Support Vector Machines

Chapter 5: pp 153 – 164

Linear SVM Classification

- Consider linearly separable dataset
- Large Margin Classifier – SVM will fit the widest possible street

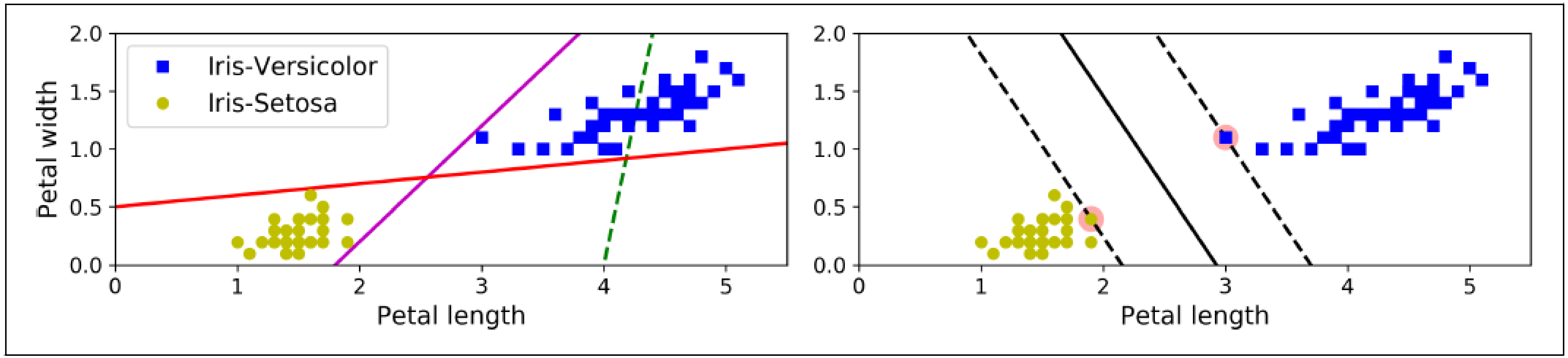

Figure 5-1. Large margin classification

- Decision boundary defined by the support vectors

- Adding more instances “off the street” does not change the decision boundary
- SVMs are sensitive to feature scales

