Support Vector Machines (SVM)

Samuel Brody

April 2010

Outline

Generative vs. Discriminative

Classification methods can be broken into two major categories:

Generative	Discriminative
$P(D\|C), P(C)$	$P(C\|D)$
deeper understanding: modeling the classes	pragmatic: focus on distinguishing
can generate data	can only classify

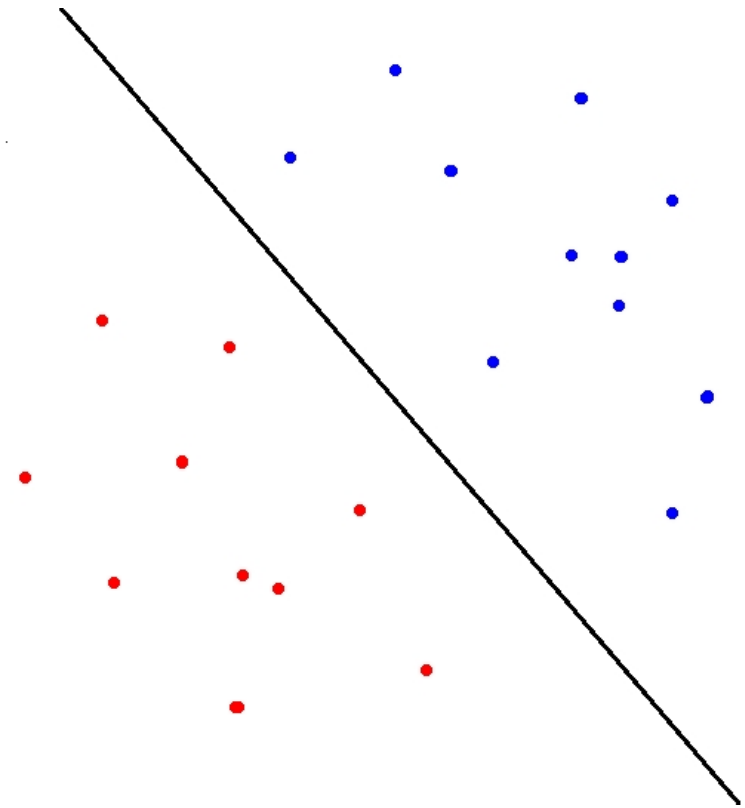
Which is which?

- Naive Bayes ? - generative
- Decision Trees ? - discriminative
- HMMs ? - generative
- K-Means Clustering ? - generative
- SVM ? - discriminative

Outline

The Idea

The Idea - Find a linear separator between classes:



Separating Hyperplane

- In 2 dimensions, the separator is a line:

  $y = a \cdot x + b \quad \Rightarrow \quad w_1 \cdot x_1 + w_2 \cdot x_2 + w_0 = 0$

- In 3 dimensions, it's a 2D plane:

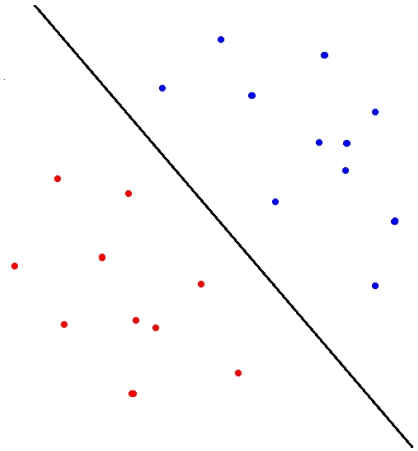
  $w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + w_0 = 0$

- In general, a separator in n dimensions is a $(n-1)$-dimensional hyperplane:

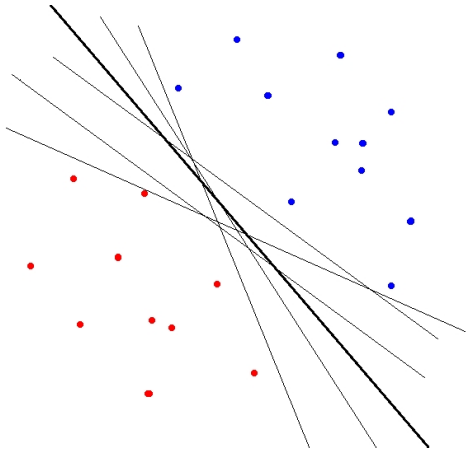
  $$\sum_{i=1}^{n} w_i \cdot x_i + w_0 = 0$$

Which Separator?

where there is one linear separator ...

