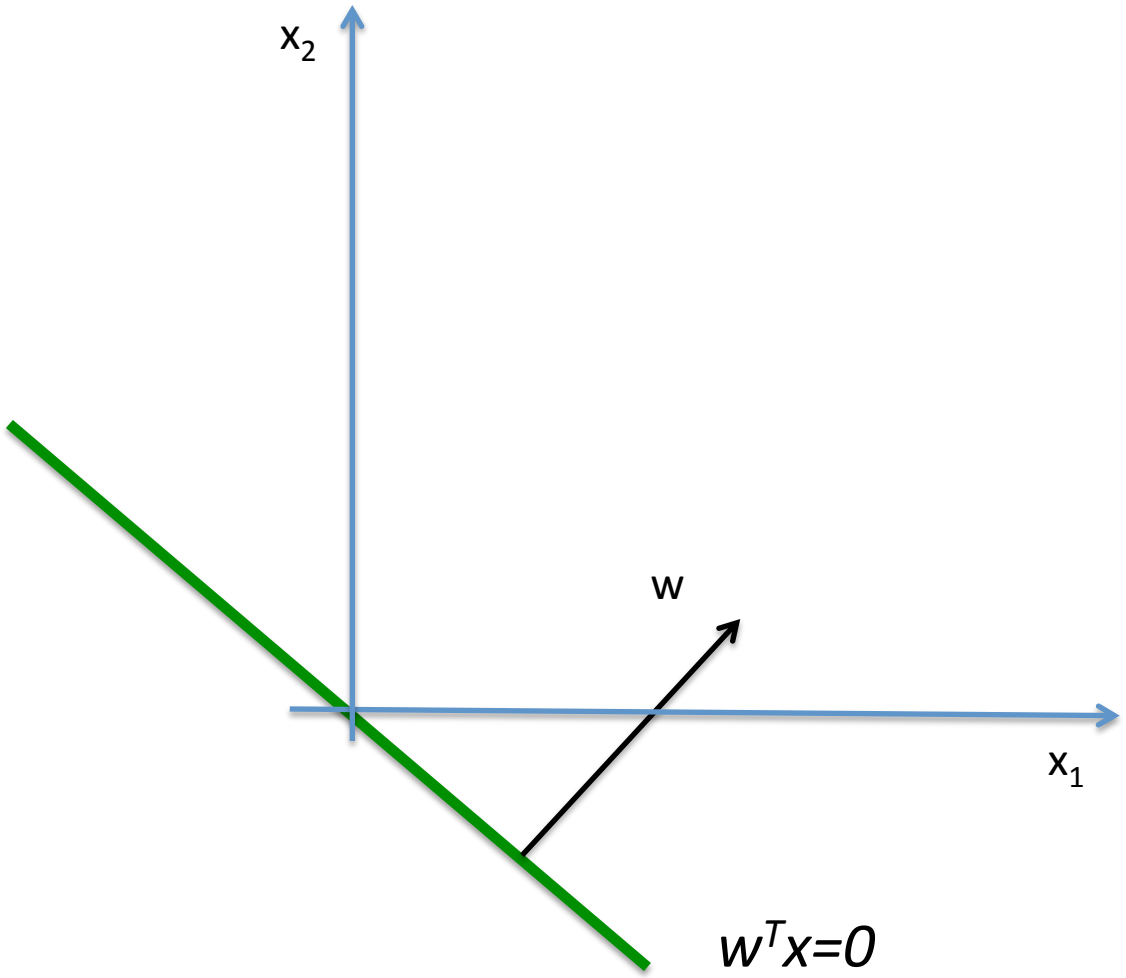
$w^T x=0$

Hyperplane

x: data point

y: label {-1, 1}

w: weight vector



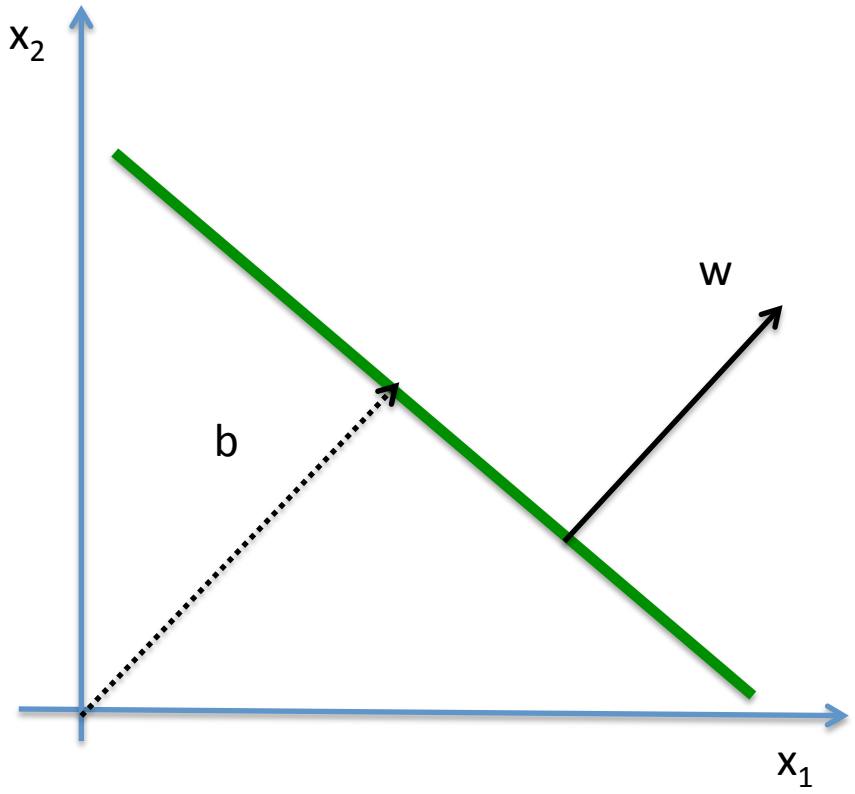
Hyperplane

x: data point

