IST 5535: Machine Learning Algorithms and Applications

Langtao Chen, Spring 2021

Support Vector Machines

Reading

- Support vector machine: book chapter 9

Outline

- What is a hyperplane?
- Maximal margin classifier
- Support vector classifier
- Support vector machine
- Extension to multi-class classification

What is a Hyperplane?

- A hyperplane in a *p*-dimensional space is a flat affine subspace of dimension *p-1*.
  - In two dimensions, a hyperplane is a line;
  - In three dimensions, a hyperplane is a plane.

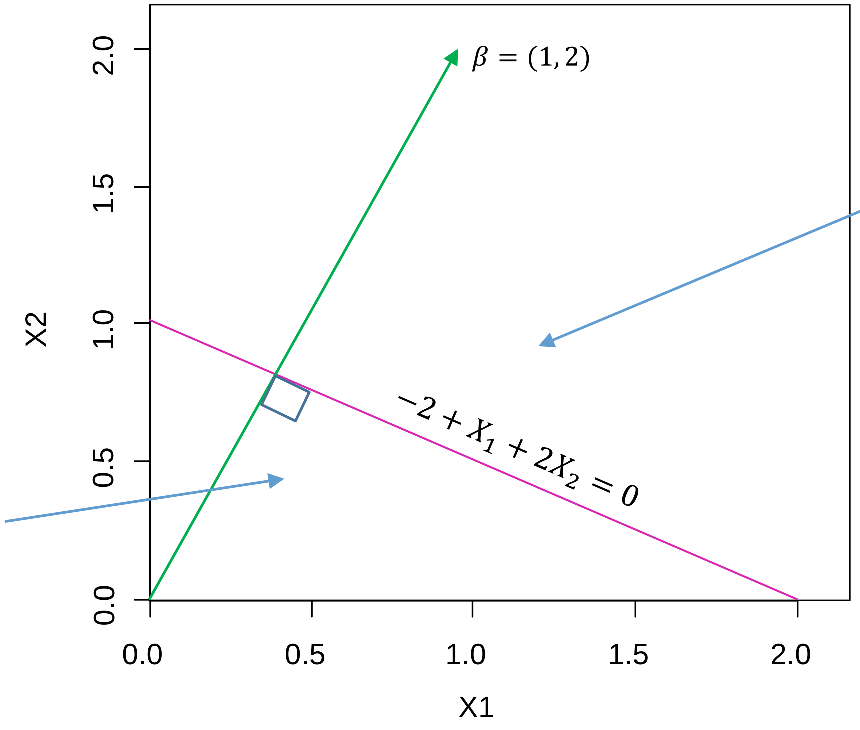
- Mathematical definition: In a *p*-dimensional space, a hyperplane is defined by:

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

  - If $\beta_0 = 0$, the hyperplane passes through the origin;
  - The normal vector $\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T$ is orthogonal to the surface of the hyperplane.

Example

- A hyperplane $-2 + X_1 + 2X_2 = 0$ in a two dimensional space.



In this region, $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$

In this region,
$\beta_0 + \beta_1 X_1 + \beta_2 X_2 < 0$

Classification Using a Separating Hyperplane

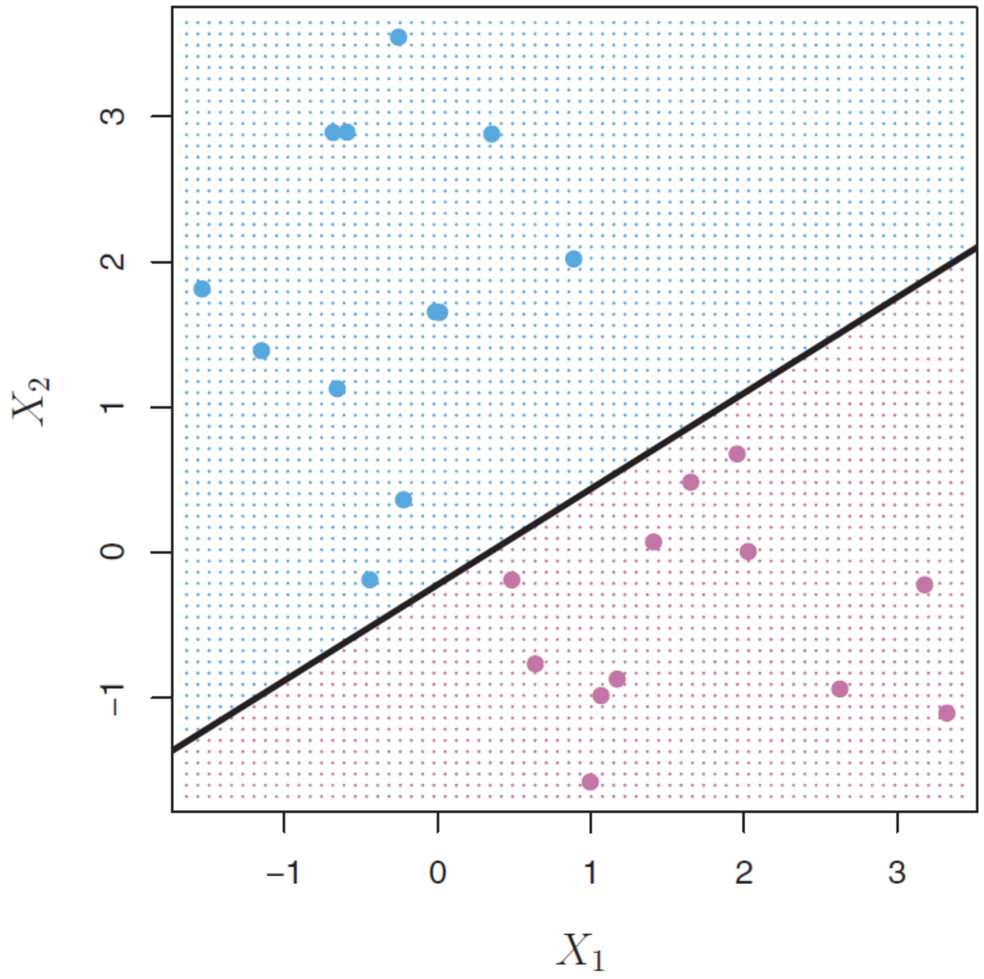
- Find a hyperplane that:

$$\begin{cases} \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} > 0, \text{if } y_i = 1 \\ \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} < 0, \text{if } y_i = -1 \end{cases}$$

Or more succinctly,

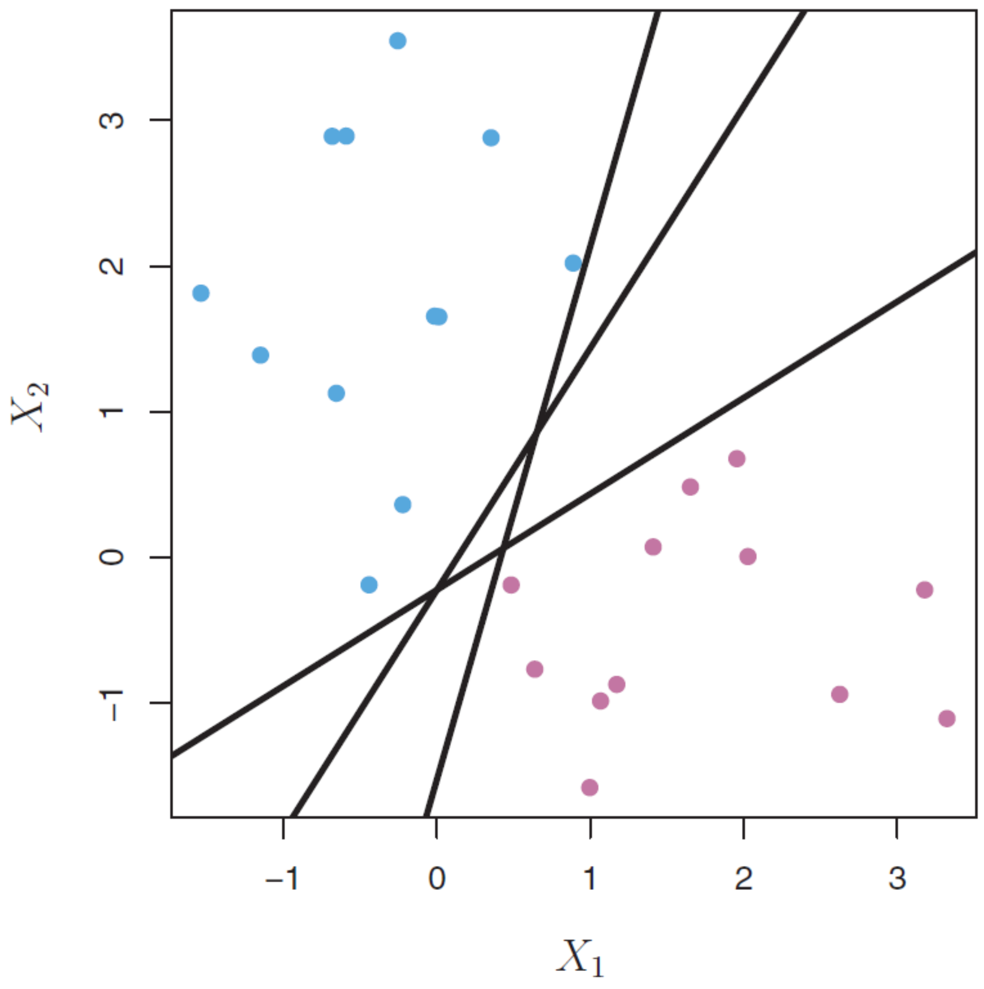
$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) > 0$$