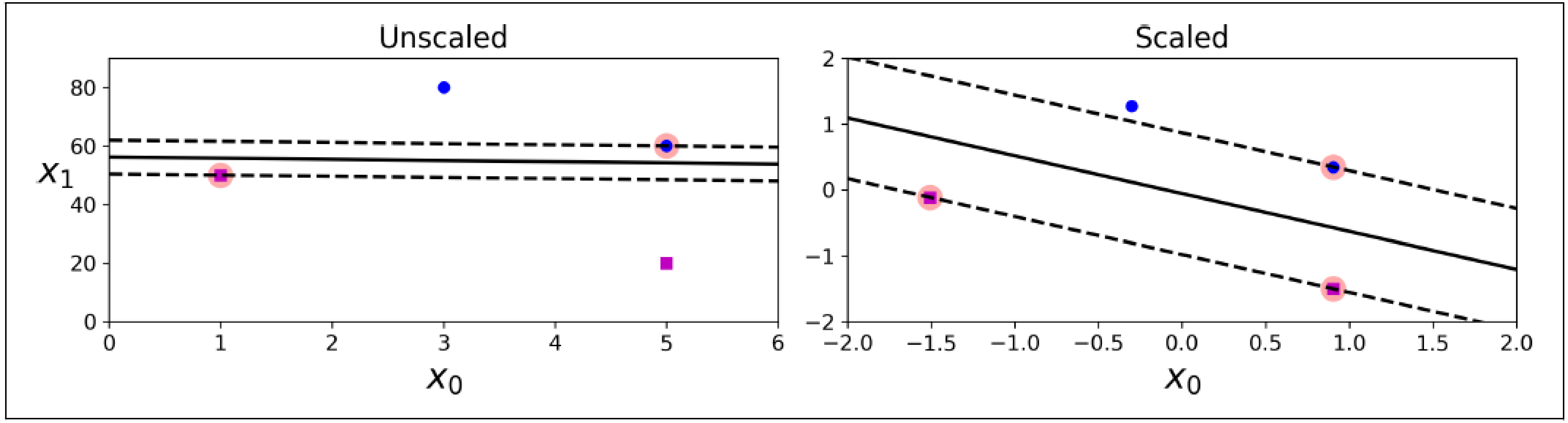

Figure 5-2. Sensitivity to feature scales

Soft Margin Classification

- Problems with hard margin classification
  - Works only if the data is linearly separable
  - Very sensitive to outliers

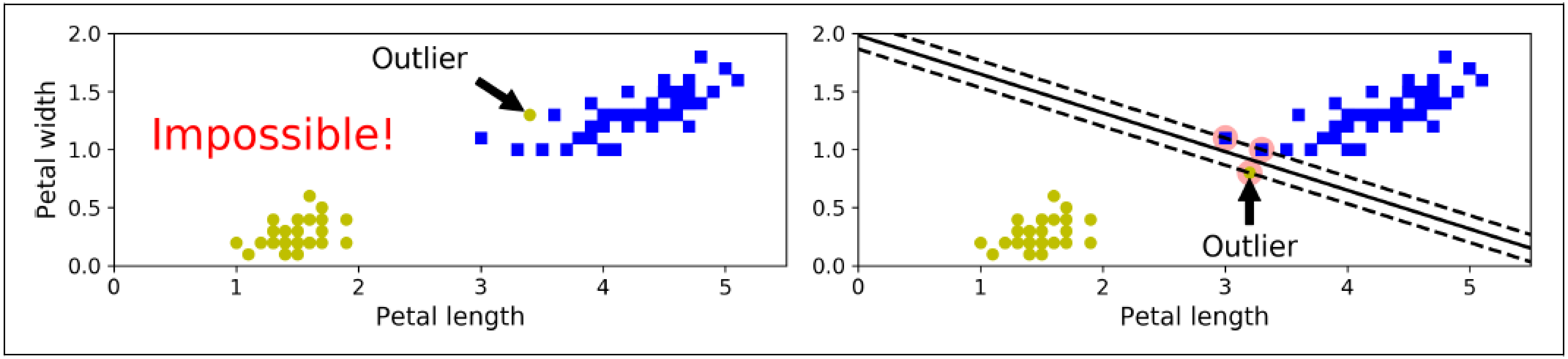

Figure 5-3. Hard margin sensitivity to outliers

- Soft Margin Classification finds a balance between maximizing the width of the street and minimizing margin violations controlled by a hyperparameter C
  - Large C: don't want misclassified points but will accept narrow margin
  - Small C: accept some misclassified points but want big margin

