Mathematical techniques in data science

Lecture 9: SVM and kernel trick

March 8th, 2019

Chapter 9: Support Vector Machines

- Maximal Margin Classifier
- Support Vector Classifiers
- Support Vector Machines

Hyperplane

- In a p-dimensional space, a hyperplane is an affine subspace of dimension p.
- In two dimensions, a hyperplane is defined by the equation

$$\beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} = 0$$

- In p dimensions:

$$\beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \ldots + \beta_p x^{(p)} = 0$$

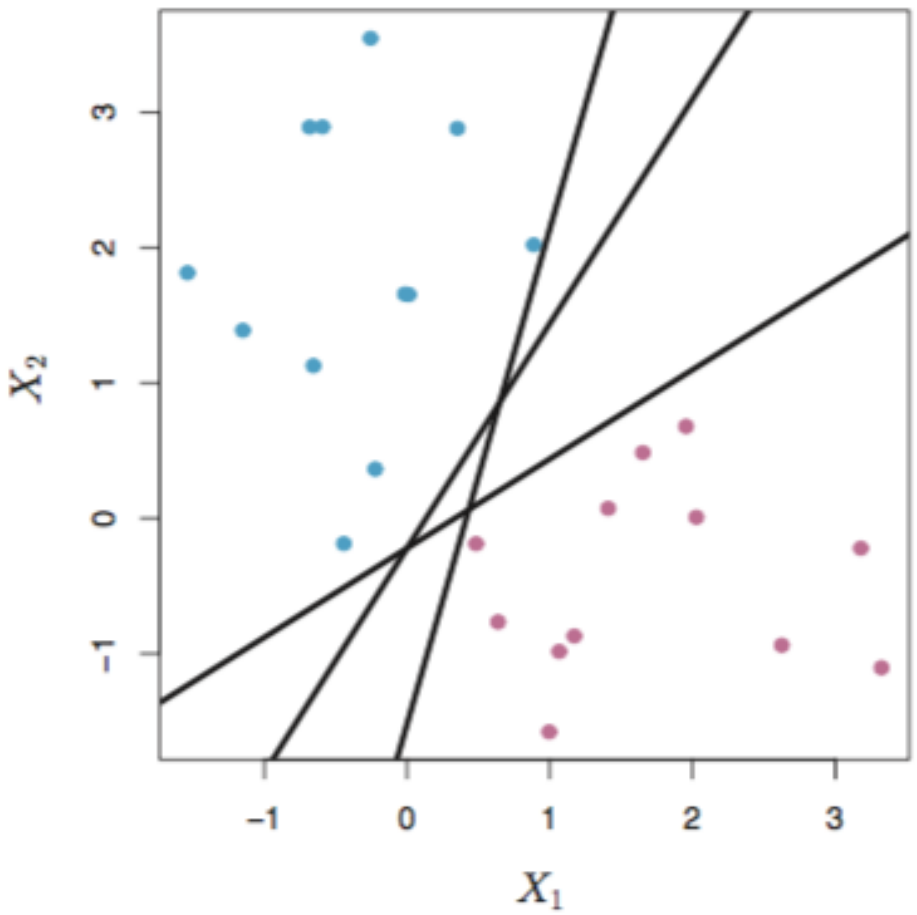
or alternatively

$$\beta_0 + \beta^T x = 0, \quad \text{where } \beta \in \mathbb{R}^p$$

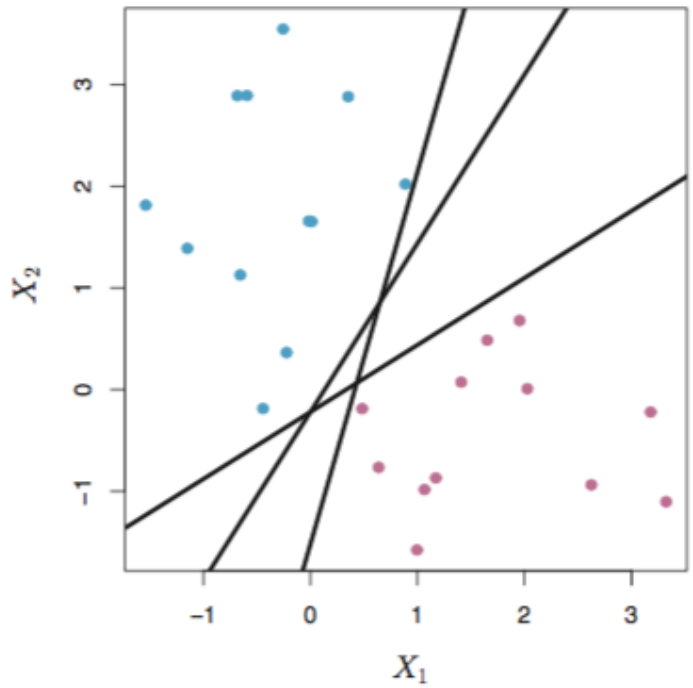
Separating hyperplane

Suppose we have data with label $\{-1, 1\}$, we want to separate the data using a hyperplane