That is, the hyperplane can ***perfectly*** separate the two classes ($y_i = 1$ and $y_i = -1$)



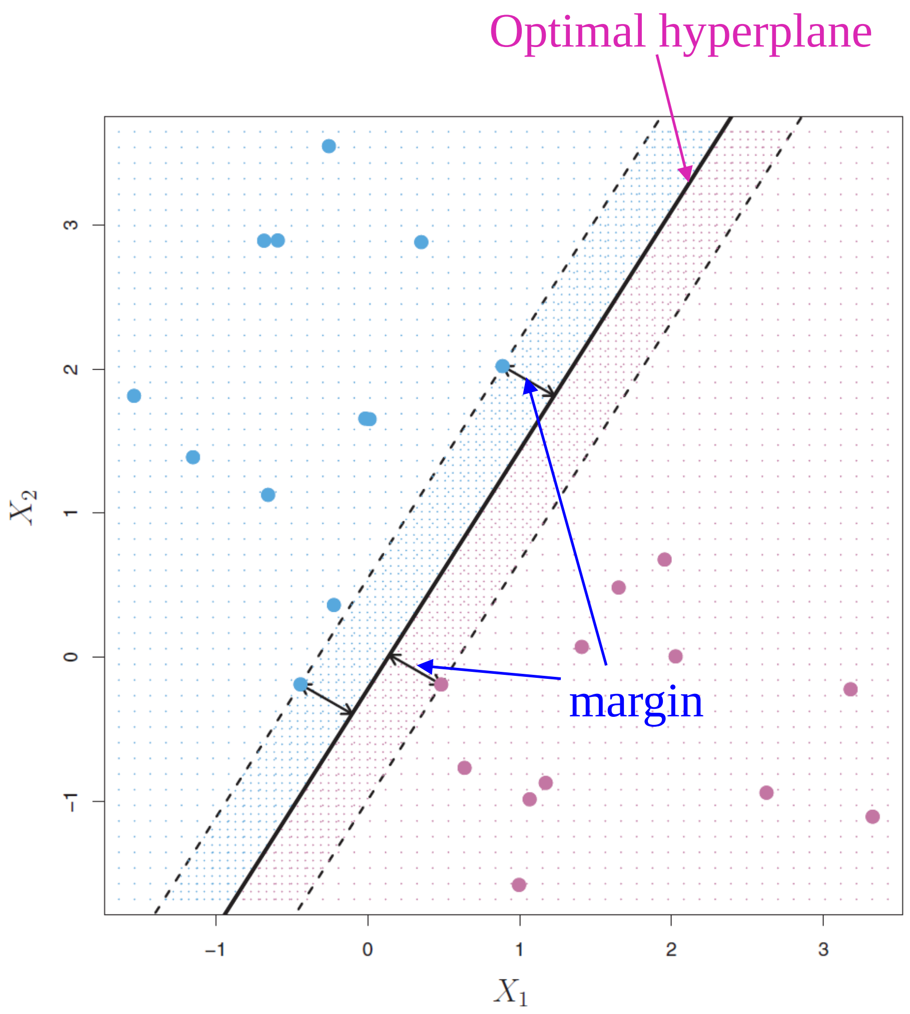
Which is the Optimal Hyperplane?

- An infinite number of hyperplanes can do the job.