y: label {-1, 1}

w: weight vector

b: bias



$w^T x+b=0$

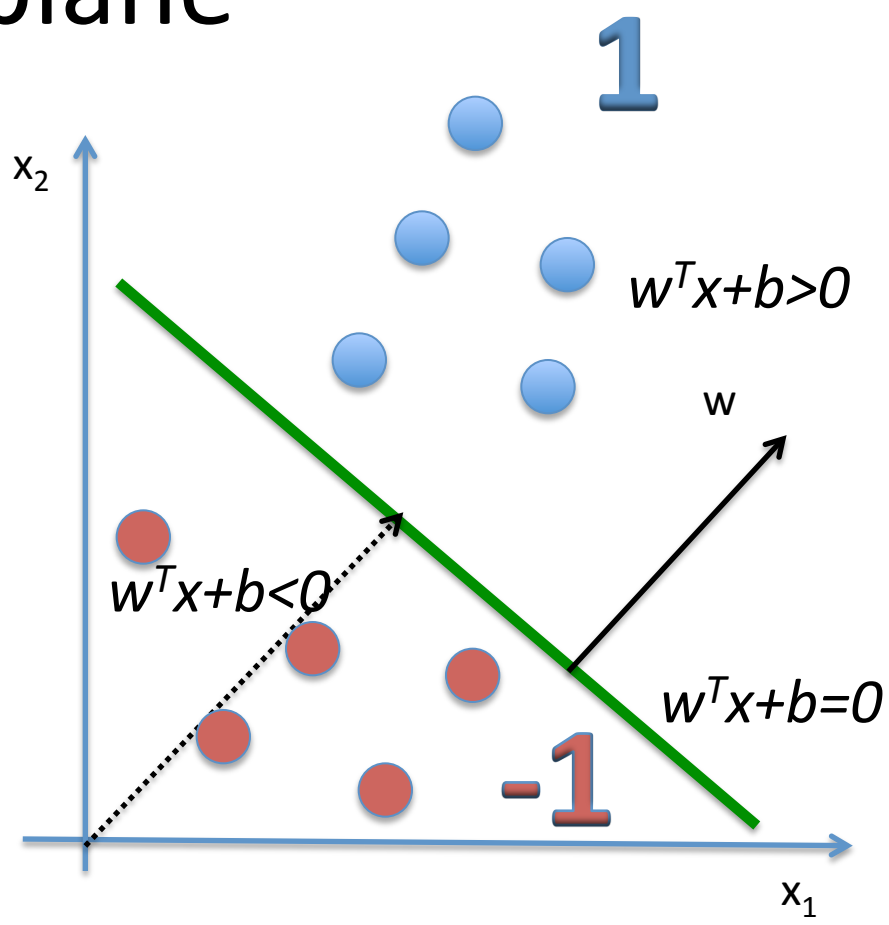
Hyperplane

x: data point

y: label {-1, 1}

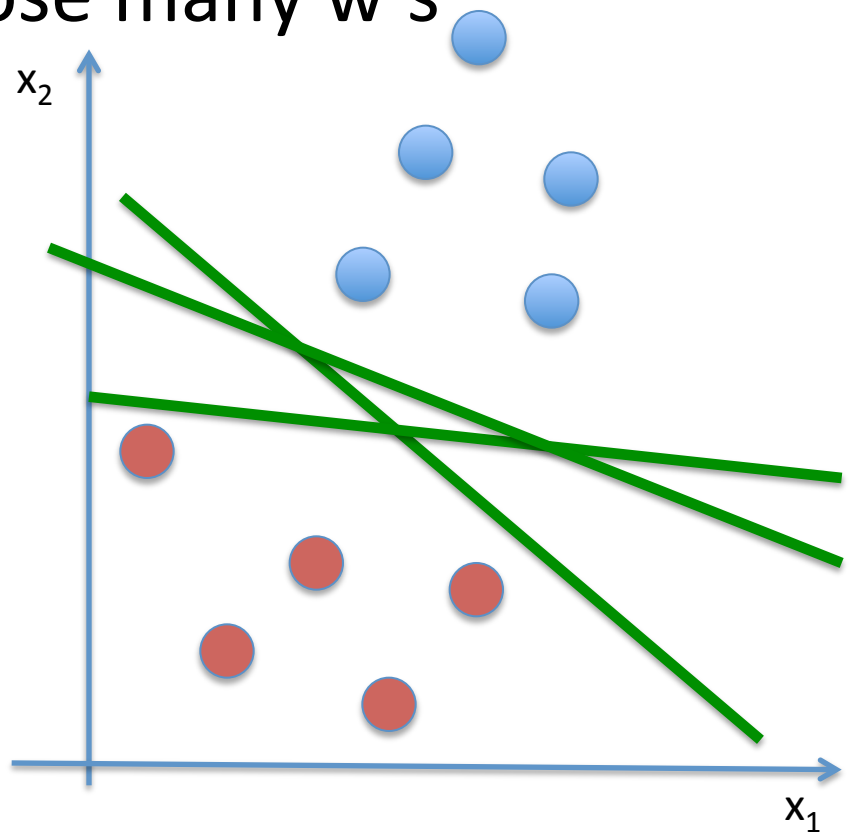
w: weight vector