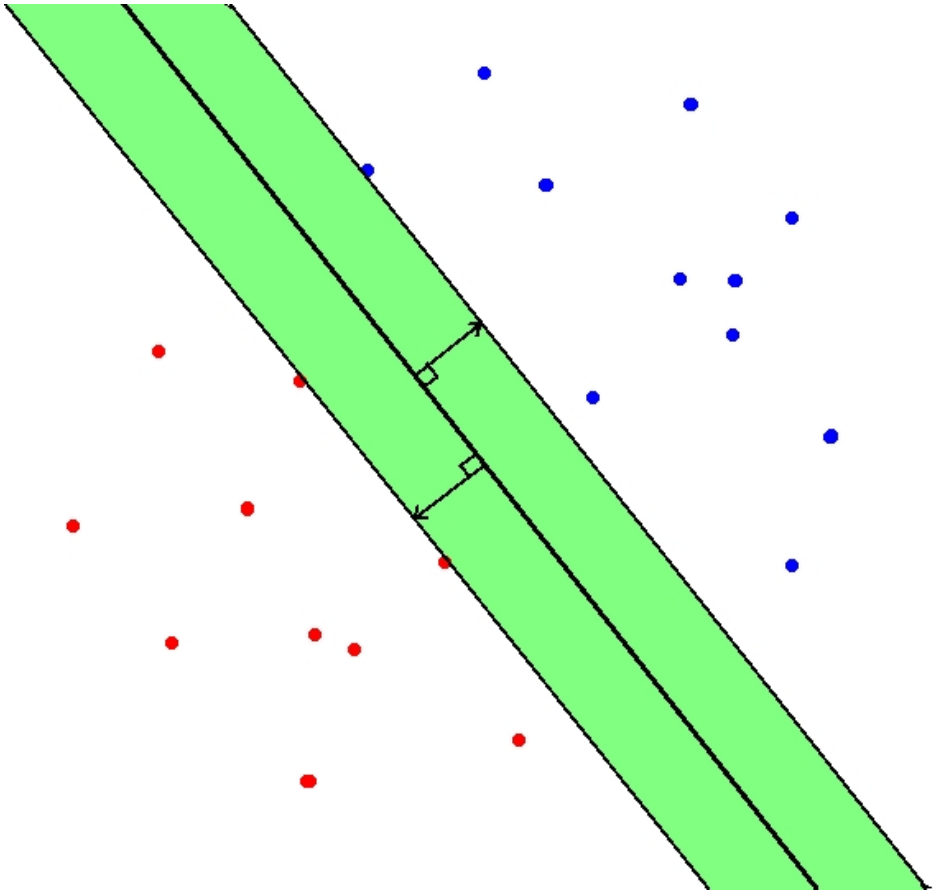


there are many



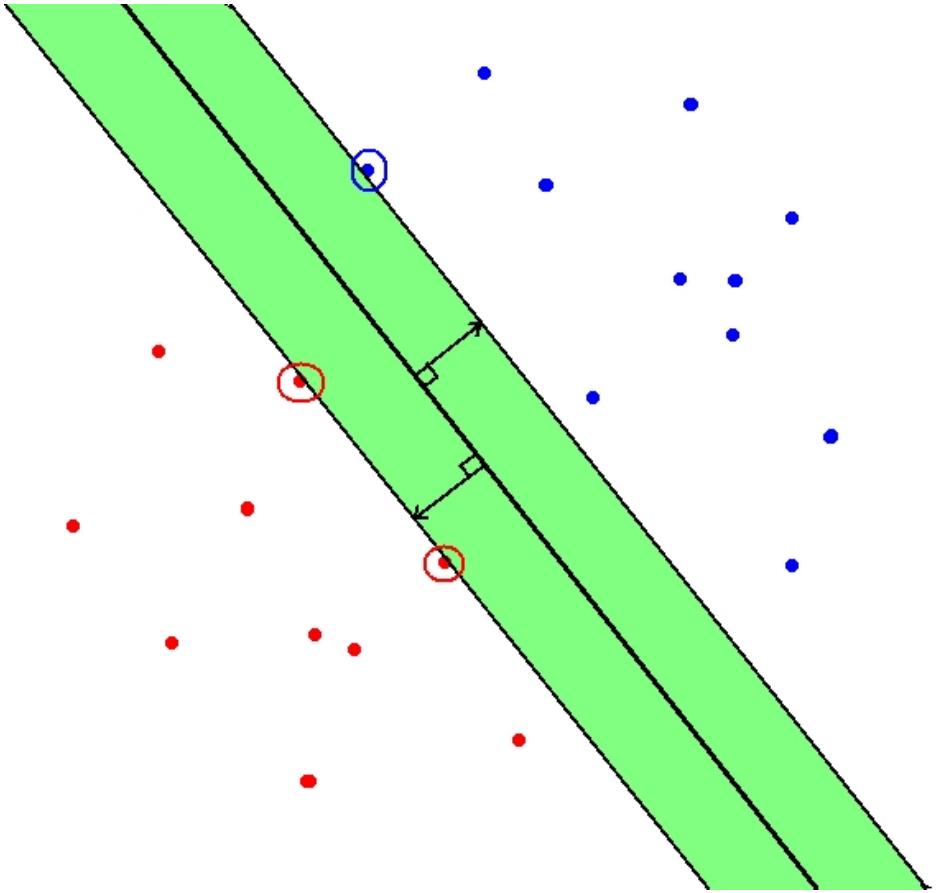
Margin

- find the separator with the largest margin
- not a “tight fit”
- allows for unseen test instances



Support Vectors

- The separator is defined by the outlying points of each class.
