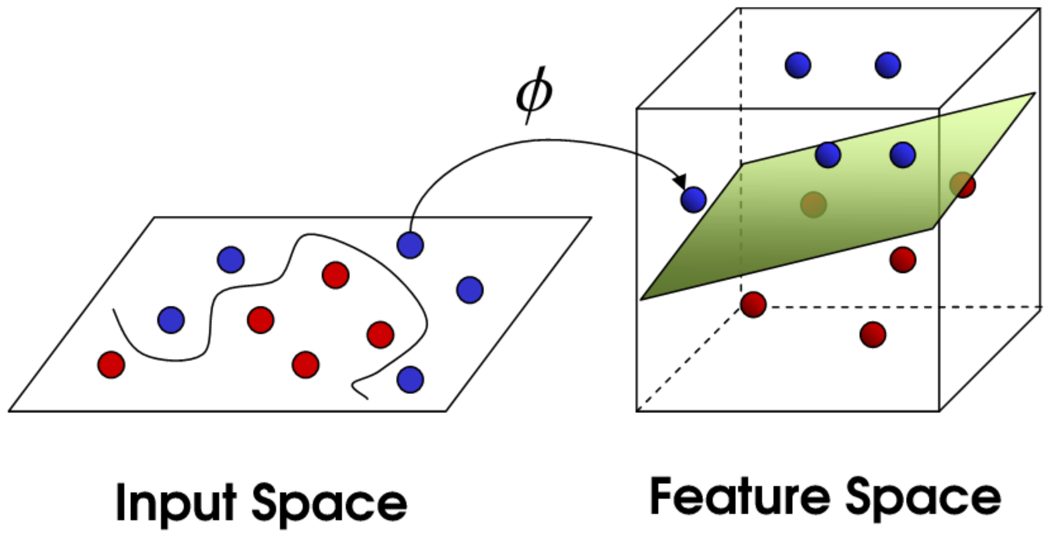
CS109 – Data Science
SVM, Performance evaluation

Hanspeter Pfister, Mark Glickman, Verena Kaynig-Fittkau



http://i.stack.imgur.com/1gvce.png

Announcements

- Midterm 1 grading under way
- HW3 grading next week
- HW4 released tomorrow, due next Wed

Classification vs. Regression

- What is the difference between classification and regression?
- Would you make a regression problem a classification problem?
- Would you make a classification problem a regression problem?

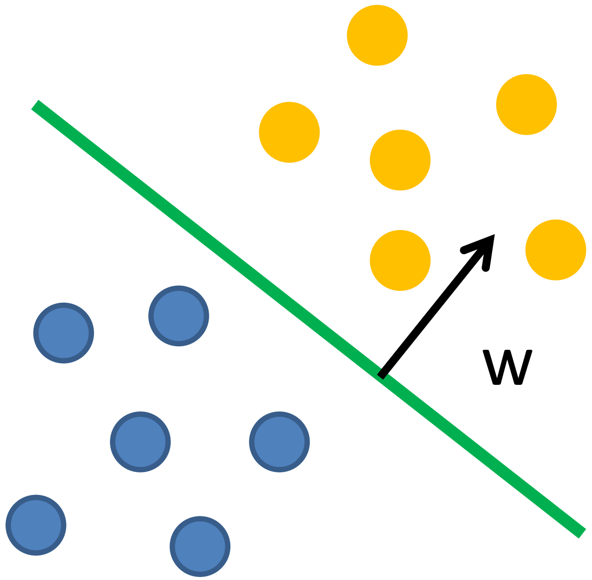
Some classifiers from last semester

- KNN
- Decision tree
- Random Forest
- Logistic regression
- Boosting ?

- Linear SVM

Separating Hyperplane

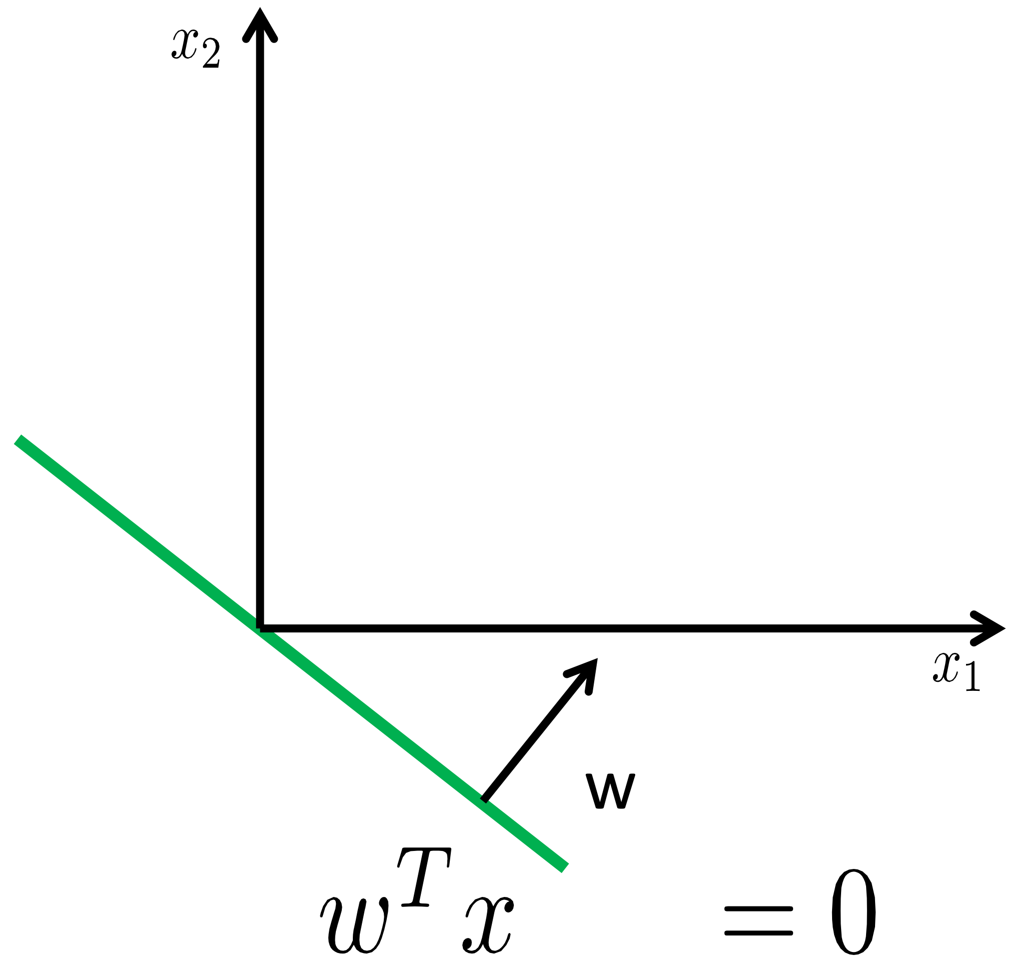
- x: data point
- y: label $\in \{-1, +1\}$
- w: weight vector