b: bias



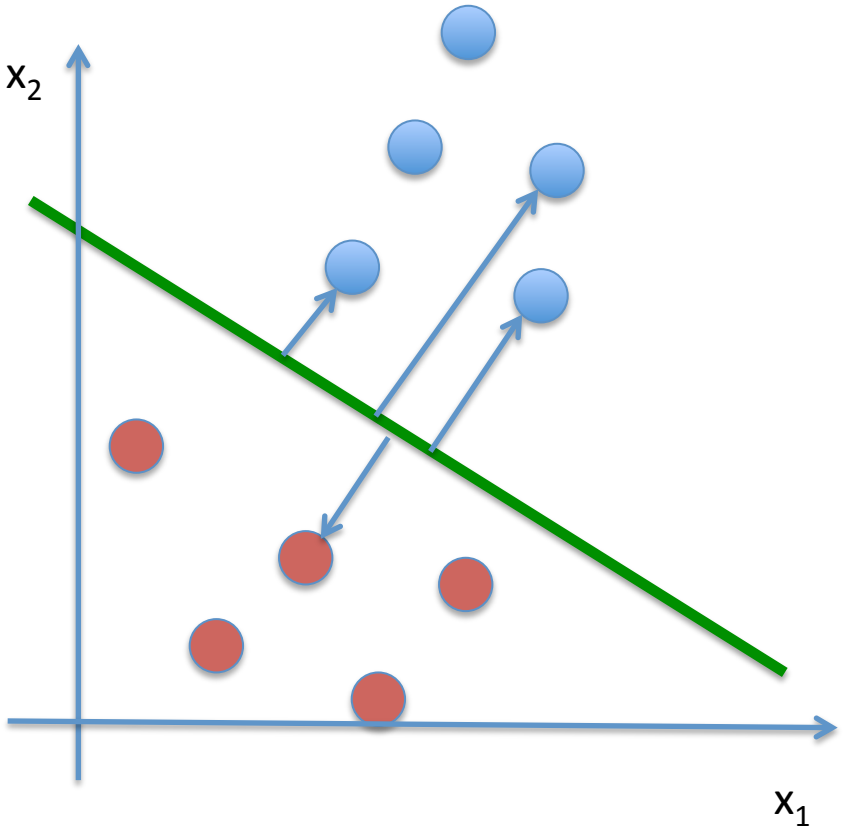
Maximal Margin Classifier

- Can choose many w's



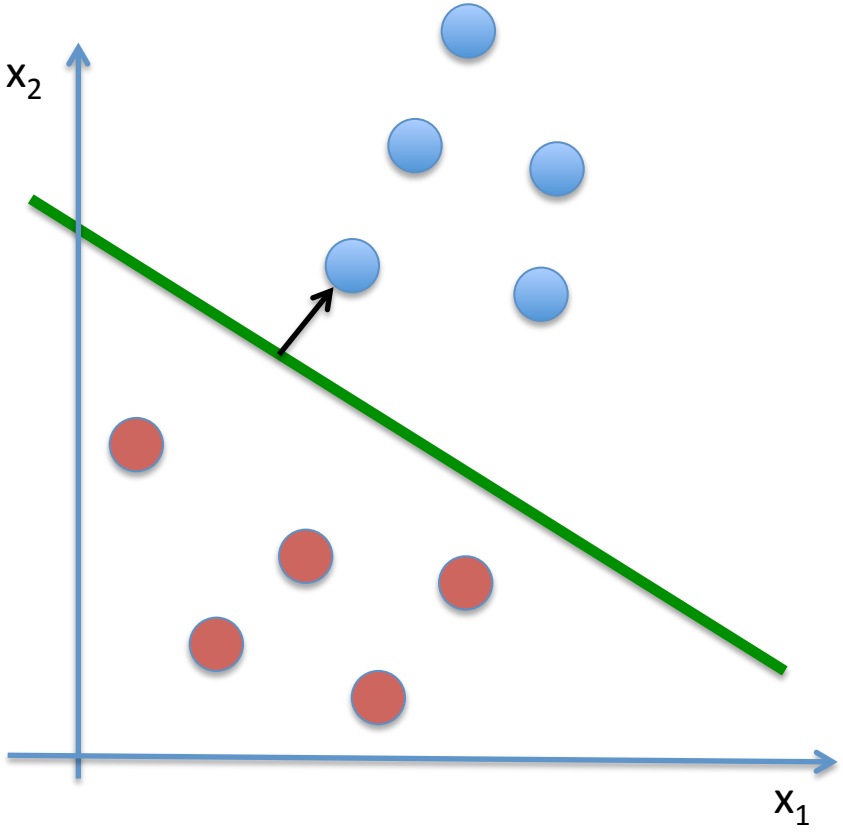
Maximal Margin Classifier