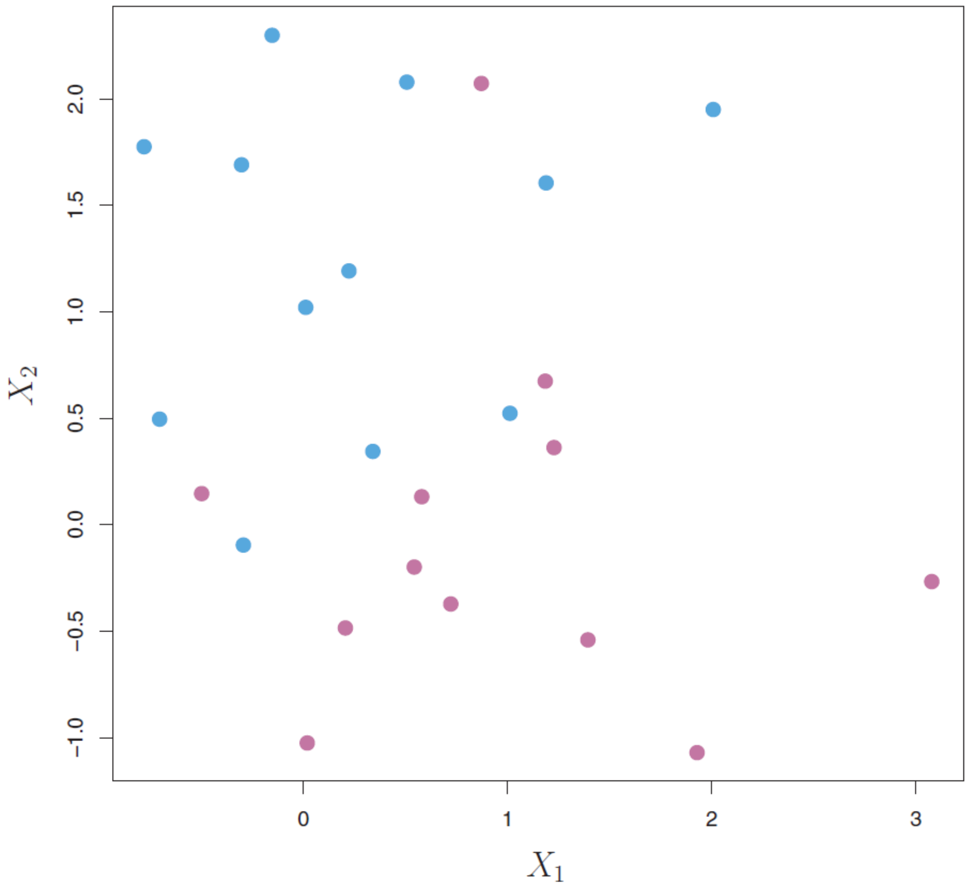
The Maximal Margin Classifier

- The maximal margin hyperplane (or optimal separating hyperplane) is the one that maximizes the margin between training data and the decision boundary.
- Maximizing the margin minimizes the chance of misclassification of new data, so that SVM would have the “best” predictive power.
- The data points closest to the margins (on the dashed lines) are called *support vectors*.
- The maximum margin classifier only depends on the support vectors in the training data.



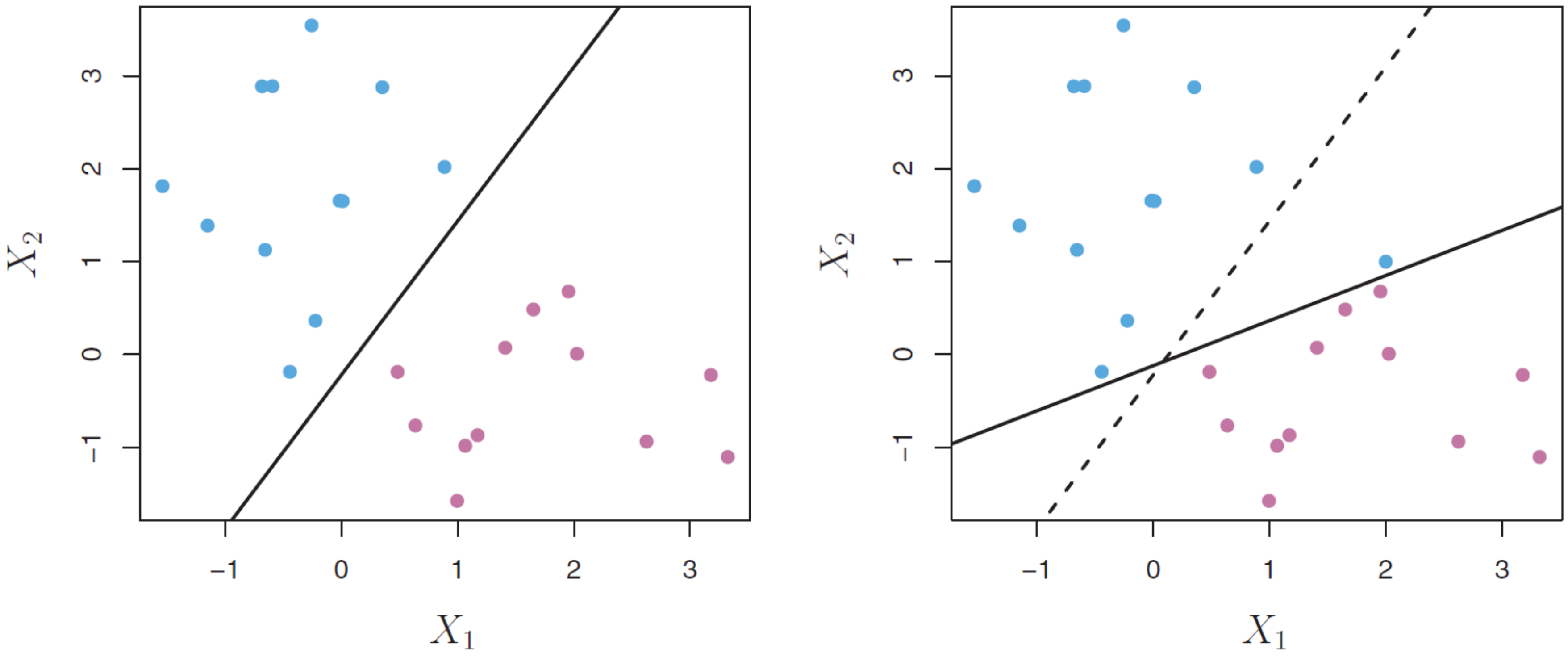
The Problem of Maximum Margin Classifier

- Case 1: A separating hyperplane does not exist!