$$w^T x \qquad = 0$$

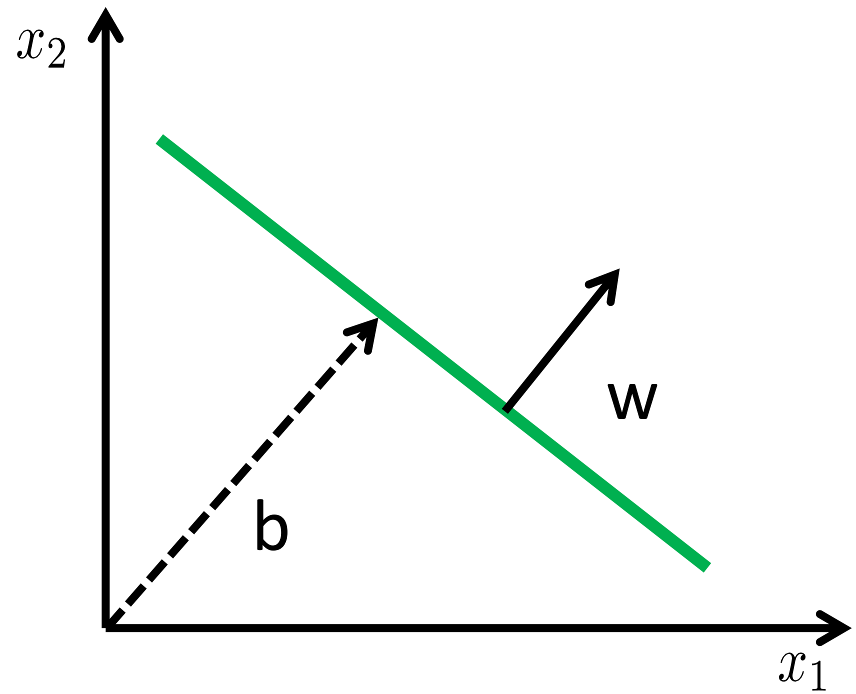
Separating Hyperplane

- x: data point
- y: label $\in \{-1, +1\}$
- w: weight vector



Separating Hyperplane

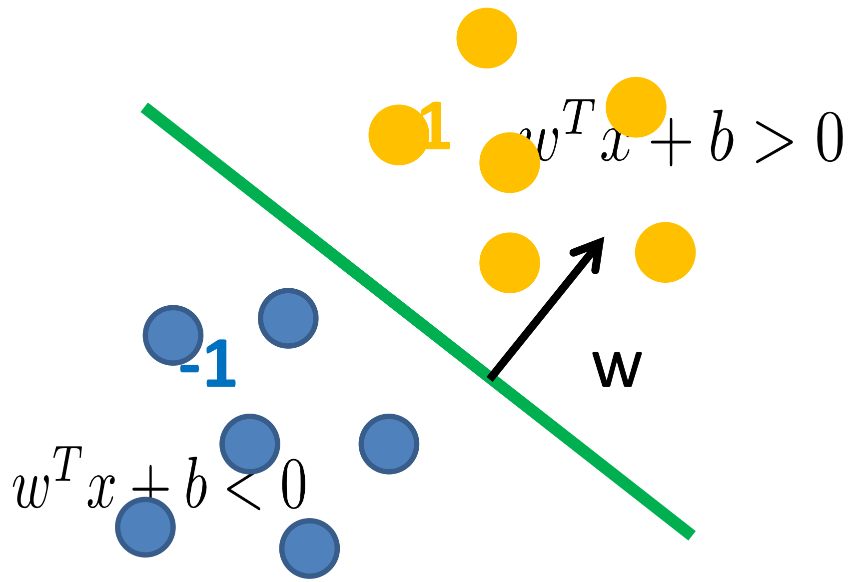
- x: data point
- y: label $\in \{-1, +1\}$
- w: weight vector
- b: bias



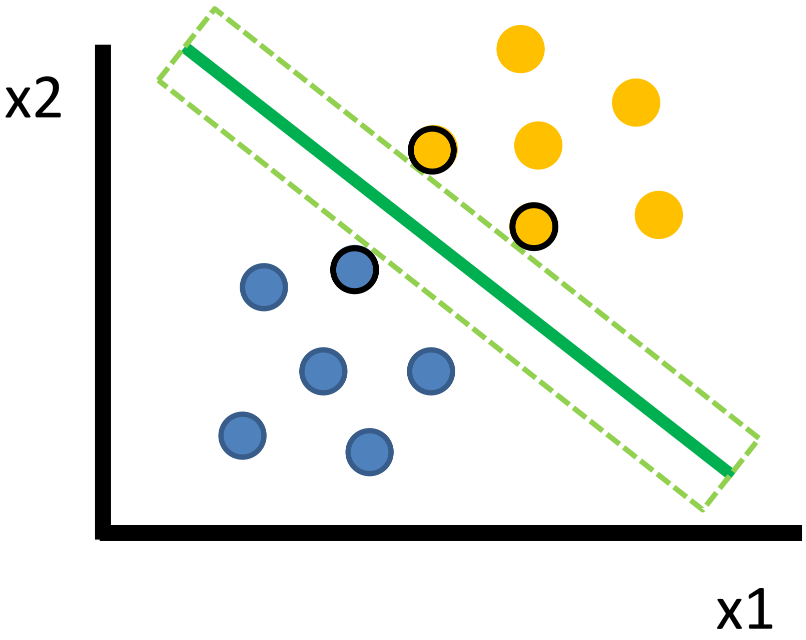
$$w^T x + b = 0$$

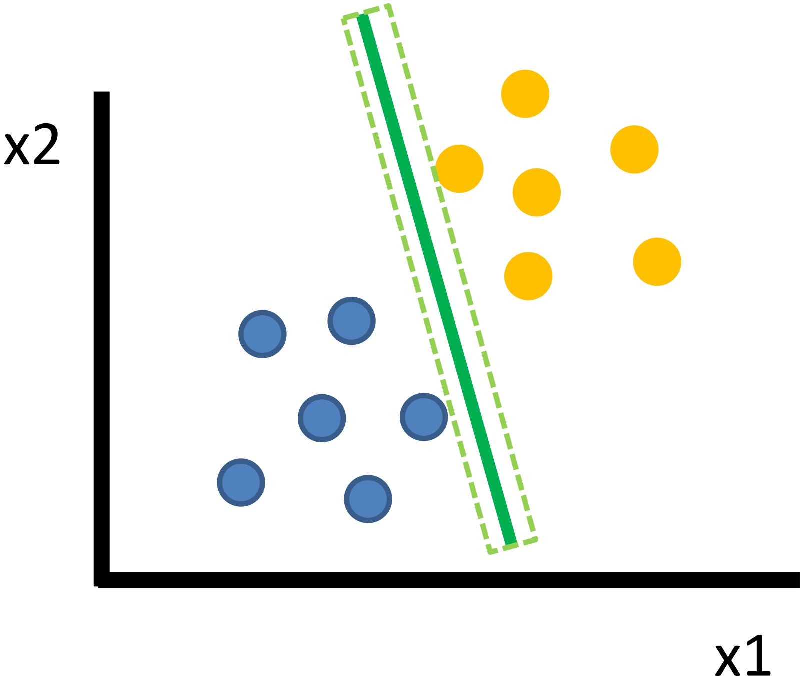
Separating Hyperplane

- x: data point
- y: label $\in \{-1, +1\}$
- w: weight vector
- b: bias



Maximum Margin Classification





Solution depends only on the support vectors!

Maximum Margin Classification

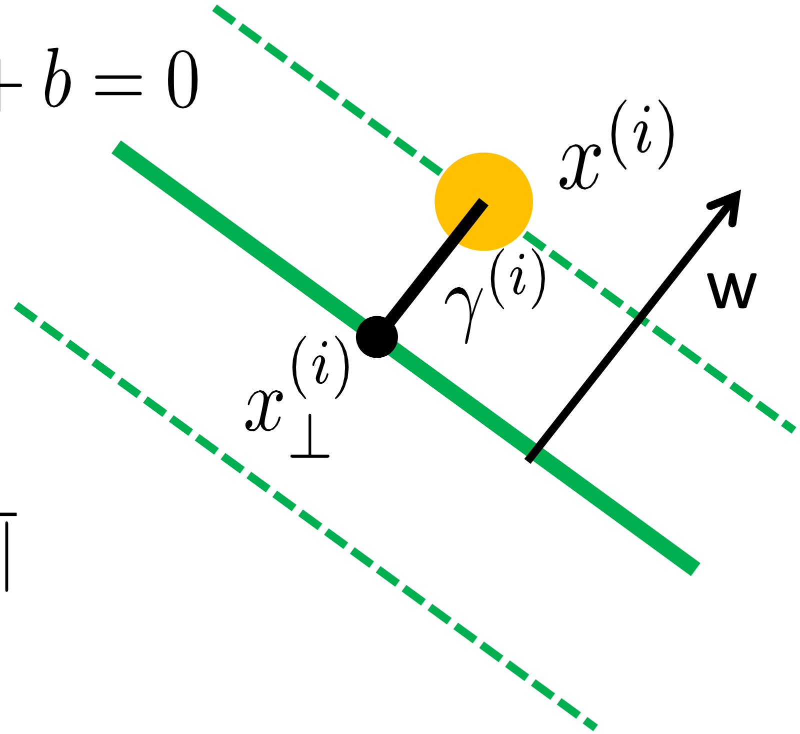
Solution depends only on the support vectors!

Support Vector Machine

- Widely used for all sorts of classification problems

- Some people say it is the best of the shelf classifier out there
- (But it is not that much of the shelf as we would like)

Maximum Margin Classification



margin:

$$x_\perp^{(i)} = x^{(i)} - \gamma^{(i)} \cdot \frac{w}{||w||}$$

$$w^T x_\perp^{(i)} + b = 0$$

$$\Rightarrow \quad \gamma^{(i)} = \quad \left(\frac{w^T x^{(i)} + b}{||w||}\right)$$

Maximum Margin Classification

$$\gamma^{(i)} = y^{(i)}\left(\frac{w^T x^{(i)} + b}{||w||}\right)$$ geometrical margin