- Given w, we measure the distances to all points



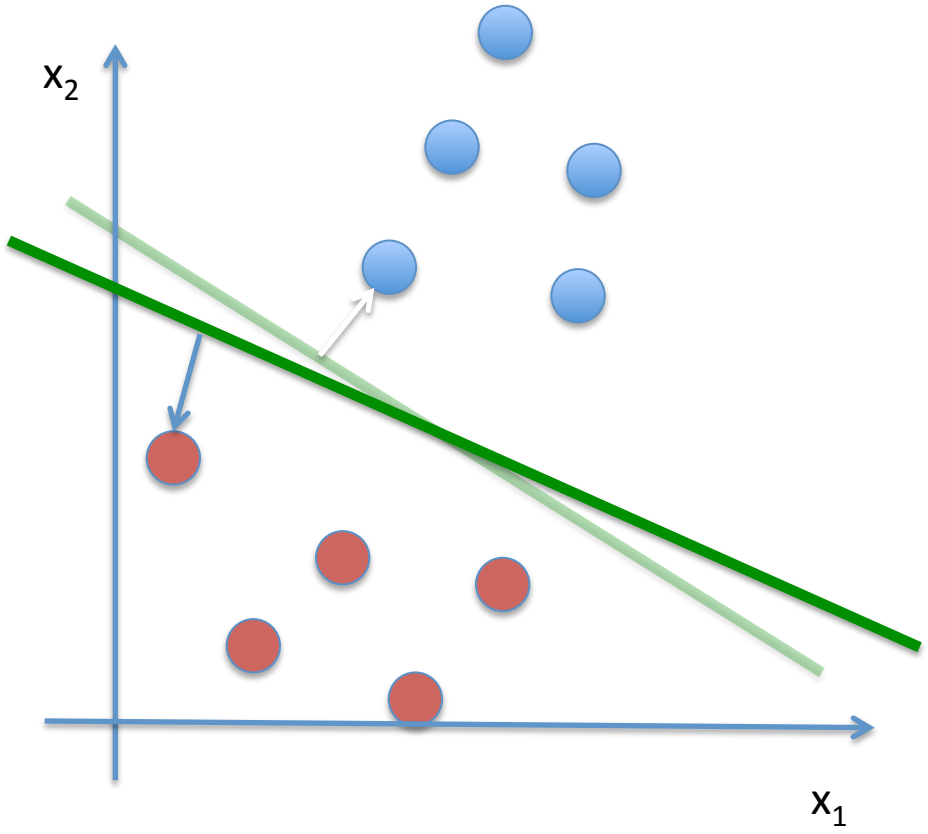
Maximal Margin Classifier

- Find the smallest such distance (the minimum) the **margin** for this hyperplane.