$$y = \text{sign}(\beta_0 + \beta^T x)$$



Separating hyperplane



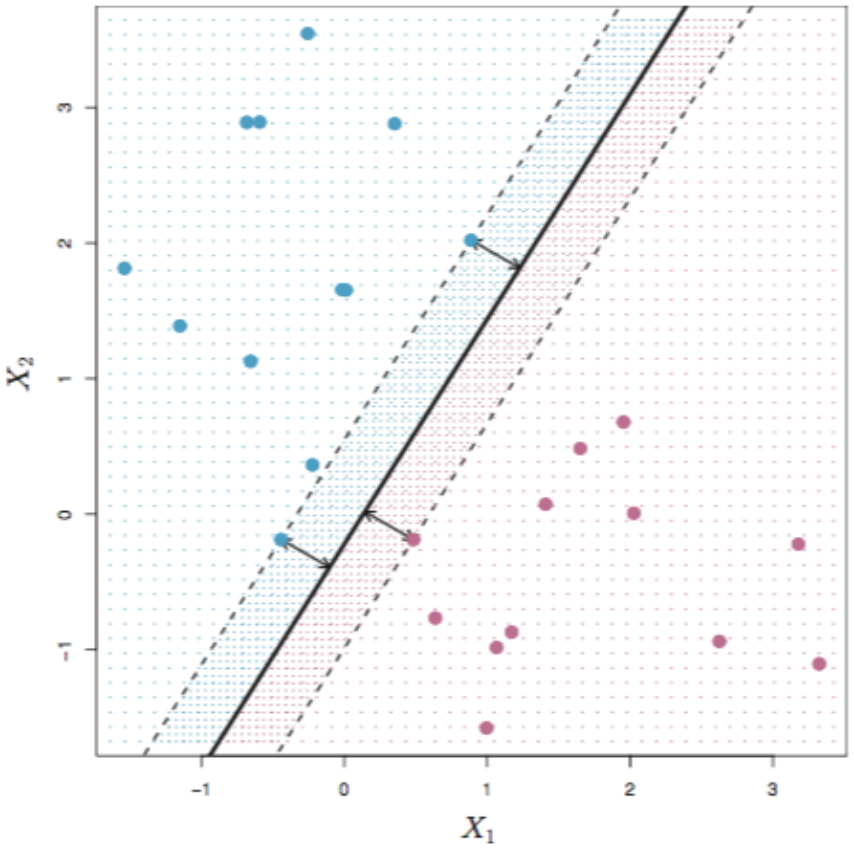
Problems:

- Separating hyperplane may not exist
- Assume that the data are perfectly separable by a hyperplane $\rightarrow$ then there might exist an infinite number of such hyperplanes

Maximal Margin Classifier

Maximal Margin Classifier

- Assume that the data are perfectly separable by a hyperplane
- The minimal distance from the data to the hyperplane is call the *margin*
- Maximal margin hyperplane: the separating hyperplane that is farthest from the training observations



Maximal Margin Classifier: formulation

- Given a set of n training observations $x_1, \ldots, x_n \in \mathbb{R}$ and associated class labels $y_i \in \{-1, 1\}$
- Maximal margin hyperplane:

$$\max_{\beta_0, \beta, M} M$$
$$\text{subject to } \|\beta\| = 1$$
$$\text{and } y_i(\beta_0 + \beta^T x_i) \geq M \quad \forall i = 1, \ldots, n.$$

MMC: Alternative form

$$\max_{\beta_0,\beta,M} M$$

$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq M\|\beta\| \quad \forall i = 1, \ldots, n.$$

- If we rescale (β_0, β) such that $M\|\beta\| = 1$, then the optimization problem becomes