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$ functional margin

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

minimal geometrical margin

non-convex

Maximum Margin Classification

$$\max_{\gamma,w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

$$\max_{\gamma,w,b} \quad \gamma$$

minimal geometrical margin

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, m$$

$$||w|| = 1.$$

non-convex

Maximum Margin Classification

$$\max_{\gamma,w,b} \quad \frac{\hat{\gamma}}{||w||}$$

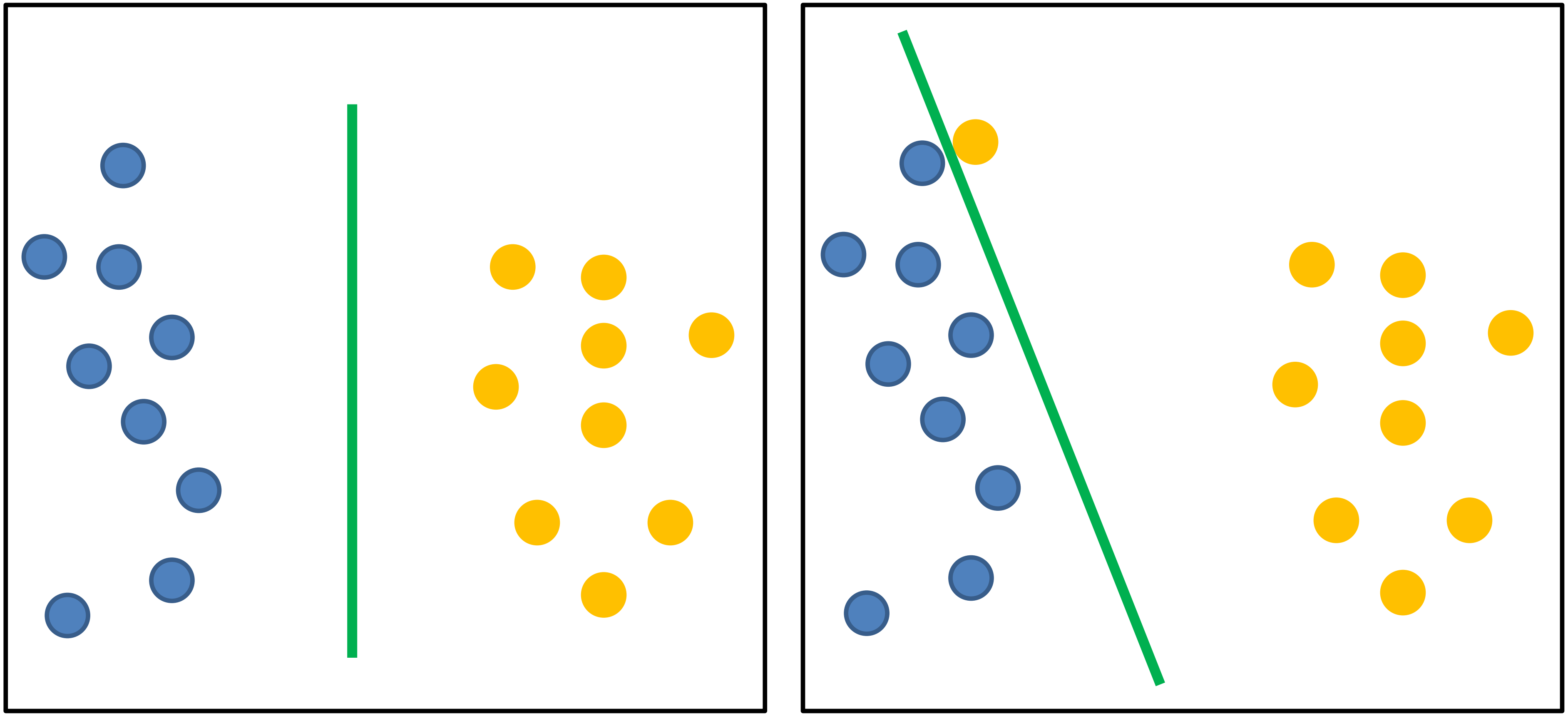
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

functional margin is not normalized – can be arbitrarily scaled

$$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m$$

Two Very Similar Problems



What about outliers?

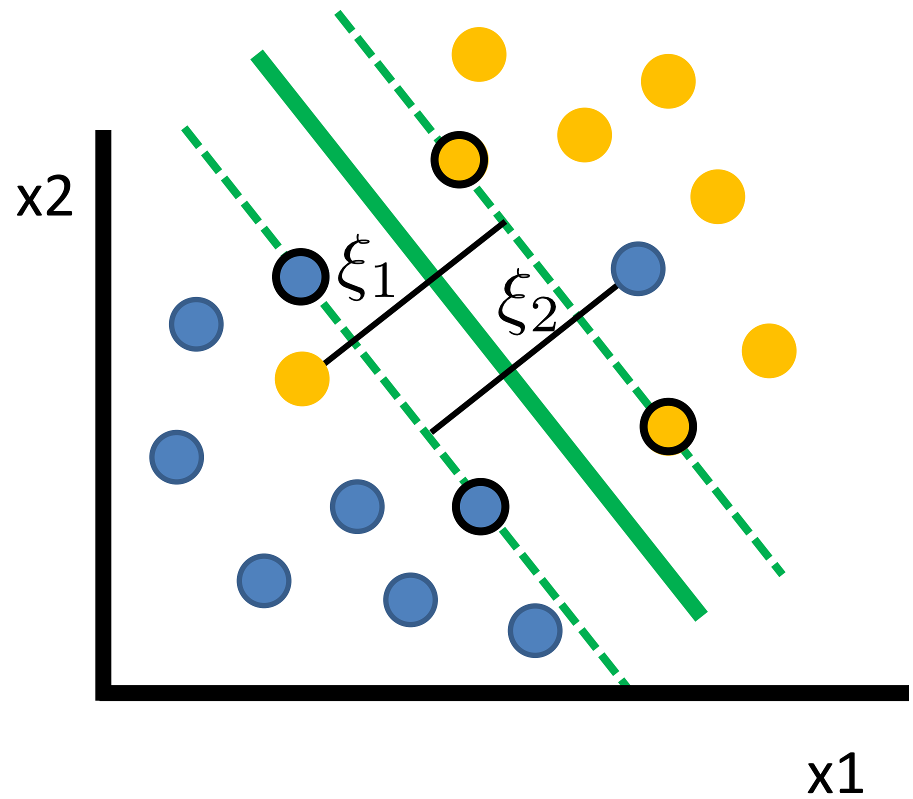
ξ_i: slack variables

$$\min_{w,b,\xi} \frac{1}{2}||w||^2$$

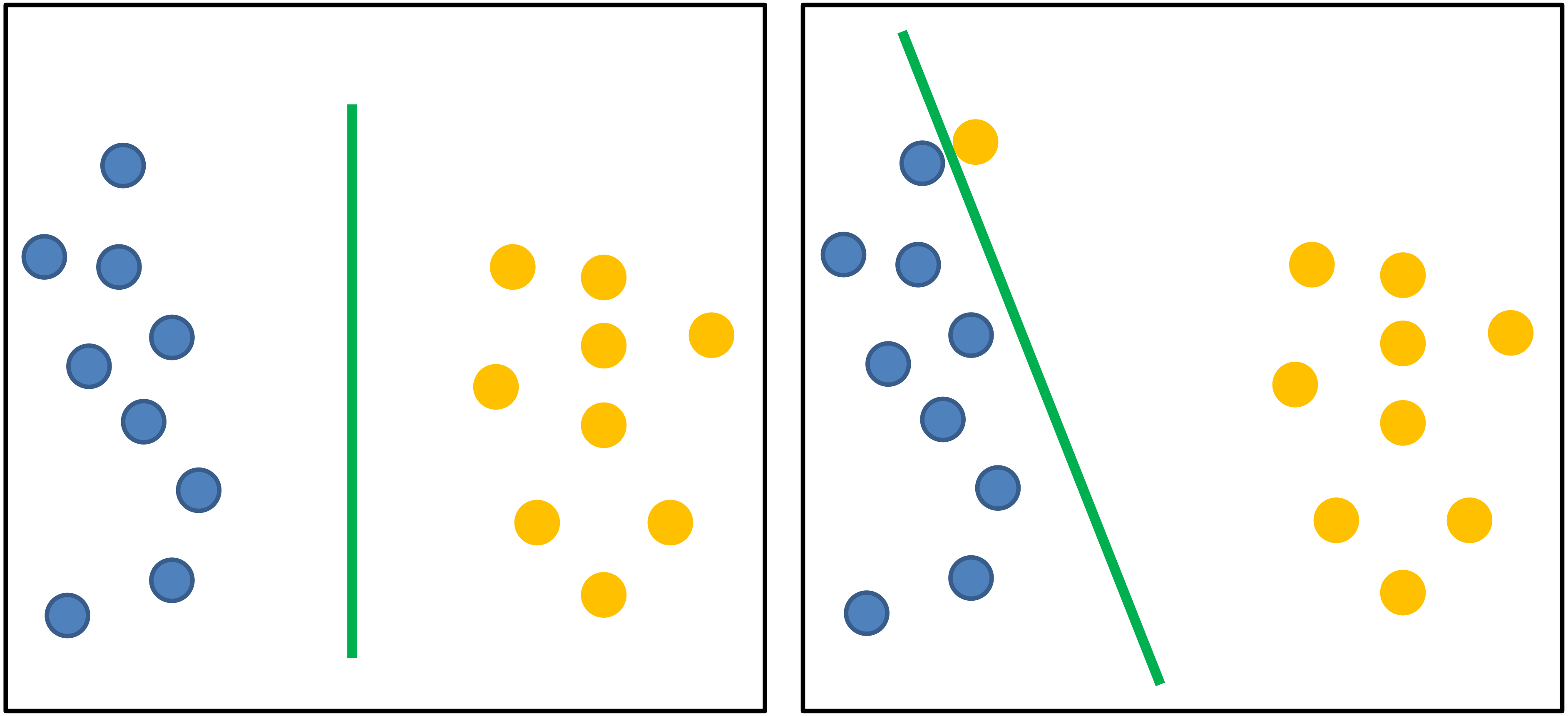
subject to:

$$y^{(i)}(w^T x^{(i)} + b) \geq 1$$

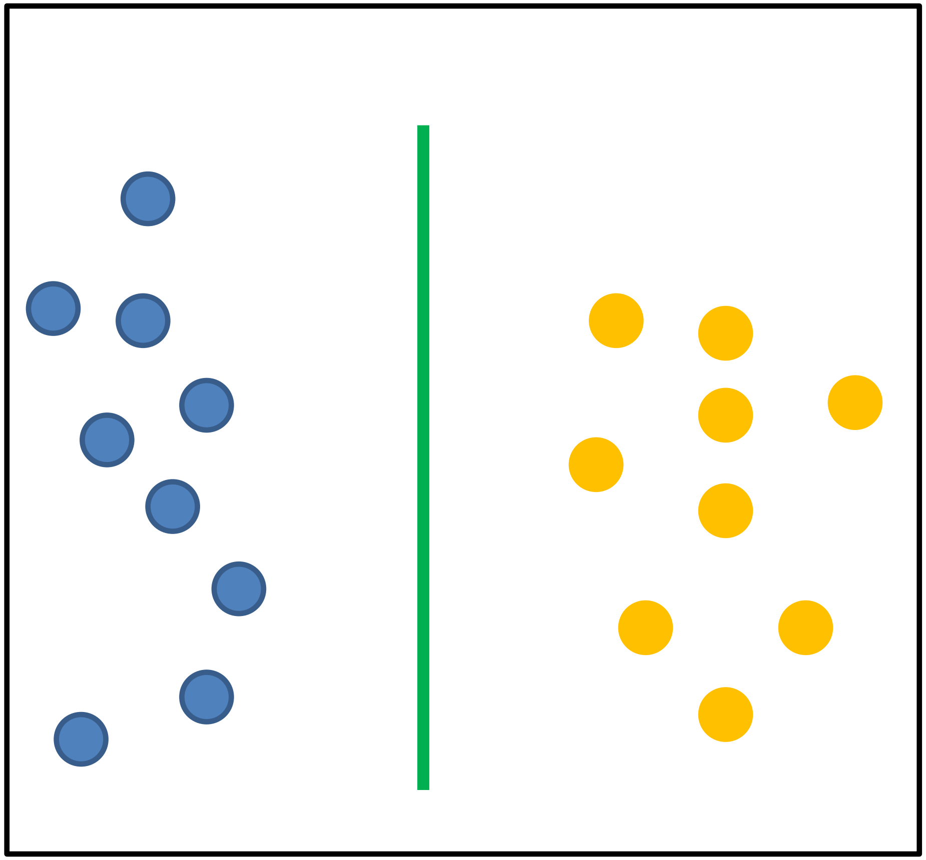
$$(i = 1, \ldots, n)$$

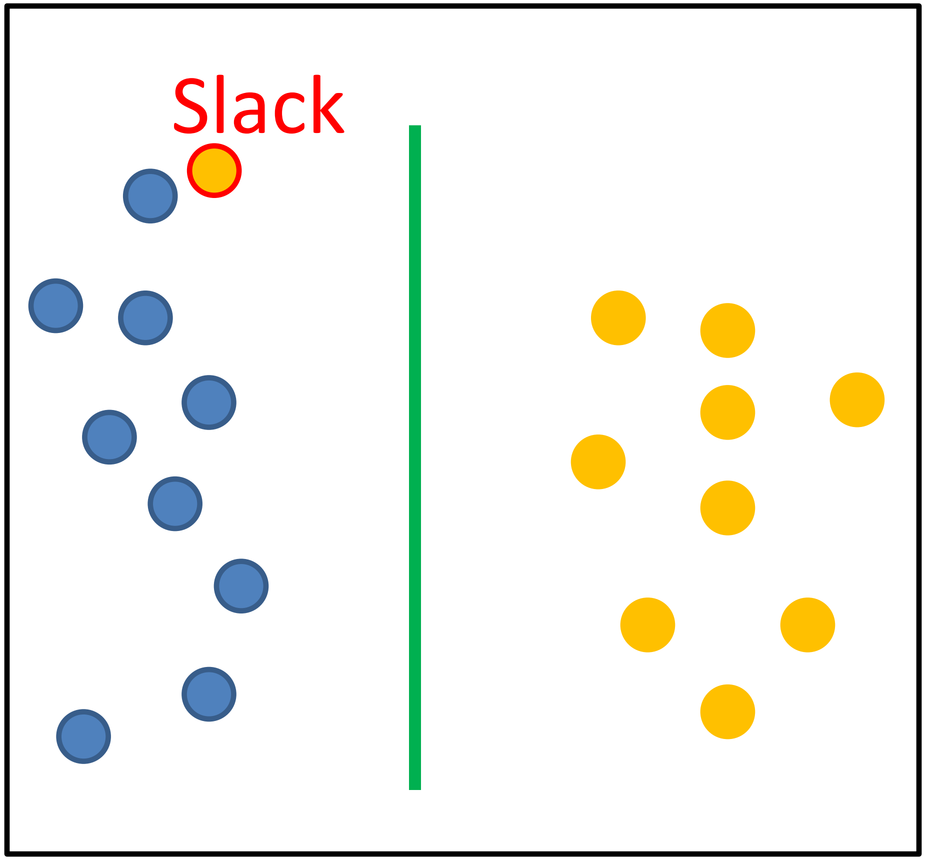


Two Very Similar Problems

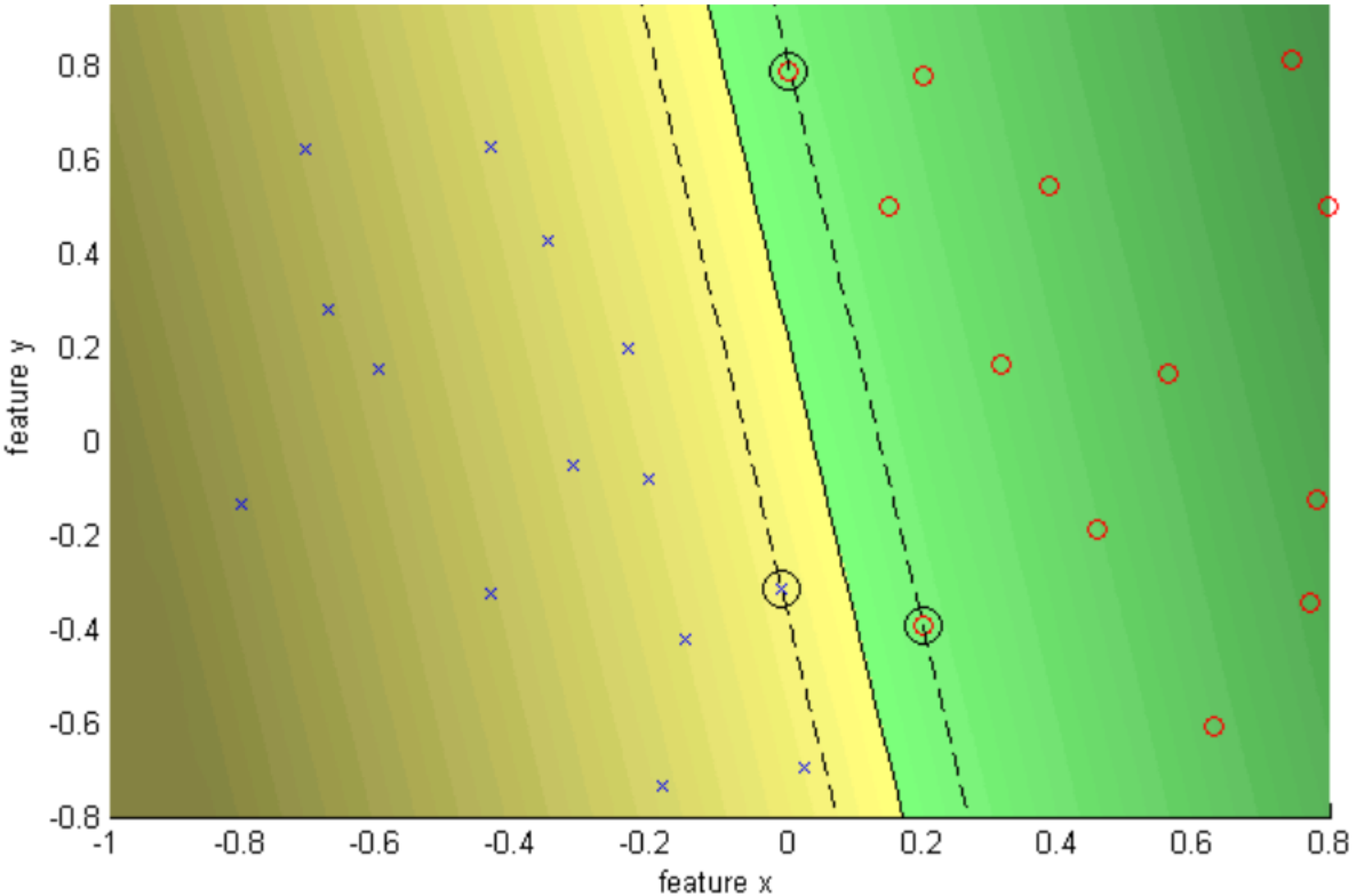