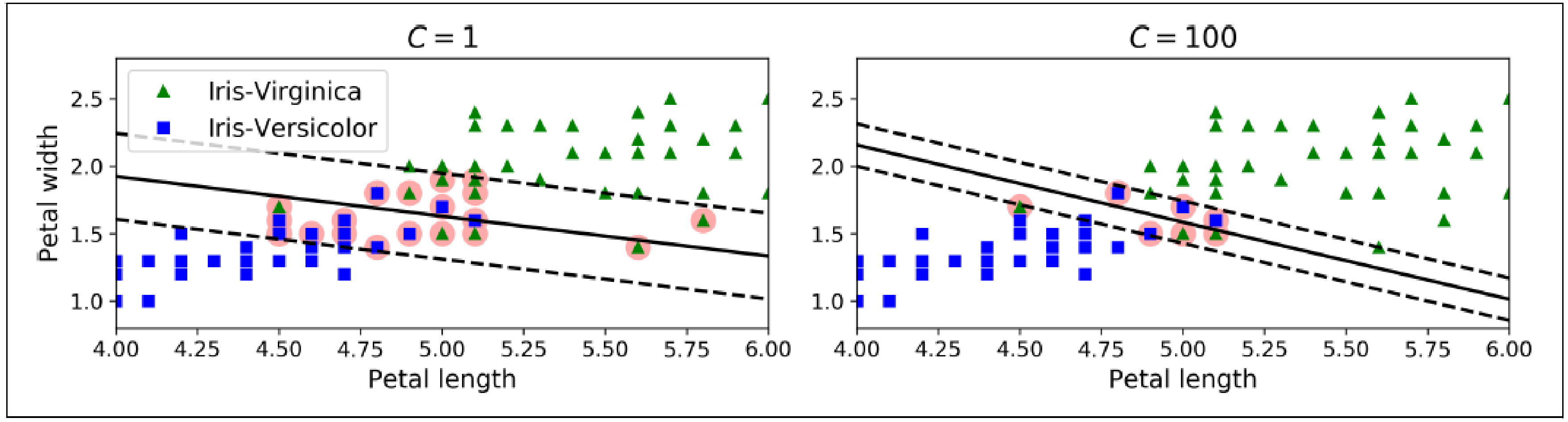

Figure 5-4. Large margin (left) versus fewer margin violations (right)

- Regularization can be done by reducing C
- SVM does not output probabilities for each class