Maximal Margin Classifier

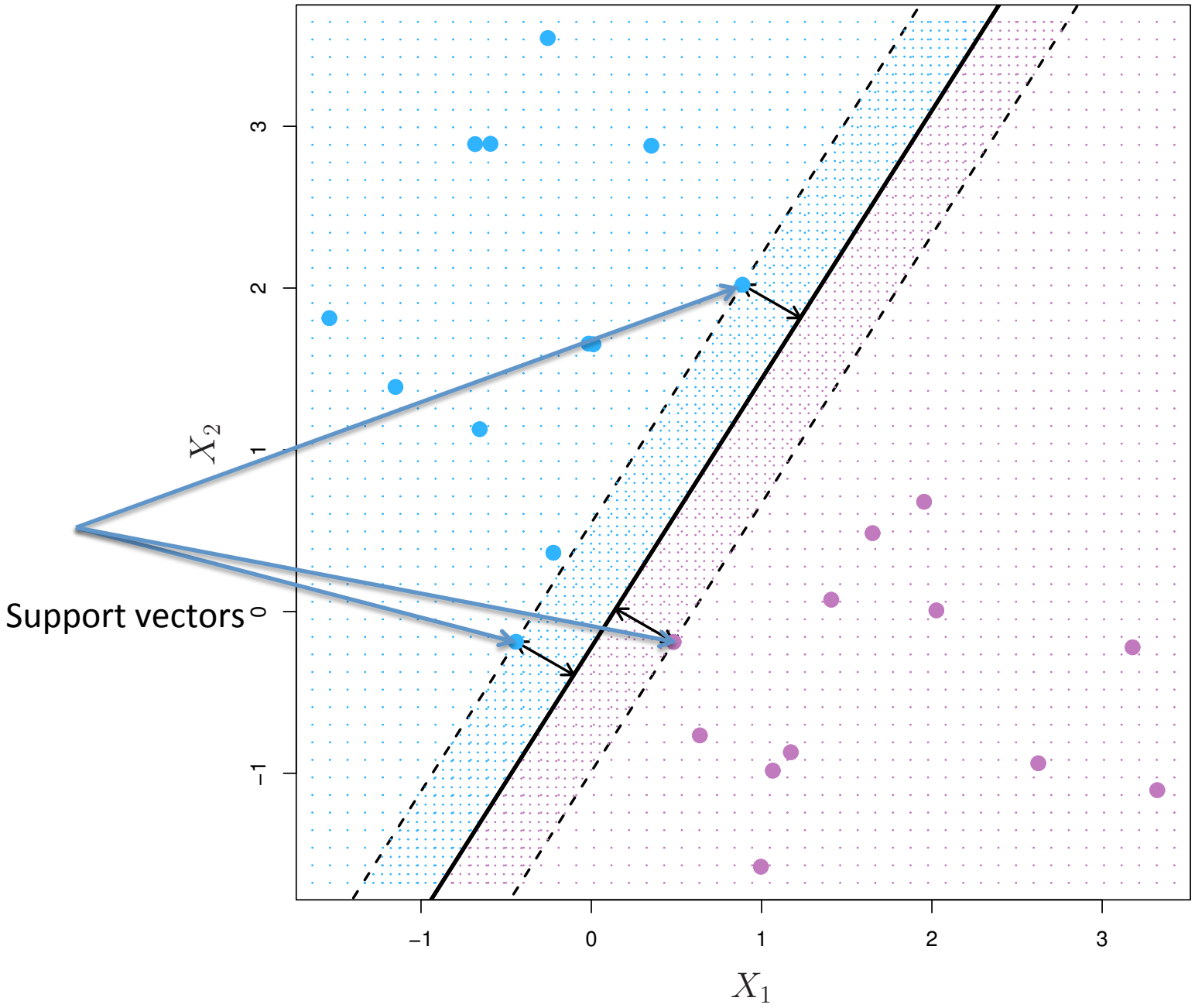
- Do that for all possible hyperplanes



Maximal Margin Classifier

The maximal margin hyperplane is the separating hyperplane for which the margin is largest

The hyperplane that has the farthest minimum distance to the training observations.