Two Very Similar Problems



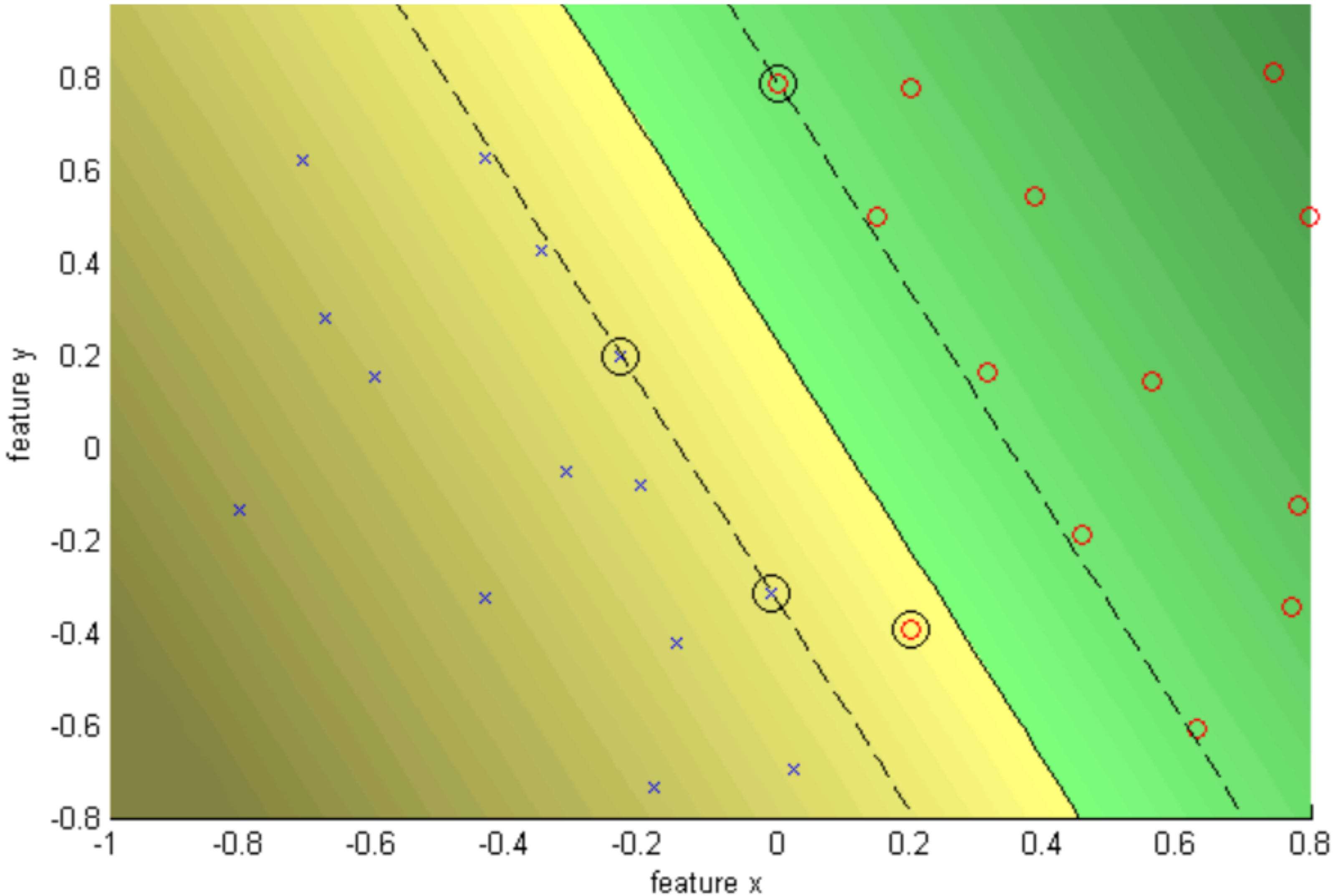


Hard Margin (C = Infinity)



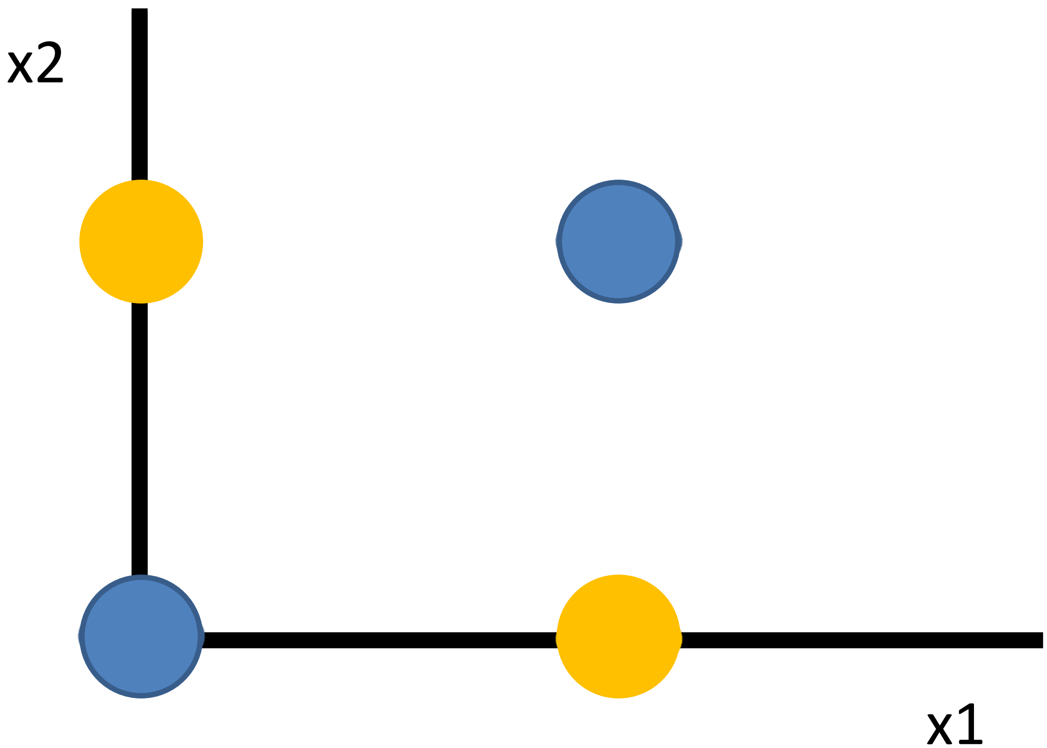
http://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

Soft Margin (C = 10)