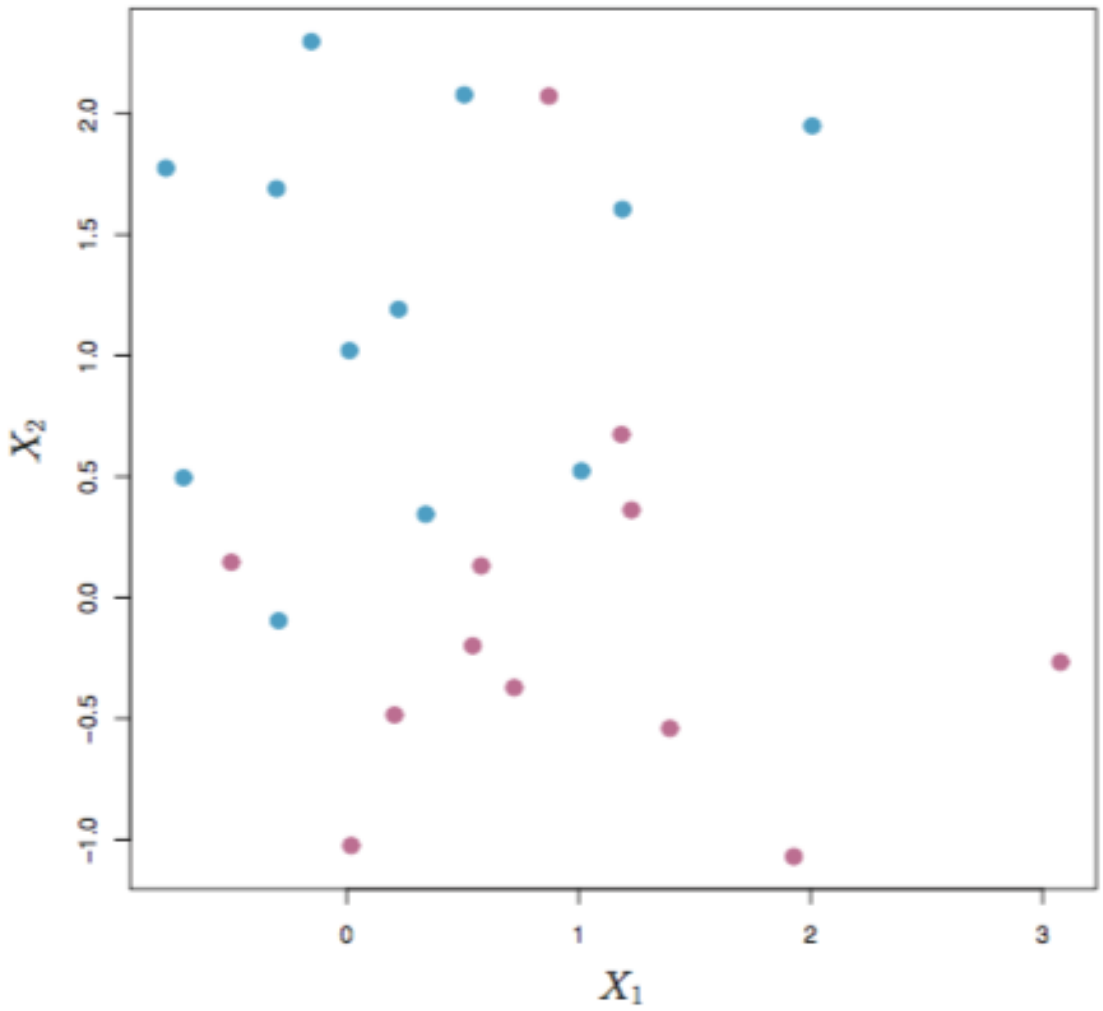
$$\min_{\beta_0,\beta} \|\beta\|^2$$

$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq 1 \quad \forall i = 1, \ldots, n.$$

- This is a convex optimization problem with a quadratic object and linear constraints