Support vectors
X_1
X_2
−1
0
1
2
3
−1
0
1
2
3

- The maximal margin hyperplane **only** depends on the support vectors

Maximal Margin Classifier

Equation for hyperplale in *p* dimensions:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

If

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0$$

The point lies on one side of the hyperplane

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0$$

Lies on the other side

Maximal Margin Classifier

If y={-1,1} then:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0$$

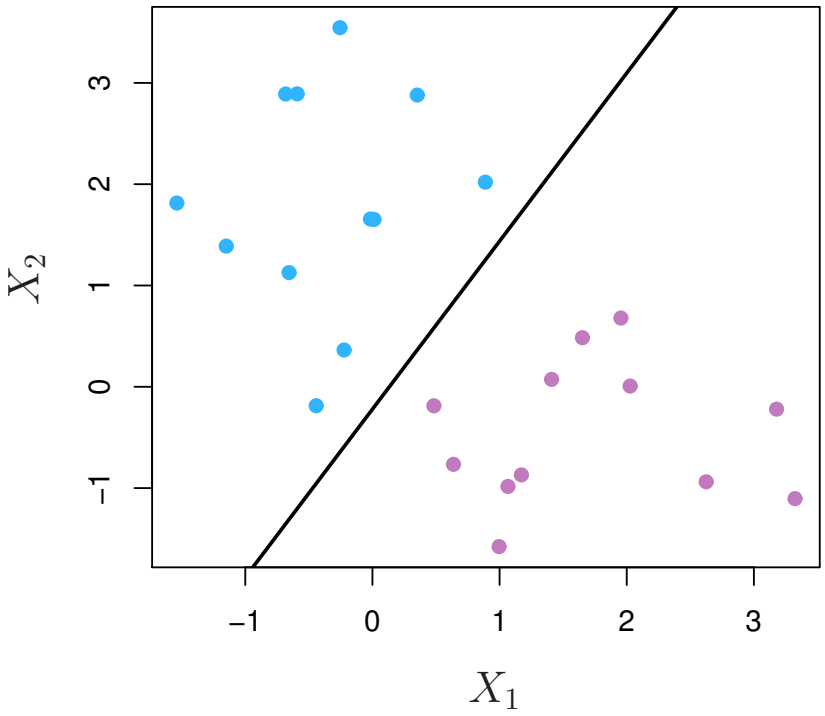
$$\text{maximize}_{\beta_0,\beta_1,\ldots,\beta_q} M$$

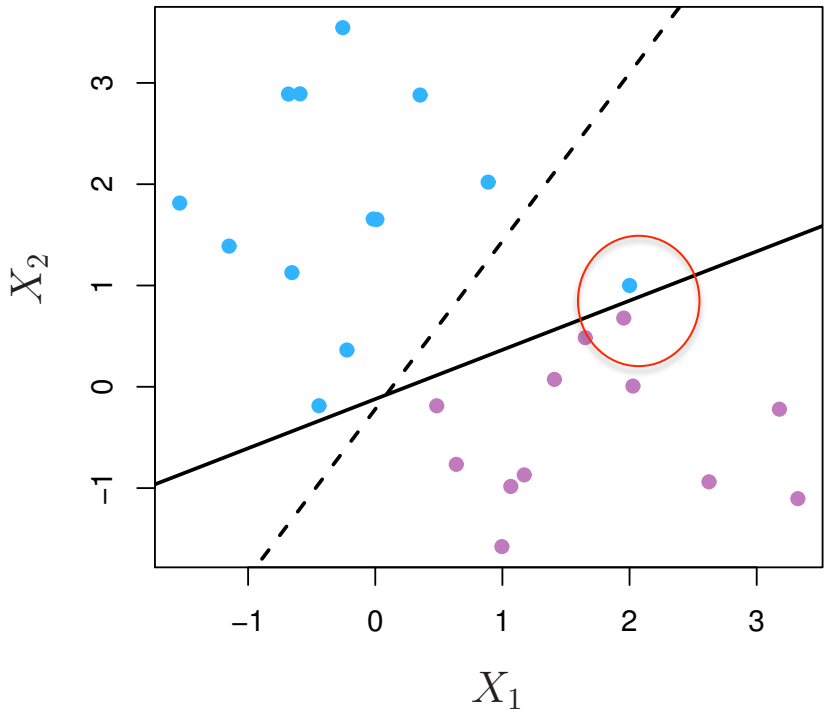
$$s.t. \sum_j \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M$$

M represents the *margin* where and the optimization finds the maximum M

Support Vectors





Support Vectors

- What if there is no hyperplane that separates the two classes?

Support Vectors

- What if there is no hyperplane that separates the two classes?