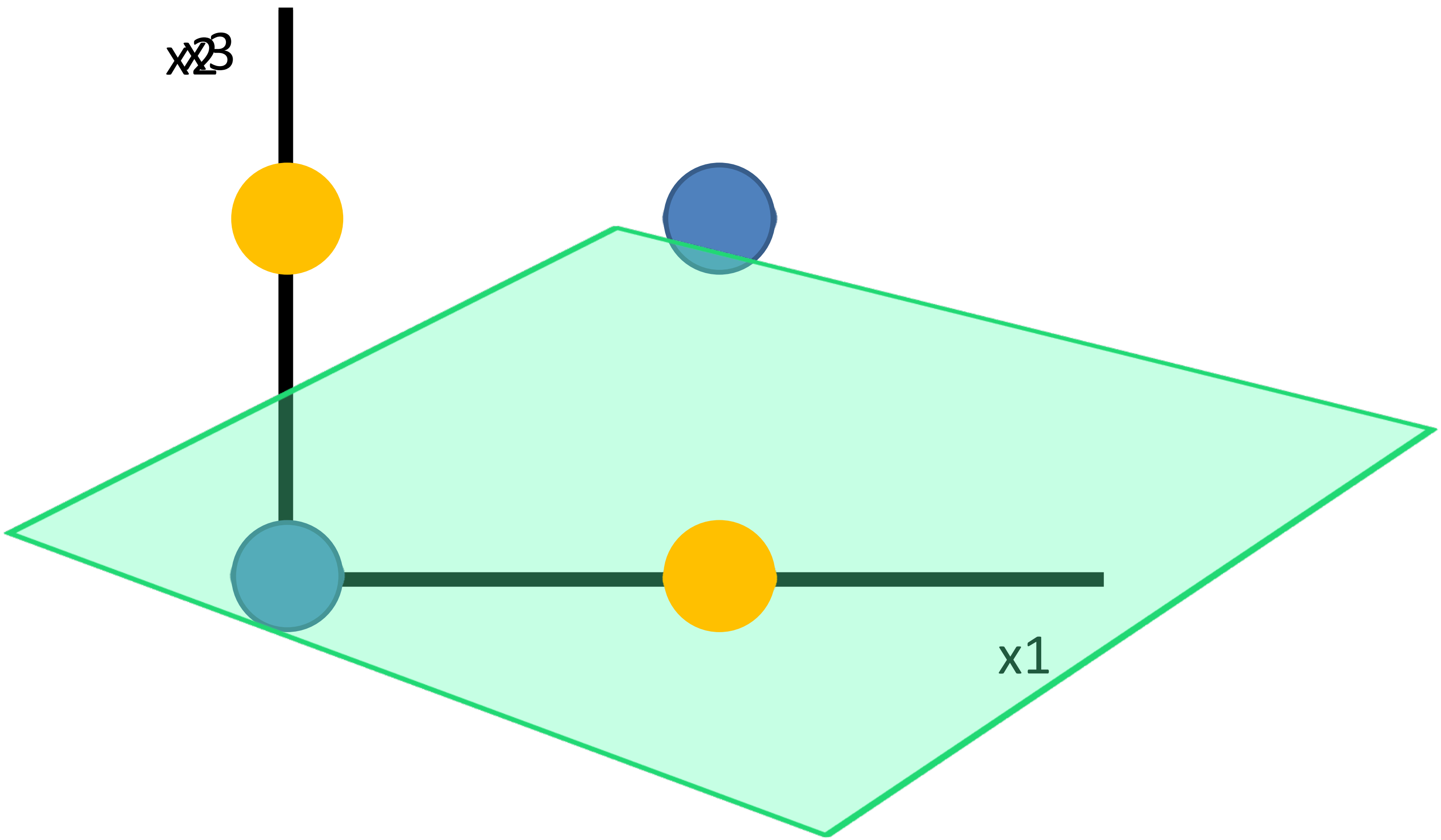


http://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

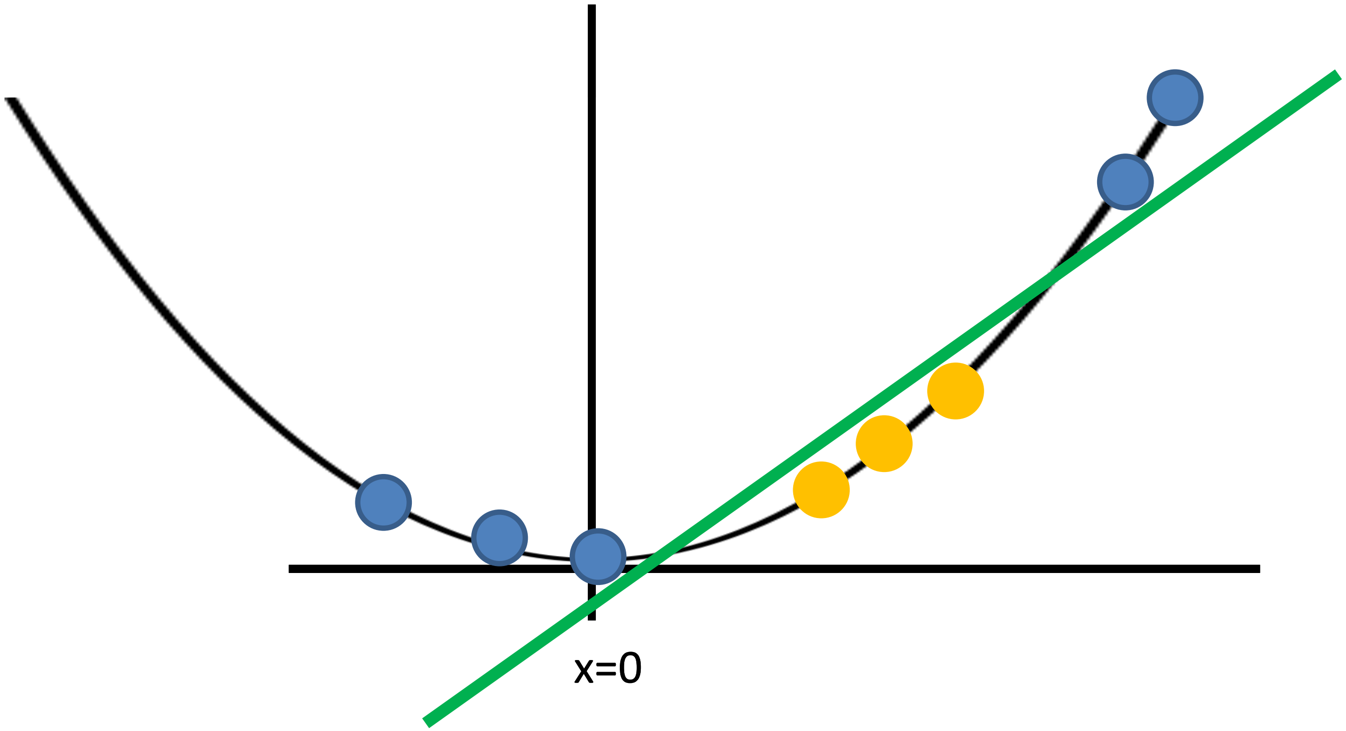
The XOR Problem



The XOR Problem

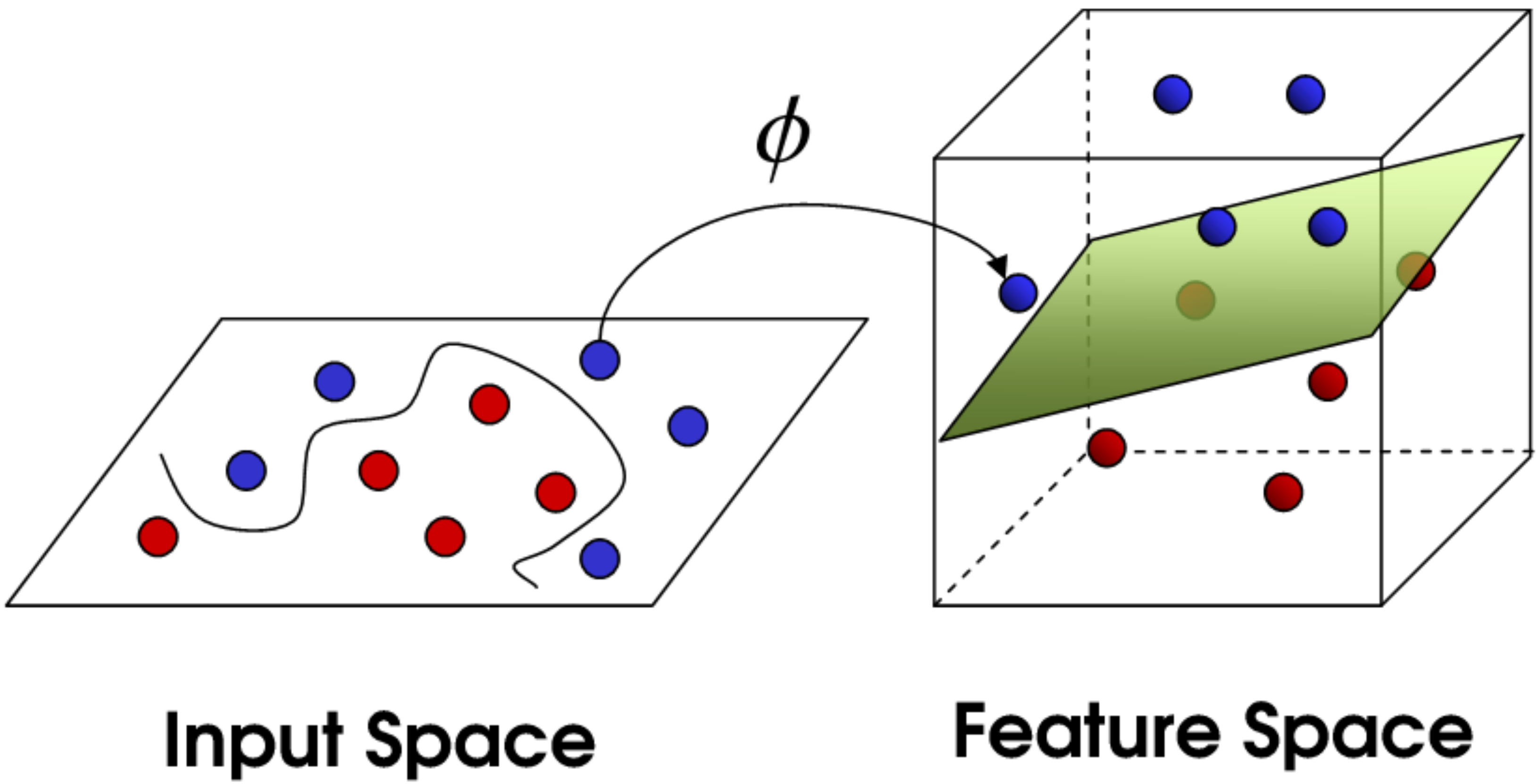


XOR problem revised



Did we add information to make the problem separable?