Non-linear SVM Classification

- Adding a feature makes a non-linearly separable dataset separable

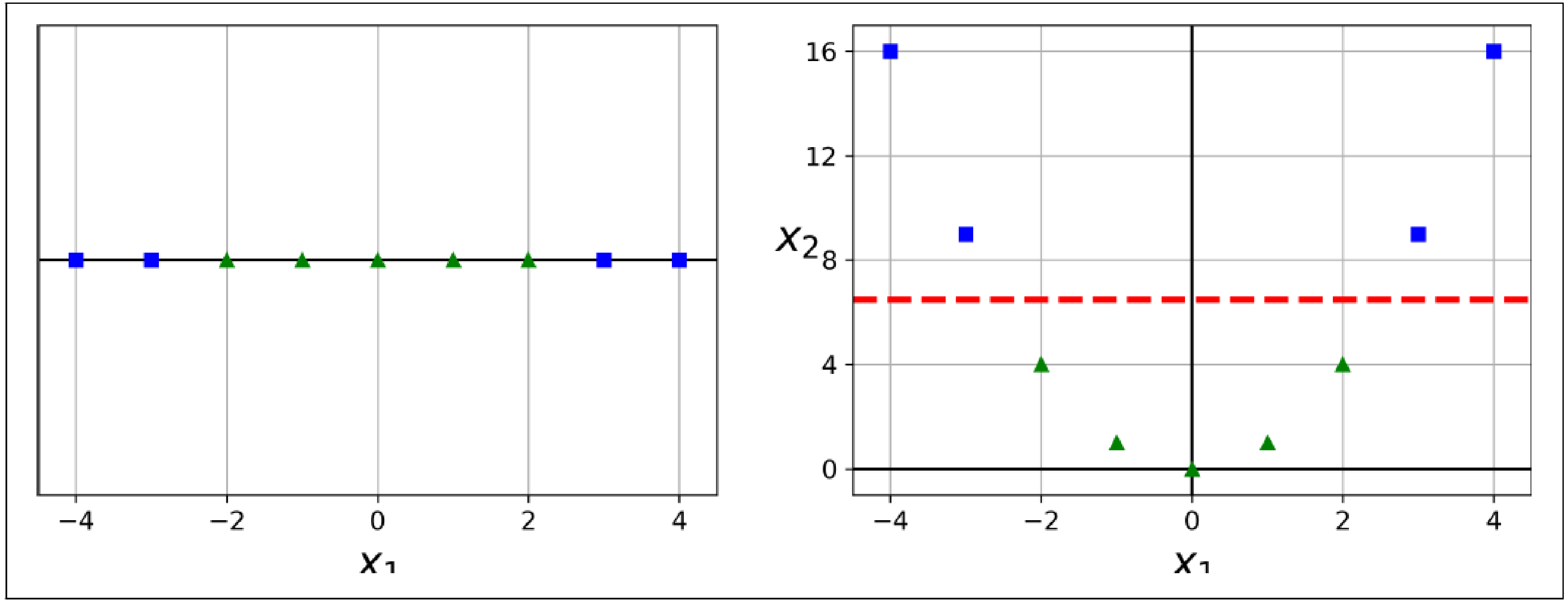

Figure 5-5. Adding features to make a dataset linearly separable

The Moons Dataset

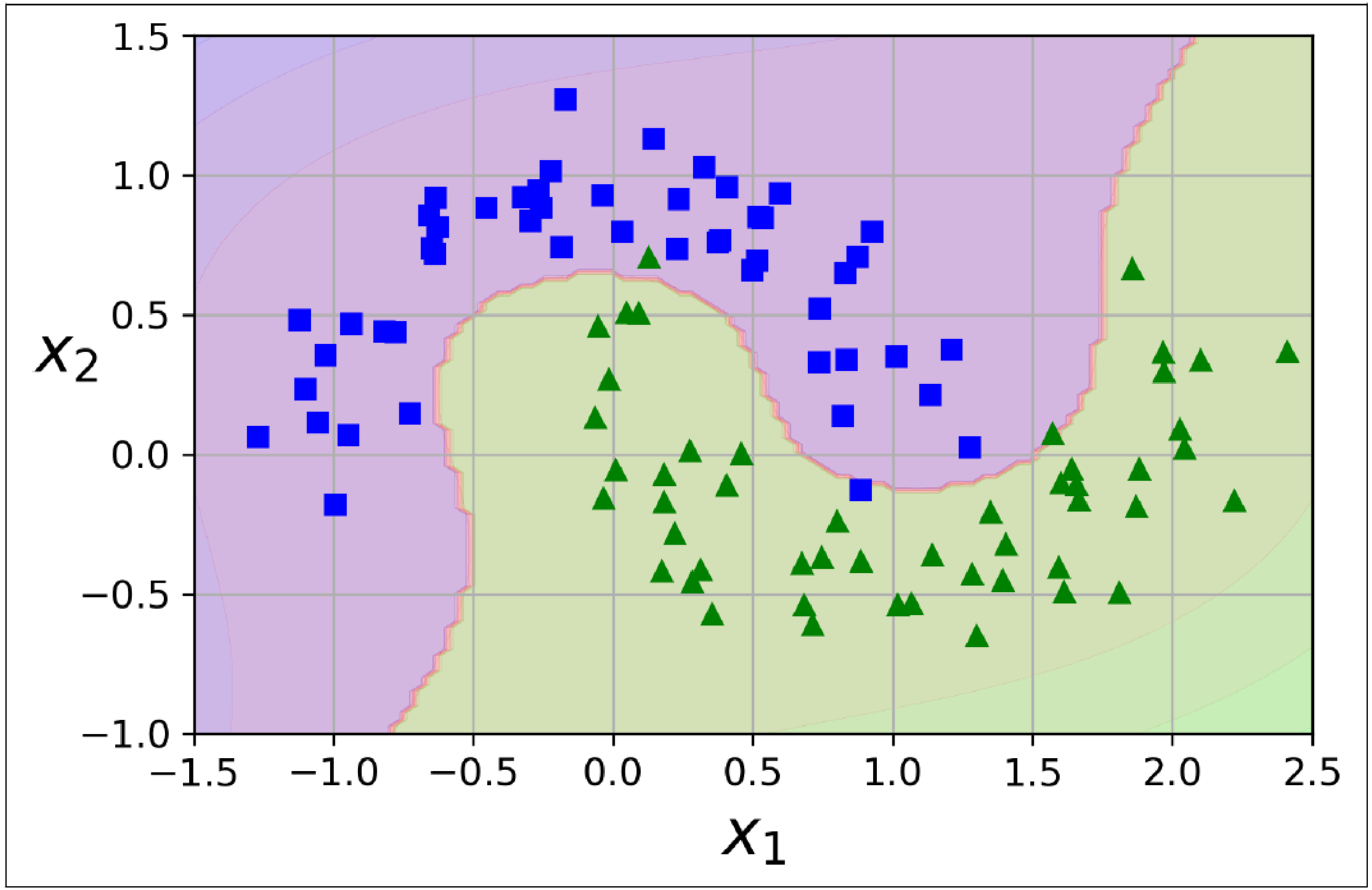

Figure 5-6. Linear SVM classifier using polynomial features

Polynomial Kernel

- Low polynomial degree cannot deal with complex datasets
- High polynomial degree makes model very slow due to a huge number of features
- Solution: use *kernel trick*
  - Maps instances into very high-dimensional space (feature space)