Support Vector Classifiers

Realistically, data are not separable by hyperplanes



MMC is not robust to noises

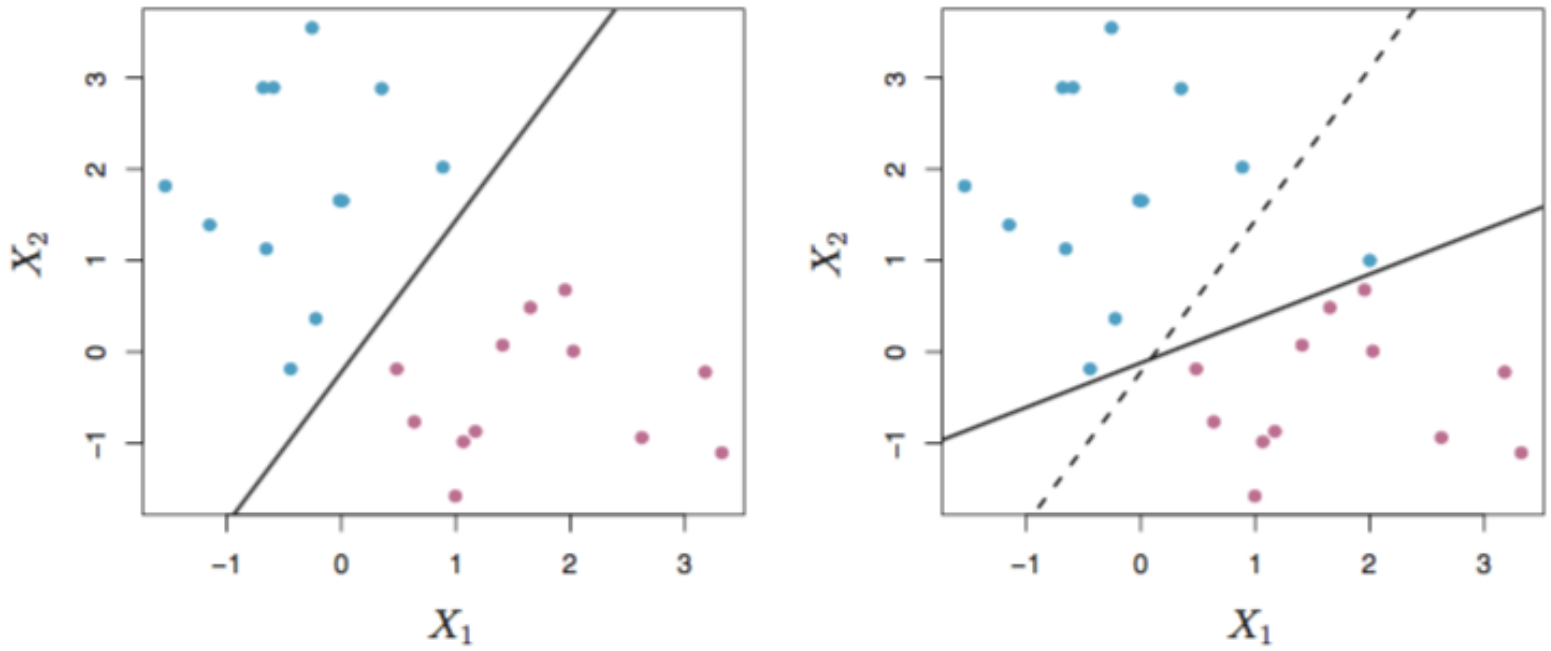


FIGURE 9.5. Left: *Two classes of observations are shown in blue and in purple, along with the maximal margin hyperplane.* Right: *An additional blue observation has been added, leading to a dramatic shift in the maximal margin hyperplane shown as a solid line. The dashed line indicates the maximal margin hyperplane that was obtained in the absence of this additional point.*

Support Vector Classifier

The hyperplane is chosen to correctly separate most of the training observations into the two classes, but may mis-classify a few observations

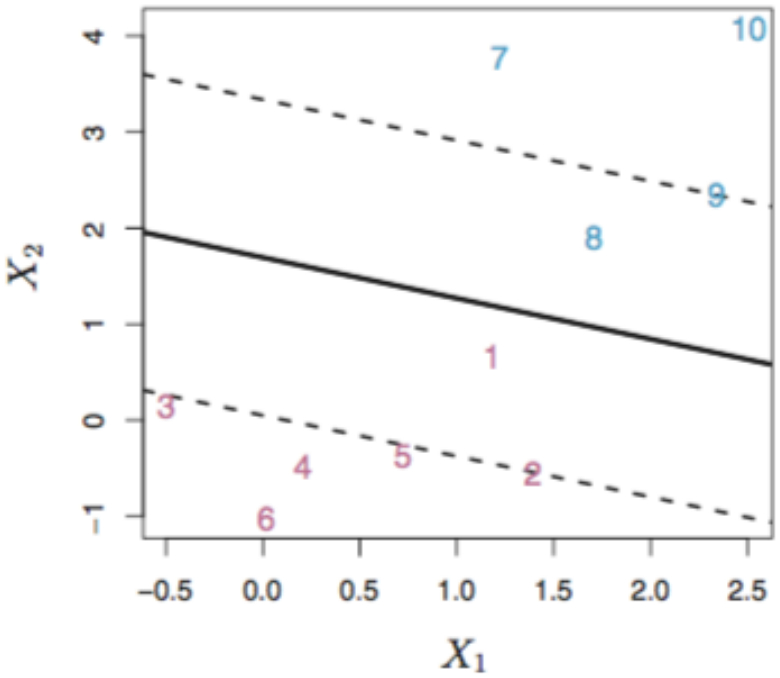
$$\max_{\beta_0,\beta,M,\epsilon_1,\epsilon_2,\ldots,\epsilon_n} M$$