- These are the *support vectors*.
- SVM doesn't care about all the other points



SVM Objective:

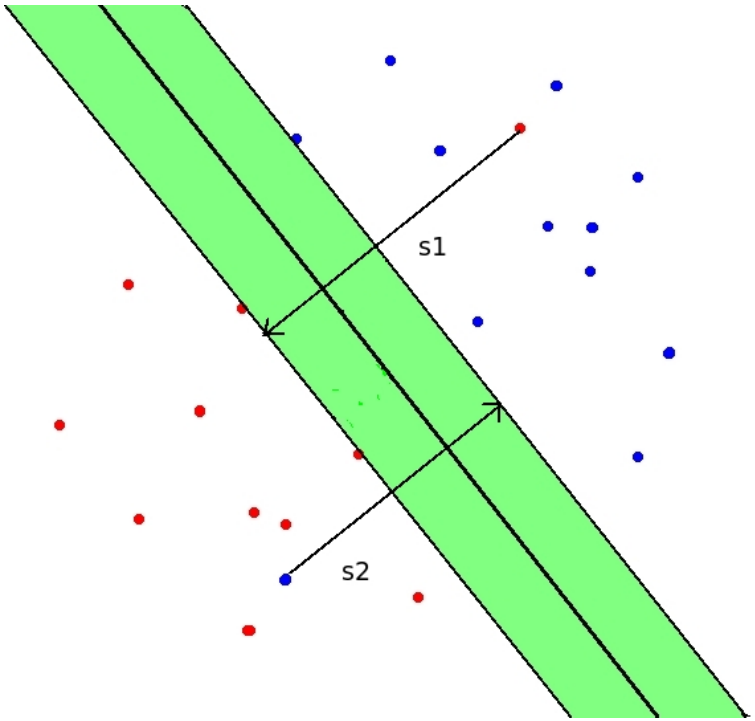
Find the hyperplane that separates the positive and negative points with the widest margin.