Non-Linear Decision Boundary



SVM with a polynomial Kernel visualization

Created by:
Udi Aharoni

Quadratic Kernel

$$x = (x_1, x_2)$$

$$\Phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)$$

$$\Phi(x) \cdot \Phi(z) = 1 + 2x_1 z_1 + 2x_2 z_2 + 2x_1 z_1 x_2 z_2$$
$$+ x_1^2 \, z_1^2 \; + x_2^2 \, z_2^2$$

$$\boxed{= (1 + x \cdot z)^2}$$

Kernel Functions

$$K(x, z) = \Phi(x) \cdot \Phi(z)$$

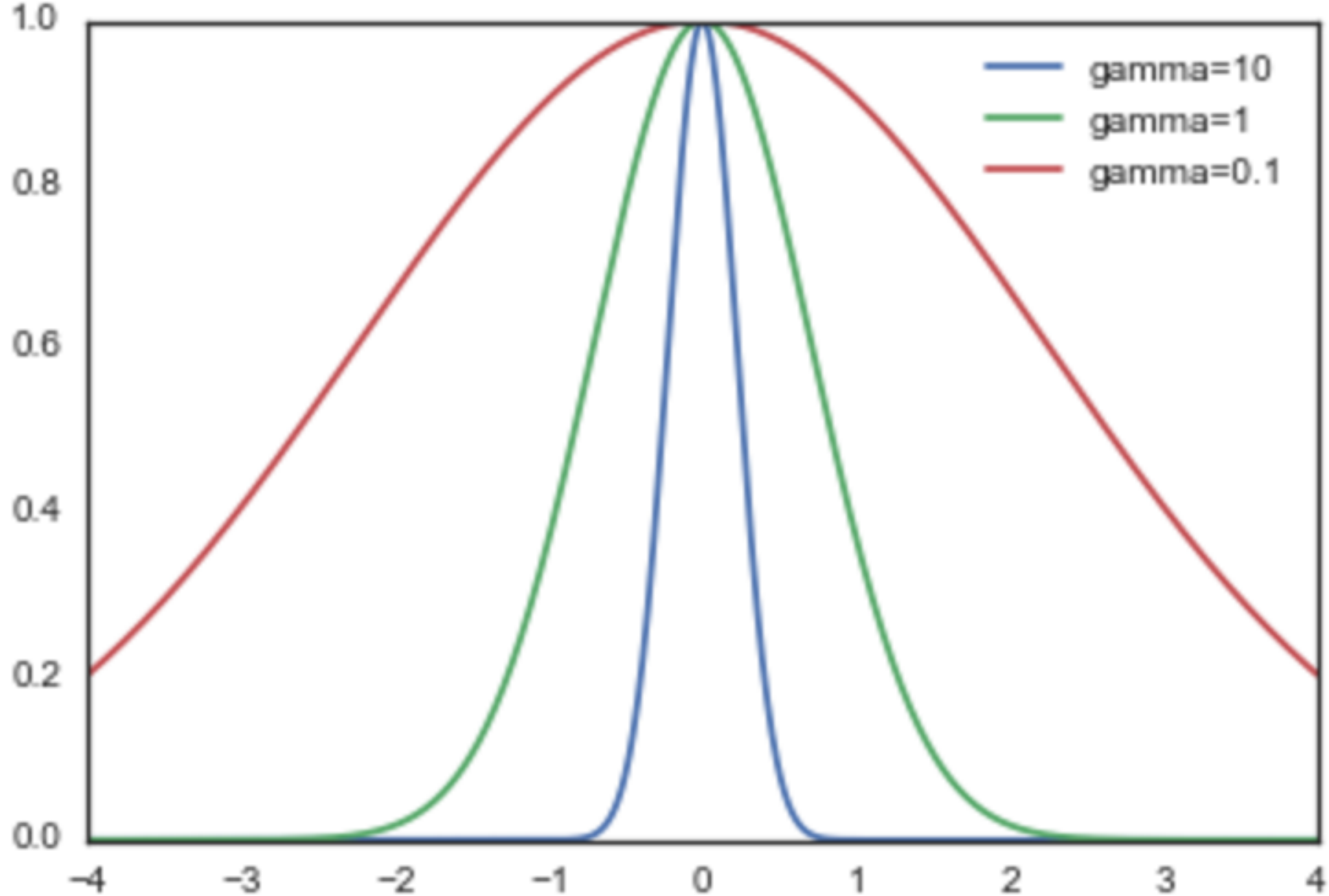
- Polynomial:

$$K(x, z) = (1 + x \cdot z)^s$$

- Radial basis function (RBF):

$$K(x, z) = \exp(-\gamma \|x - z\|^2)$$

RBF Kernel



So what is the excitement?

$$\max_\alpha \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j {x^{(i)}}^T x^{(j)}$$

$$\text{s.t. } \alpha_i \geq 0,\ i = 1, \ldots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

$$\arg\min_{w,b} \frac{1}{2} ||w||^2$$

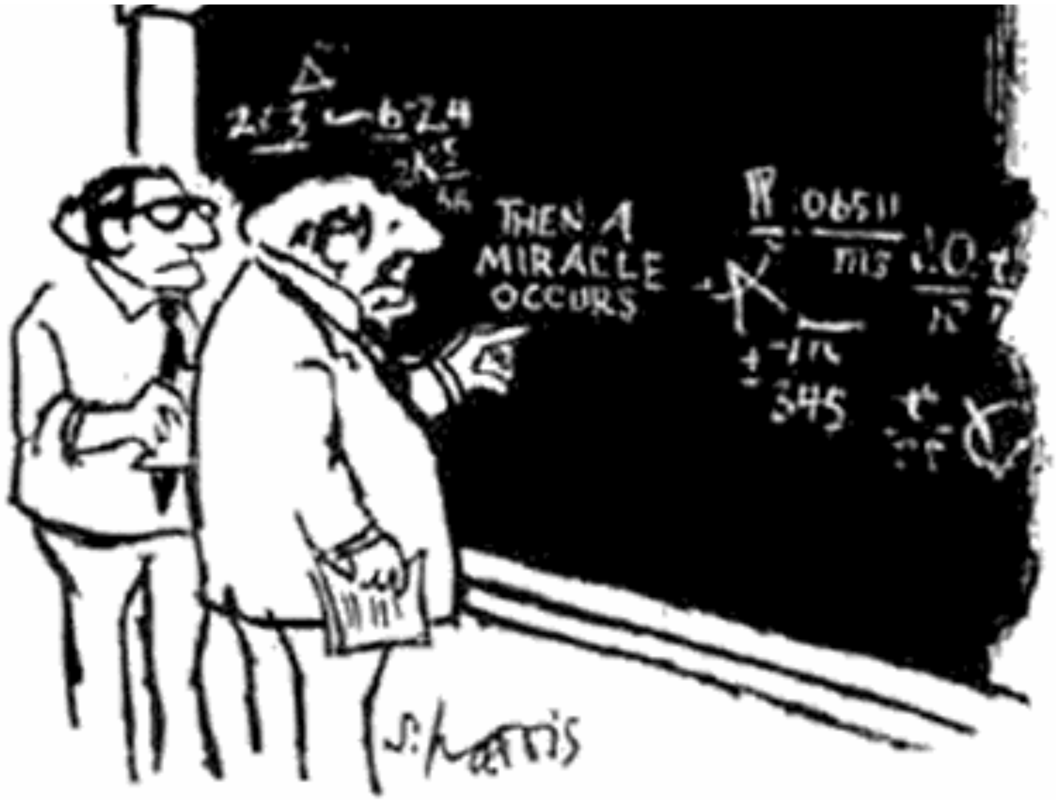
$$\text{s.t. } y^{(i)} (w^T x^{(i)} + b) \geq 1$$

So what is the excitement?