The Problem of Maximum Margin Classifier

- Case 2: The maximum margin classifier is not robust.



Adding a new data point dramatically change the optimal separating hyperplane.

The Support Vector Classifier

- The support vector classifier is a generalization of the maximum margin classifier.
  - It adopts a soft margin that may not perfectly separate the classes;
  - Greater robustness to individual observations;
  - Better classification of *most* of the training observations.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

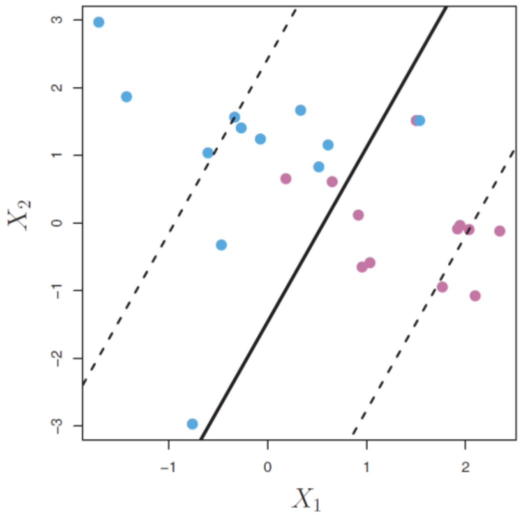
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

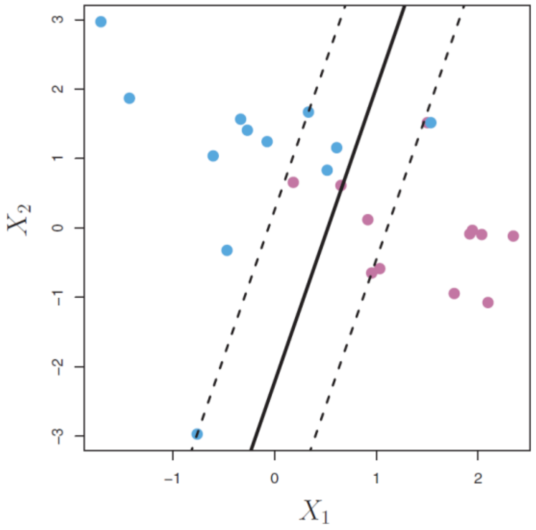
$$\epsilon_i \geq 0, \ \sum_{i=1}^{n} \epsilon_i \leq C,$$

C is a nonnegative tuning parameter, usually chosen by cross-validation.