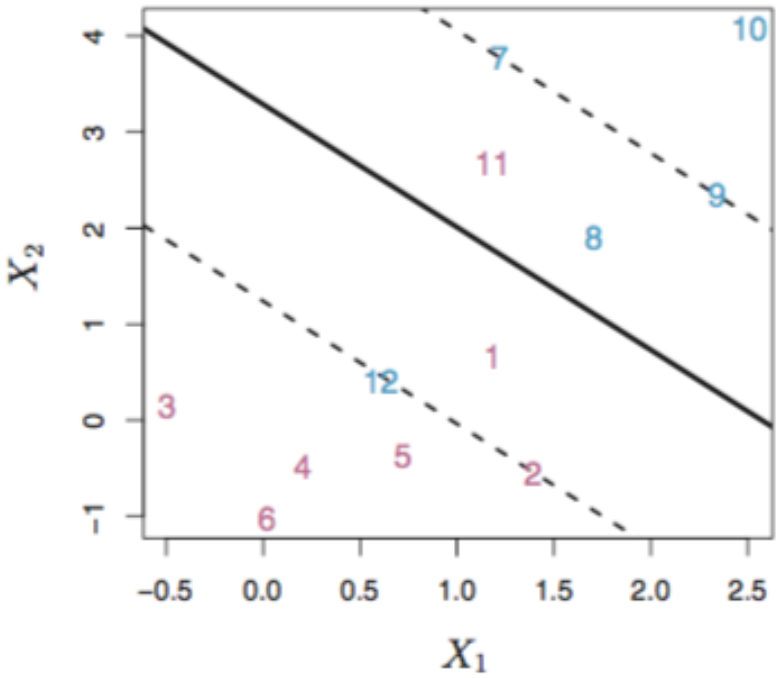
$$\text{subject to } \|\beta\| = 1$$

$$y_i(\beta_0 + \beta^T x_i) \geq M(1-\epsilon_i) \quad \forall i = 1,\ldots,n$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C.$$

Support Vector Classifier





SVC: alternative form

The hyperplane is chosen to correctly separate most of the training observations into the two classes, but may misclassify a few observations

$$\min_{\beta_0,\beta,\epsilon_1,\epsilon_2,\ldots,\epsilon_n} \|\beta\|^2$$

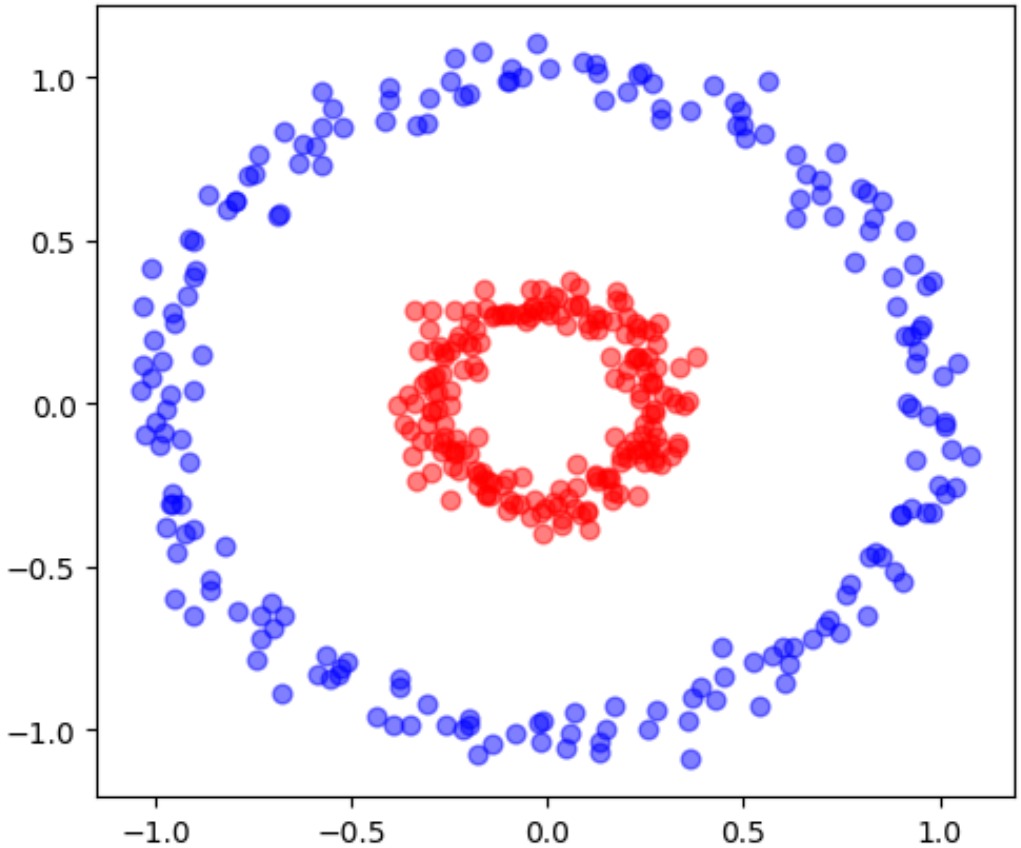
$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq (1 - \epsilon_i) \quad \forall i = 1, \ldots, n$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C.$$