$\max_\alpha \sum$ $\quad {}^{(i)^T} x^{(j)}$

s.t. α_i

$\sum$



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

$\arg\mathrm{r}$

s.t. y

So what is the excitement?

$$\max_\alpha \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \boxed{{x^{(i)}}^T x^{(j)}}$$

$$\text{s.t. } \alpha_i \geq 0, \; i = 1, \ldots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

$$\boxed{K(x^{(i)}, x^{(j)})}$$

$$\arg\min_{w,b} \frac{1}{2} ||w||^2$$

$$\text{s.t. } y^{(i)} (w^T x^{(i)} + b) \geq 1$$

Prediction

$$w^T x + b = \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.$$

- Again we can use the kernel trick!
- Prediction speed depends on number of support vectors

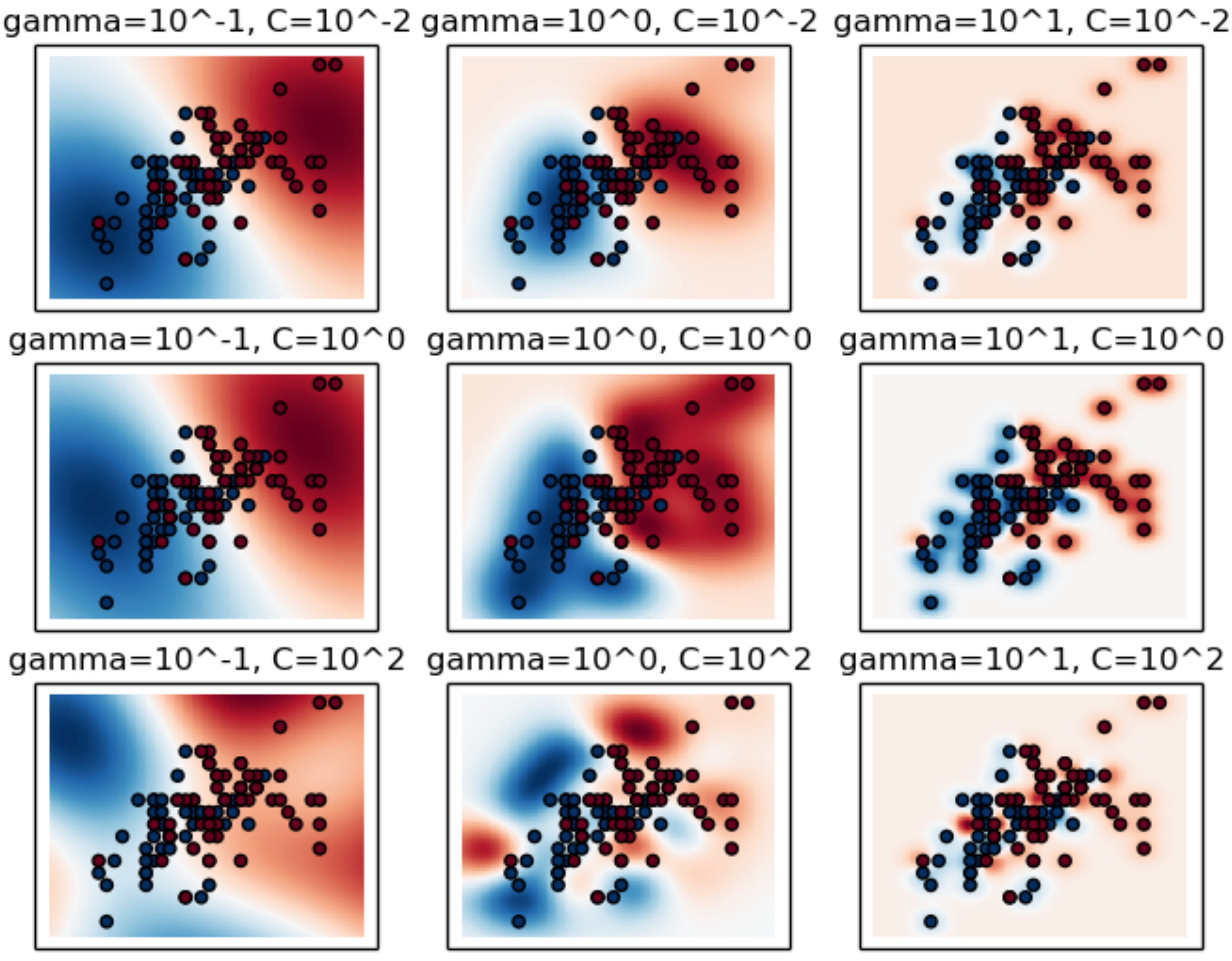
The Miracle Explained

- Andrew Ng does this really well

- http://cs229.stanford.edu/notes/cs229-notes3.pdf
- Course is also on Youtube, ItunesU, etc.

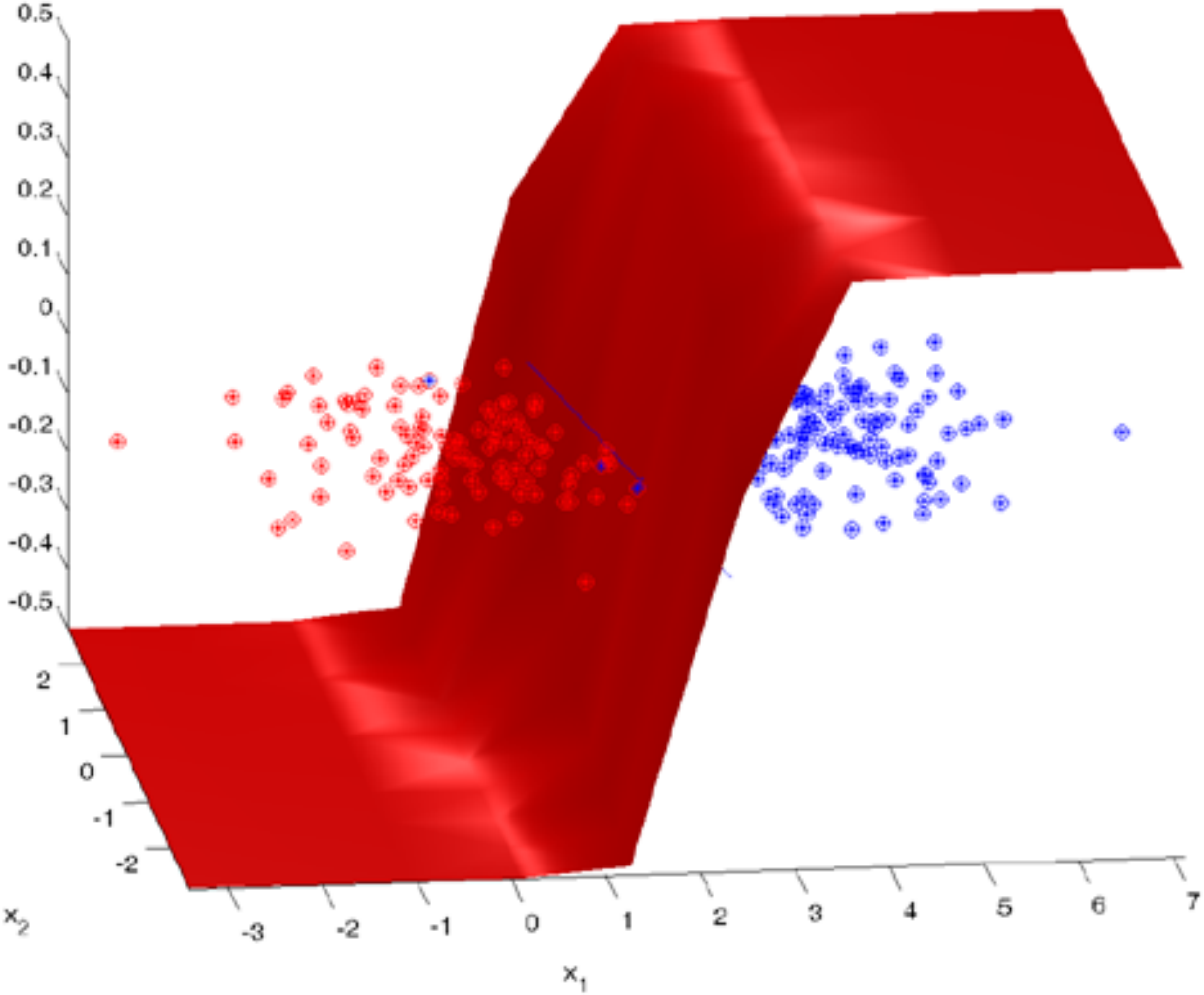
Kernel Trick for SVMs

- Arbitrary many dimensions
- Little computational cost
- Maximal margin helps with curse of dimensionality



http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

Logistic Regression Recap



http://strijov.com/sources/demoDataGen.php

KLR vs SVM

- The classification performance is very similar.
- Has limiting optimal margin properties
- Provides estimates of the class probabilities.
- Generalizes naturally to multiclass problems

http://web.stanford.edu/~hastie/TALKS/svm.pdf

KLR vs SVM

- KLR is computationally more expensive

  $O(N^3)$ versus $O(N^2m)$, where m is the number of support points.
- In noisy problems, m can be large, approx $N/2$.
- SVMs are hot right now, while logistic regression is a traditional statistical tool.

http://web.stanford.edu/~hastie/TALKS/svm.pdf

Tips and Tricks

- SVMs are not scale invariant
- Check if your library normalizes by default
- Normalize your data
  - mean: 0 , std: 1
  - map to [0,1] or [-1,1]
- Normalize test set in same way!

Tips and Tricks

- RBF kernel is a good default
- For parameters try exponential sequences
- Read:

    Chih-Wei Hsu et al., "A Practical Guide to Support Vector Classification", Bioinformatics (2010)