Regularization Parameter C

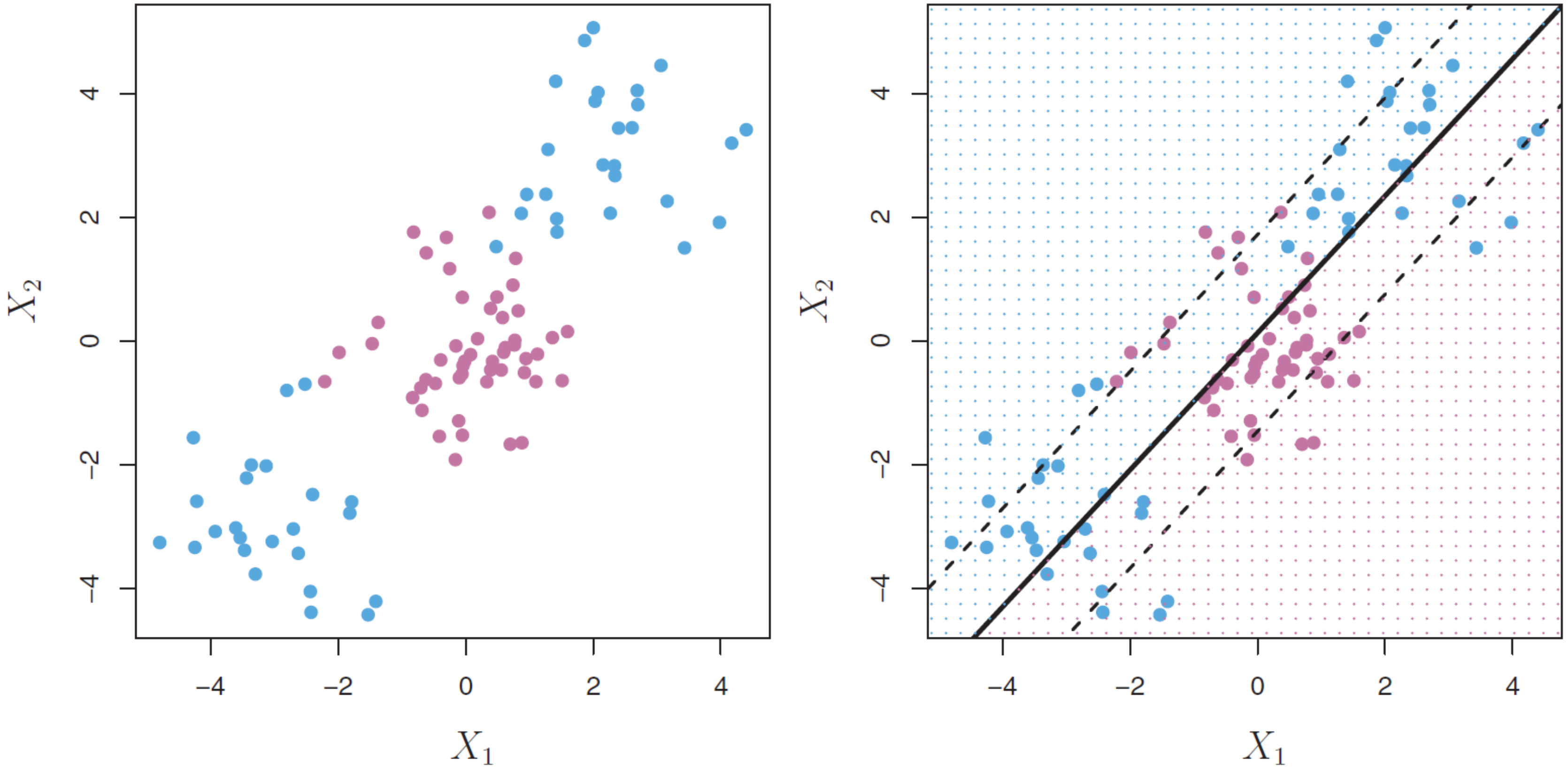
- C is the *budget* for the amount that the margin can be violated.
  - Maximum margin classifier = SVC with C = 0
- C controls the bias-variance trade-off.
  - When C is large:
    - Many observations violate the margin.
    - There are many support vectors.
    - SVC has low variance and potentially high bias.
  - When C is small:
    - Fewer observations violate the margin.
    - There are fewer support vectors.
    - SVC has high variance and potentially low bias.