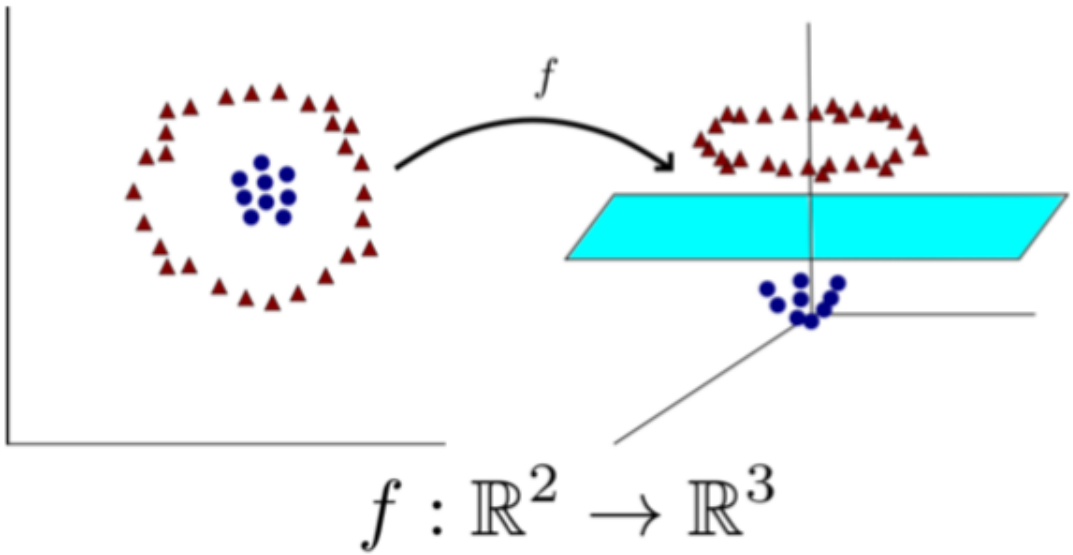
Can be solved using standard optimization packages.

Support Vector Machine

Realistically, the boundary may be non-linear



Idea: map the learning problem to a higher dimension



$$f(x, y) = (x, y, x^2 + y^2)$$

Idea: map the learning problem to a higher dimension

More rigorously,

$$f(x, y) = (x, y, x^2, y^2, xy)$$

A hyperplane on $\mathbb{R}^5$, modeled by the equation $\beta_0 + \beta^T x = 0$ will classify the points based on the sign of

$$\beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy$$

This corresponds to a quadratic boundary on the original space $\mathbb{R}^2$

How to solve SVM's optimization

Problem:

$$\min_{\beta_0,\beta} \|\beta\|^2$$

$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq 1 \qquad \forall i = 1, \ldots, n.$$

Alternative form

Lagrange multiplier:

$$L(\beta, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\beta_0 + \beta^T x_i) - 1], \quad \text{where } \alpha_i \geq 0$$

New problem:

$$\min_{\beta} \max_{\alpha} L(\beta, \alpha)$$

Idea:

- Consider a game with two players, Mindy and Max,
- Mindy goes first, choosing β. Max, observing Mindy's choice, selects α to maximize $L(\beta, \alpha)$
- Mindy, aware of Max's strategy, makes her initial choice to minimize $L(\beta, \alpha)$

Minimax theory

Minimax theory: for some class of functions:

$$\min_{\beta} \max_{\alpha} L(\beta, \alpha) = \max_{\alpha} \min_{\beta} L(\beta, \alpha)$$

Recall:

$$L(\beta, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\beta_0 + \beta^T x_i) - 1], \quad \text{where } \alpha_i \geq 0$$

Question: Given α, what is the optimal value of β?