Non-separable Data

Q. What happens when that's not possible?

A. Introduce *slack variables*.

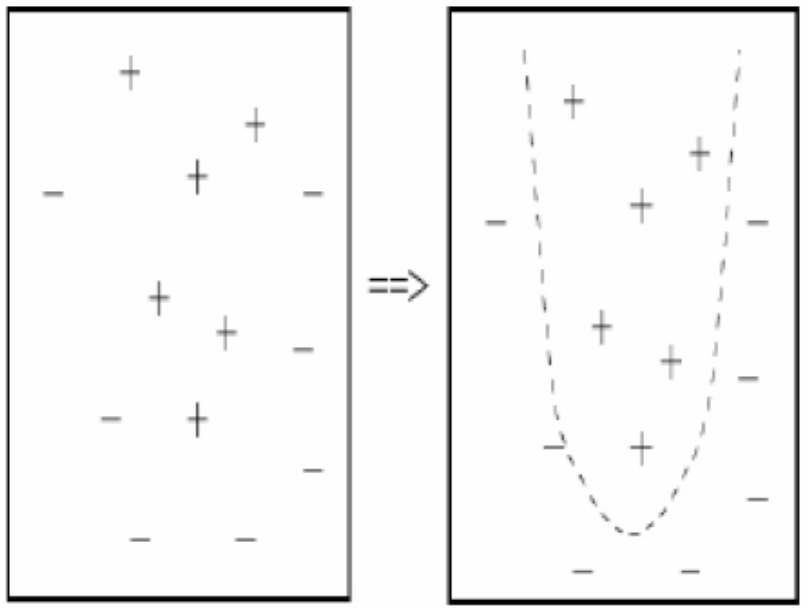
New Goal: Find the widest margin hyperplane, with minimal use of slack variables.



Non-linear Separators

Q. What happens when there's a separator, but it's not linear?

- the equation is:
  $x_2 = a \cdot x_1^2 + b \cdot x_1 + c$
- if we had an extra feature $x_3 = x_1^2$
  it would be linear:
  $x_2 = a \cdot x_3 + b \cdot x_1 + c$



Polynomial Mapping

- In order to find non-linear separators, we want to add extra features/dimensions.
- For example, all polynomials with degree ≤ 2:
- $map(\{x_1, x_2, x_3\}) = \{x_1, x_2, x_3, x_1^2, x_1 \cdot x_2, x_1 \cdot x_3, x_2^2, x_2 \cdot x_3, x_3^2\}$
- Our function maps from a 3-dimensional problem to a 9-dimensional one.

Kernel Functions

- To find non-linear separators, we add extra features that are functions of the originals.

  Pro: allows us to find non-linear separators

  Con: huge overhead in space and computation

- Kernels are mapping functions with special properties that circumvent the problem.

Kernel Example

- Suppose we have a 2D space, with instances $X = \{x_1, x_2\}$
- The separator will have the form $W = \{w_0, w_1, w_2\}$
- In order to learn a non-linear separator, we'd like to map to a polynomial space:

  e.g.*, $map(\{x_1, x_2\}) = \{x_1^2, \sqrt{2} \cdot x_1 \cdot x_2, x_2^2\}$

- The main calculation in SVM is $\sum_{i=1}^{n} w_i \cdot x_i + w_0$

* note: multiplying by the constant $\sqrt{2}$ helps with the kernel, and doesn't affect the ability to separate.

Kernel Example - cont.

- $map(\{x_1, x_2\}) = \{x_1^2, \sqrt{2} \cdot x_1 \cdot x_2, x_2^2\}$
- The main calculation in SVM is $\sum_{i=1}^{n} w_i \cdot x_i + w_0$
- To perform this calculation in the mapped space, we need to:
  - map W and X throught the map function (3 multiplications each)
  - do the calculation with $n = 3$ (another 3 multiplications, 9 in total)
- instead, we use the kernel function $K(X, W) = (x_1 \cdot w_1 + x_2 \cdot w_2)^2$
- this produces the same result, with only 2 multiplications and a squaring operation

Kernel Functions

- kernels are mapping functions with special properties
- allow learning and classifying without ever explicitly mapping to a higher dimension
- give the advantages of higher dimensions, without the overhead

Support Vector Machines

- extremely powerful (discriminative) classifier
- learns from edge cases
- finds a linear separator

- can be extended to handle:
  - a few misfits via slack variables
  - non linear separators via kernel functions