Support Vectors

- Shall we consider a classifier based on a hyperplane that does not perfectly separate the two classes:
  - Greatest robustness to individual measurements
  - Better classification of most of the training observations

  Misclassify a few training observations in order to do a better job in classifying the remaining observations

  Support vector classifier, aka **soft margin classifie**r, does this

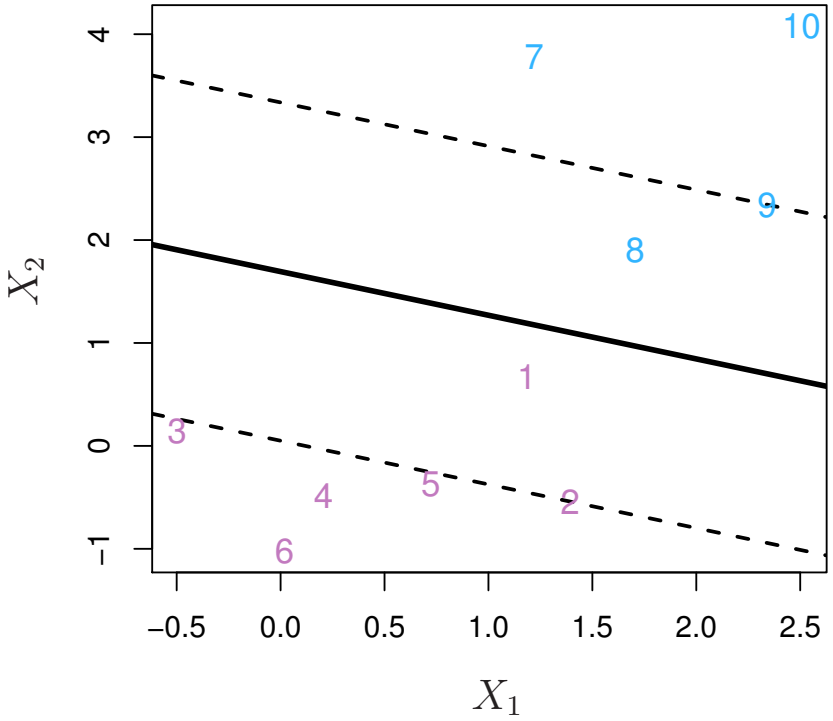
Support Vectors

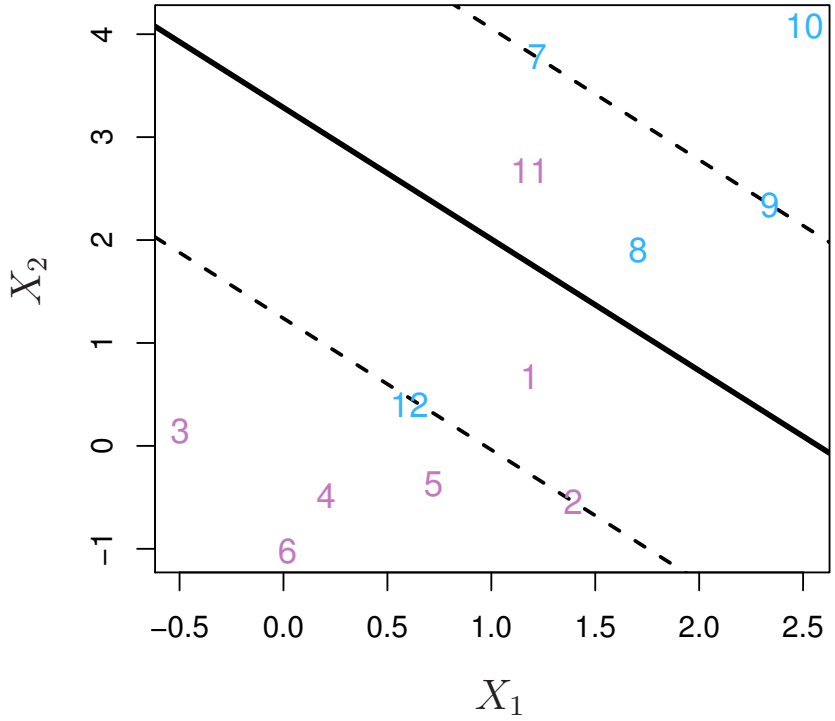
- We consider a classifier based on a hyperplane that does not perfectly separate the two classes:
  - Greatest robustness to individual measurements
  - Better classification of most of the training observations

  Misclassify a few training observations in order to do a better job in classifying the remaining observations

  Support vector classifier, aka **soft margin classifie**r, does this

Support Vector Classifier





Support Vector Classifier

Simply speaking instead of looking at the

maximum of the minimum of the distances

we redefine the margin not to be the minimum of the distances but we allow some points to be on the other side of the margins or even the other side of the hyperplane.

C, a tuning parameter controls how much we allow.