Recall:

$$L(\beta, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\beta_0 + \beta^T x_i) - 1], \quad \text{where } \alpha_i \geq 0$$

Question: Given α, what is the optimal value of β?

$$\frac{\partial L}{\partial \beta^{(j)}} = \beta^{(j)} - \sum_{i=1}^{n} \alpha_i y_i x_i^{(j)}$$

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n} \alpha_i y_i$$

Conclusion

$$\beta^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Conclusion

$$\beta^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Put this back into the expression of L:

$$\max_{\alpha \geq 0} \quad \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i y_i x_i^T x_j,$$

Conclusion: To solve the MMC's optimization problem, we just need to have information about

$$x_i^T x_j = \langle x_i, x_j \rangle \quad \forall i, j$$

Note that

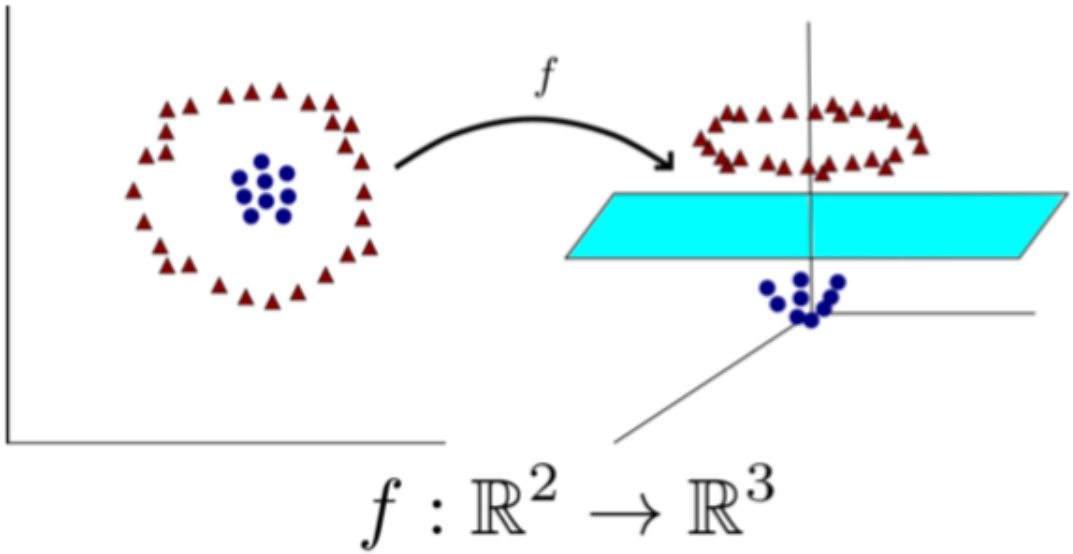
$$\beta^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

Decision function

$$\beta_0 + \sum_{i=1}^{n} \alpha_i^* y_i \langle x_i, x \rangle$$

...back to SVM

Idea: map the learning problem to a higher dimension



When mapping x to $f(x)$ in a higher dimensions, make sure you can compute

$$\langle f(x_i), f(x_j) \rangle \qquad \forall i, j$$

Previous lecture

More rigorously,

$$f(x, y) = (x, y, x^2, y^2, xy)$$

A hyperplane on $\mathbb{R}^5$, modeled by the equation $\beta_0 + \beta^T x = 0$ will classify the points based on the sign of

$$\beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy$$

This corresponds to a quadratic boundary on the original space $\mathbb{R}^2$

A more careful mapping

Define

$$f(x, y) = (1, \sqrt{2}x, \sqrt{2}y, x^2, y^2, \sqrt{2}xy)$$

A hyperplane on $\mathbb{R}^6$, modeled by the equation $\beta_0 + \beta^T x = 0$ will classify the points based on the sign of

$$\beta_0 + \beta_1 + \beta_2 x + \beta_3 y + \beta_4 x^2 + \beta_5 y^2 + \beta_6 xy$$

This corresponds to a quadratic boundary on the original space $\mathbb{R}^2$

A more careful mapping

Moreover:

$$\begin{aligned}\langle f(x,y), f(u,v)\rangle &= 1 + 2xu + 2yv + x^2u^2 + x^2v^2 + 2xyuv \\ &= (1 + xu + yv)^2 \\ &= (1 + \langle (x,y),(u,v)\rangle)^2\end{aligned}$$

In other the words,

$$K(x_i, x_j) = \langle f(x_i), f(x_j)\rangle = (1 + x_i^T x_j)^2$$

can be computed quite easily.