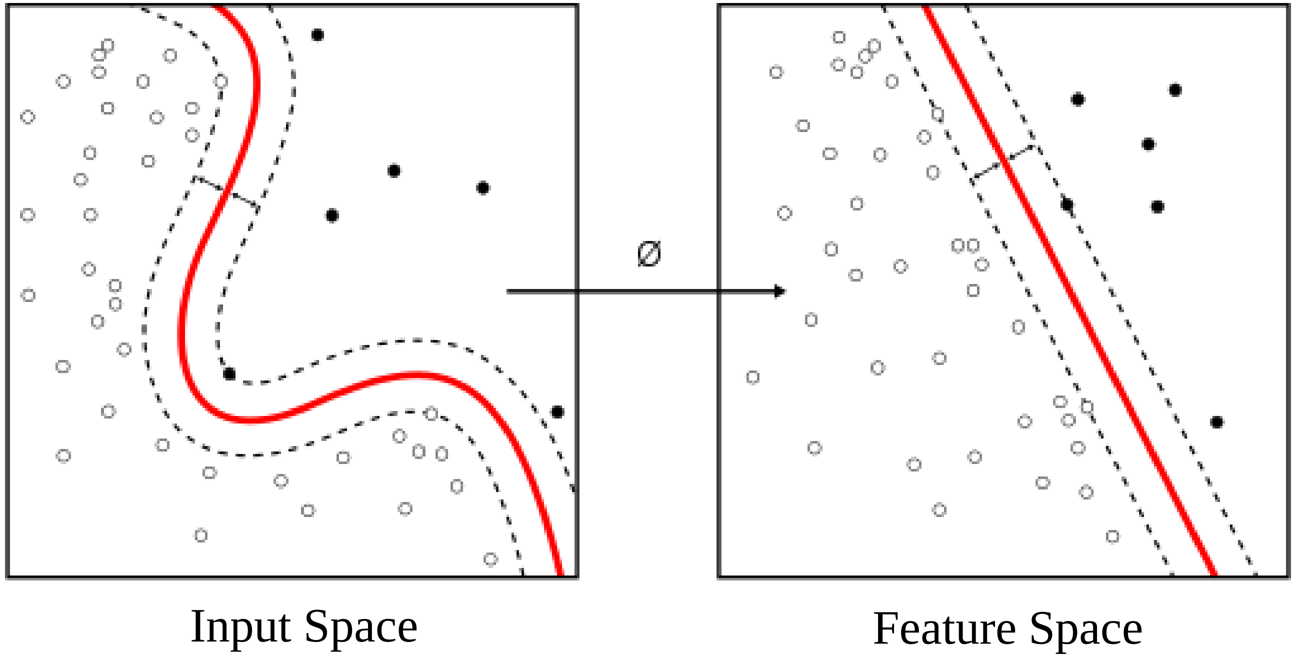
Problem of Support Vector Classifier

- When the true boundary is non-linear



Deal with Nonlinear Data: To Enlarge Feature Space

- Most classification tasks cannot be perfectly separated by a linear boundary.
- Support vector machine (SVM) uses a mathematical function, known as kernel, to map (transform) from input space to high dimensional feature space, such that the mapped objects are linearly separable in the transformed space.



Kernel Functions

- There are a couple of kernels
  - Linear
  - Polynomial
  - Radial basis function (RBF)
  - Sigmoid

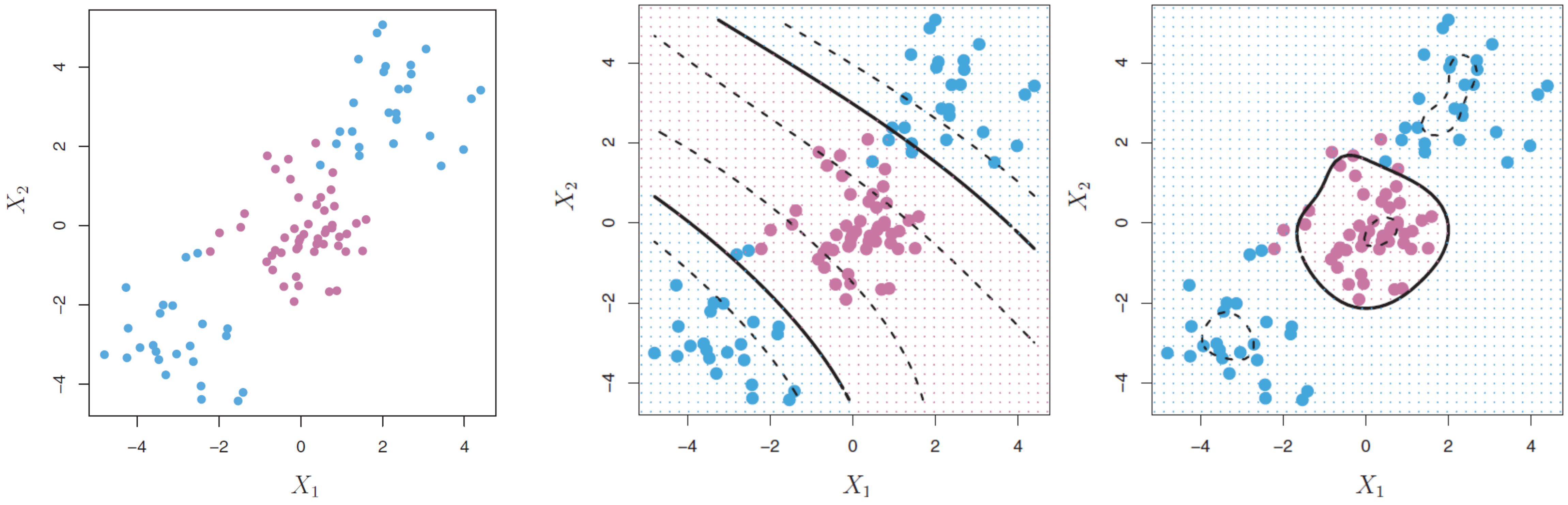
Kernel function is a dot product of input data points mapped to feature space by transformation Φ.

When SVC is combined with a non-linear kernel, it's called SVM.

$$K(X_i, X_j) = \Phi(X_i) \cdot \Phi(X_j) = \begin{cases} X_i \cdot X_j & Linear \\ (\gamma X_i \cdot X_j + C)^d & Polynomial \\ exp\left(-\gamma |X_i - X_j|^2\right) & RBF \\ tanh(\gamma X_i \cdot X_j + C) & Sigmoid \end{cases}$$

Two hyper-parameters except for linear SVM: (1) gamma γ; (2) C.