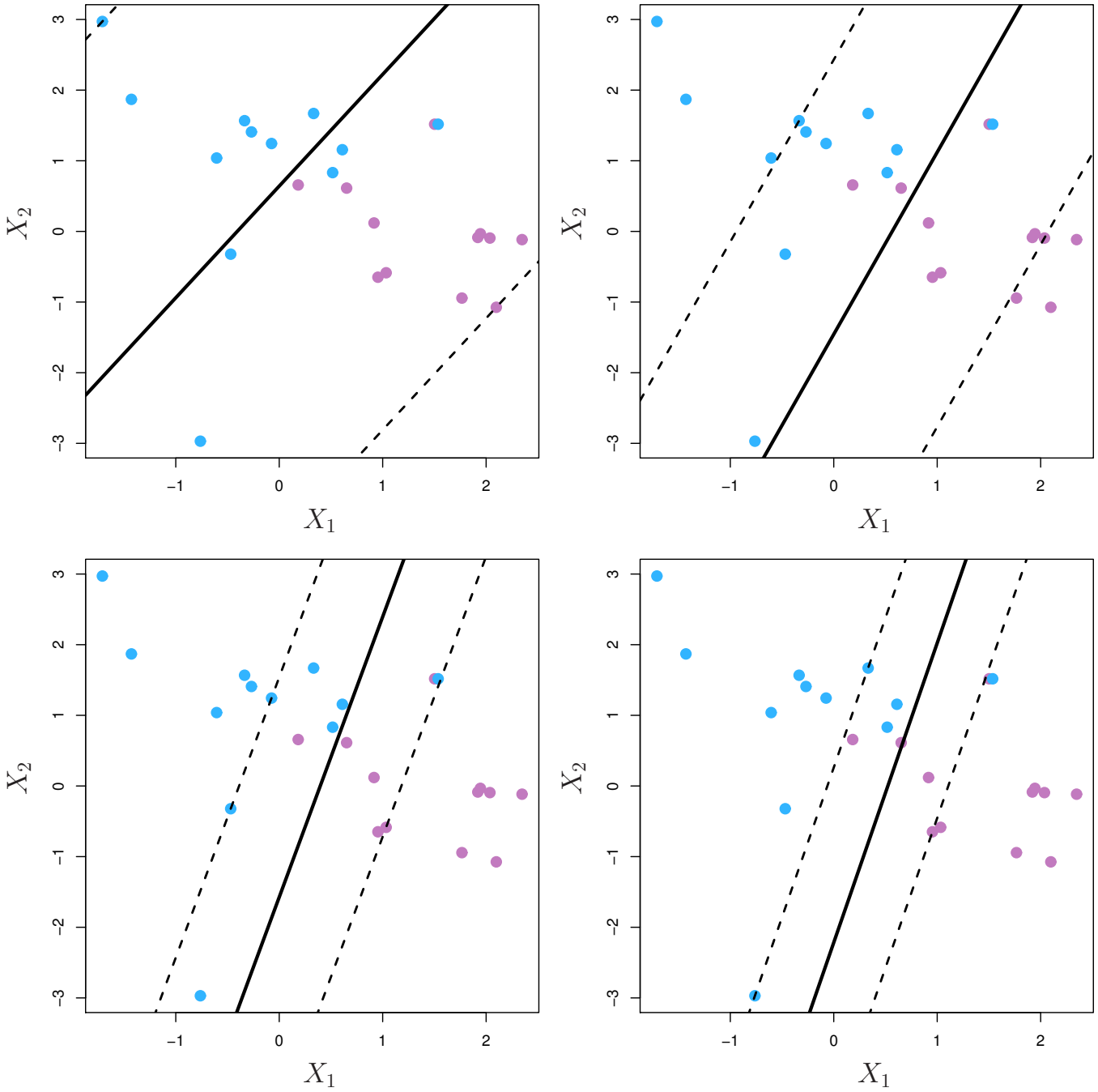
If C=0 then we are back to *Maximal Margin Classifier* (if it is possible)

If C>0 we become more tolerant of violations to the margin. C tells us how many points can be violating the margin.

Support Vector Classifier

We determine *C* using Cross-Validation (as usual) and it is again a tradeoff between bias and variance.

For very small *C* (no mistakes) we fit the training set very well but we will not do well in the testing. If *C* is too large we gave a many hyperplanes that can do the job.


X_1
X_2
−3
−2
−1
0
1
2
3

Support Vector Classifier

$$\text{maximize}_{\beta_0,\beta_1,\ldots,\beta_q} M$$

$$s.t. \sum_j \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \sum_i \epsilon_i \leq C$$

$\varepsilon_1,\ldots,\varepsilon_n$ are slack variables.
If $\varepsilon_i > 0$ then the *i-th* observation is on the wrong side of the margin, and we say that the *i-th* observation has violated the margin. If $\varepsilon_i > 1$ then it is on the wrong side of the hyperplane.