SVM on a higher dimensional space

Recall that in order to solve the optimization of SVM on the original space, we need to optimize

$$\max_{\alpha \geq 0} \quad \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i y_i x_i^T x_j,$$

If we want to do the same thing with the mapped data

$$\max_{\alpha \geq 0} \quad \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i y_i K(x_i, x_j),$$

Bonus: we don't need to know the form of f at all!

The kernel trick

We don't need to know the form of f, only need

$$K(x, y) = \langle f(x_i), f(x_j) \rangle$$

Question: Given $K : \mathbb{R}^p \times \mathbb{R}^p$, when can we guarantee that

$$K(x, y) = \langle h(x_i), h(x_j) \rangle$$

for some function h?

Kernel: condition

Question: Given $K : \mathbb{R}^p \times \mathbb{R}^p$, when can we guarantee that

$$K(x, y) = \langle h(x_i), h(x_j) \rangle$$

for some function h?

Definition

Let X be a set. A symmetric kernel $K : X \times X \to R$ is said to be a positive definite kernel if the matrix

$$[K(x_i, x_j)]_{i,j=1}^n$$

is positive semi-definite for all $x_1, \ldots, x_n$ and $n \in \mathbb{N}$, i.e.

$$\sum_{i,j} K(x_i, x_j) c_i c_j \geq 0$$

for any $c \in \mathbb{R}^n$.

Popular kernels

- Polynomials

$$K(x, u) = [1 + \langle x, u \rangle]^d$$

- RBF (Gaussian) kernels

$$K(x, u) = e^{-\gamma \|x-u\|^2}$$

- Neural network

$$K(x, u) = tanh(\kappa_1 \langle x, u \rangle + \kappa_2)$$

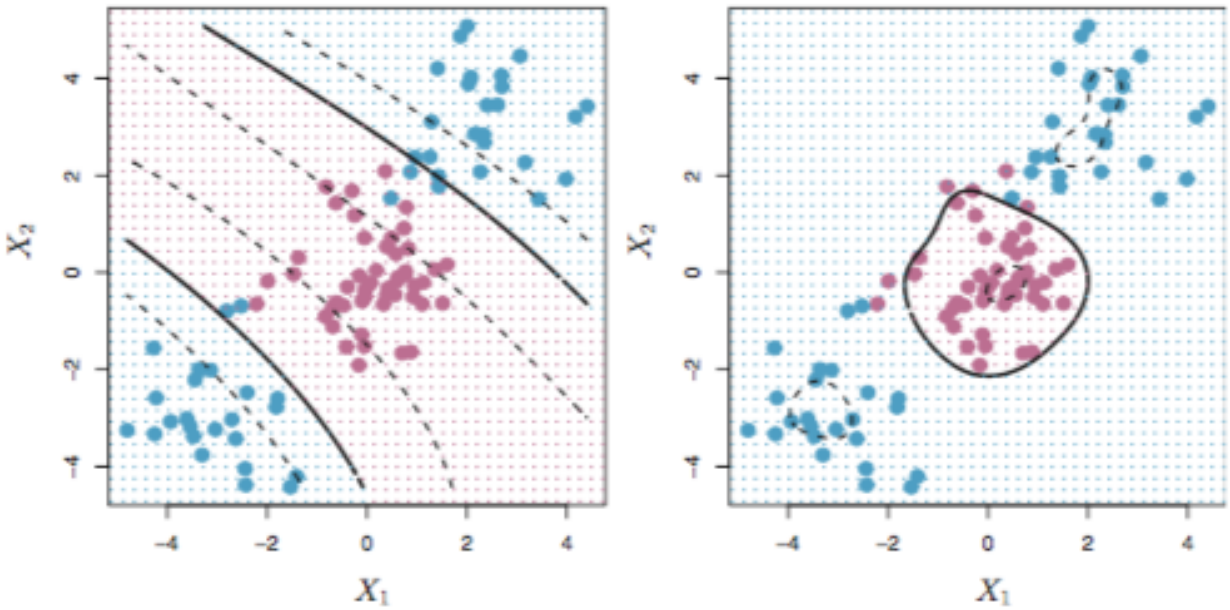


FIGURE 9.9. Left: *An SVM with a polynomial kernel of degree 3 is applied to the non-linear data from Figure 9.8, resulting in a far more appropriate decision rule.* Right: *An SVM with a radial kernel is applied. In this example, either kernel is capable of capturing the decision boundary.*