Example



Original data

Polynomial kernel of degree 3

Radial kernel

The Choice of Kernels Matters

- "Kernel Trick"
  - The training set is not linearly separable in the input data space;
  - The training set is linearly separable in the feature space.
- Choosing the right kernel based on the problem or application can improve the performance of SVM.
- In practice, we usually choose kernel by trial and error on the test set.
- RBF might be the most popular kernel.

Advantages of SVM

- Advantage
  - Effective in high dimensional spaces
  - Uses a subset of training points (i.e., support vectors), so it is memory efficient
  - Provides different kernels to handle different decision problems
- Disadvantage
  - Usually has poor performance when the number of features is much greater than the number of samples
  - Does not directly provide probability estimates

SVMs with More than Two Classes

- Separating hyperplane cannot naturally handle more than two classes.

- Two most popular approaches:
  - One-Versus-One Classification
    - For response with K classes, construct $\frac{K*(K-1)}{2}$ binary classifiers, each for one of the all $\binom{K}{2}$ pairs;
    - Count the number of times that the test observation is assigned to each of the K classes;
    - The final classification is to assign the most frequently assigned class.

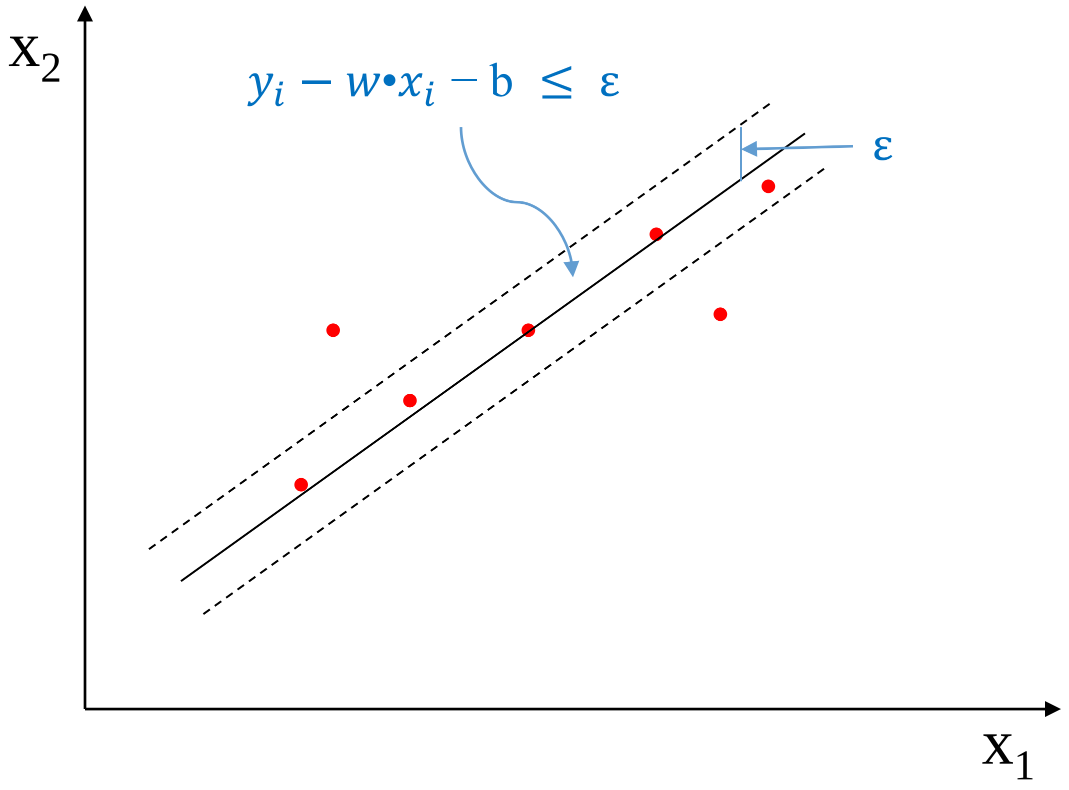
  - One-Versus-All Classification
    - Fit K SVMs, each time comparing one of the K classes to the remaining $K-1$ classes;
    - Let $\beta_{0k}, \beta_{1k}, \beta_{2k}, \dots, \beta_{kp}$ denote the parameters resulting from the K-th classifier, X^* denotes the new observation;
    - Choose the class with the largest $\beta_{0k} + \beta_{1k}X_1^* + \beta_{2k}X_2^* + \cdots + \beta_{kp}X_p^*$ (higher confidence of correct classification).

Support Vector Regression

- To find a linear function f(x) = w•x + b

Minimize $\frac{1}{2}\|w\|^2$

Subject to $y_i - w \bullet x_i - \mathrm{b} \leq \varepsilon$



Unlike SVC that tries to find a hyperplane to separate classes, support vector regression (SVR) tries to find a hyperplane that minimizes the coefficients (L2 norm).

Q & A