  - Possible to get the same result as adding high degree polynomials without actually adding them

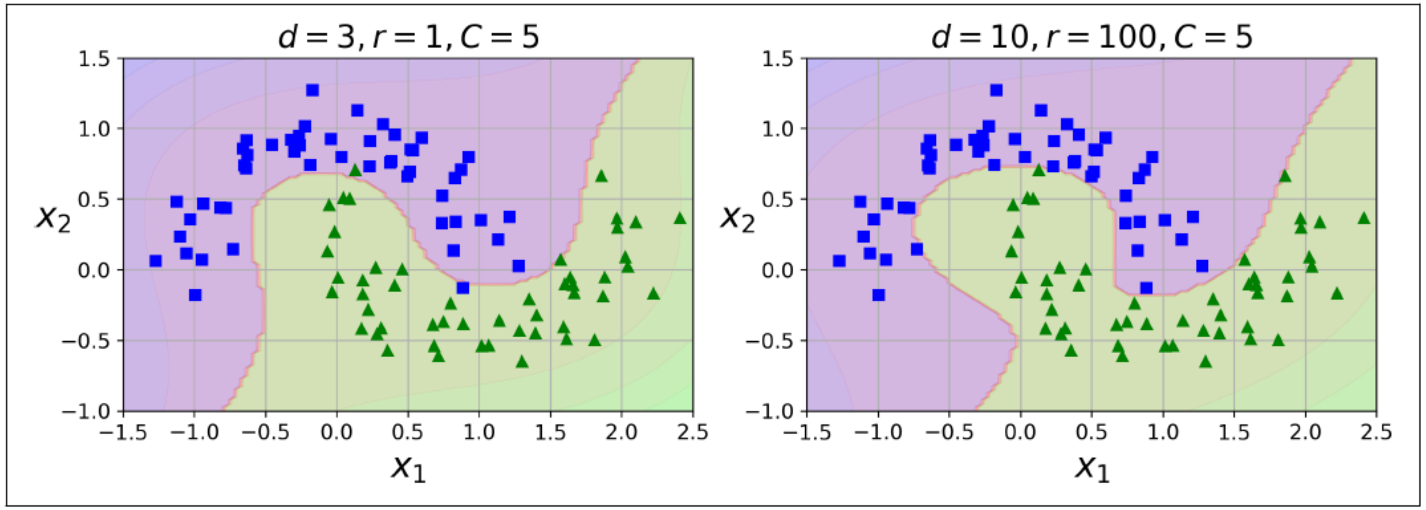

Figure 5-7. SVM classifiers with a polynomial kernel

Similarity Features

- Similarity function to measure the distance of an instance to a landmark
- Landmark: Gaussian Radial Basis Function (RBF) with $\gamma = 0.3$ for $x_1 = -2, x_1 = 1$

Equation 5-1. Gaussian RBF

$$\phi_{\gamma}(\mathbf{x}, \ell) = \exp\left(-\gamma \| \mathbf{x} - \ell \|^2\right)$$

$$x_2 = \exp\left(-0.3 \times 1^2\right) \approx 0.74$$

$$x_3 = \exp\left(-0.3 \times 2^2\right) \approx 0.30$$

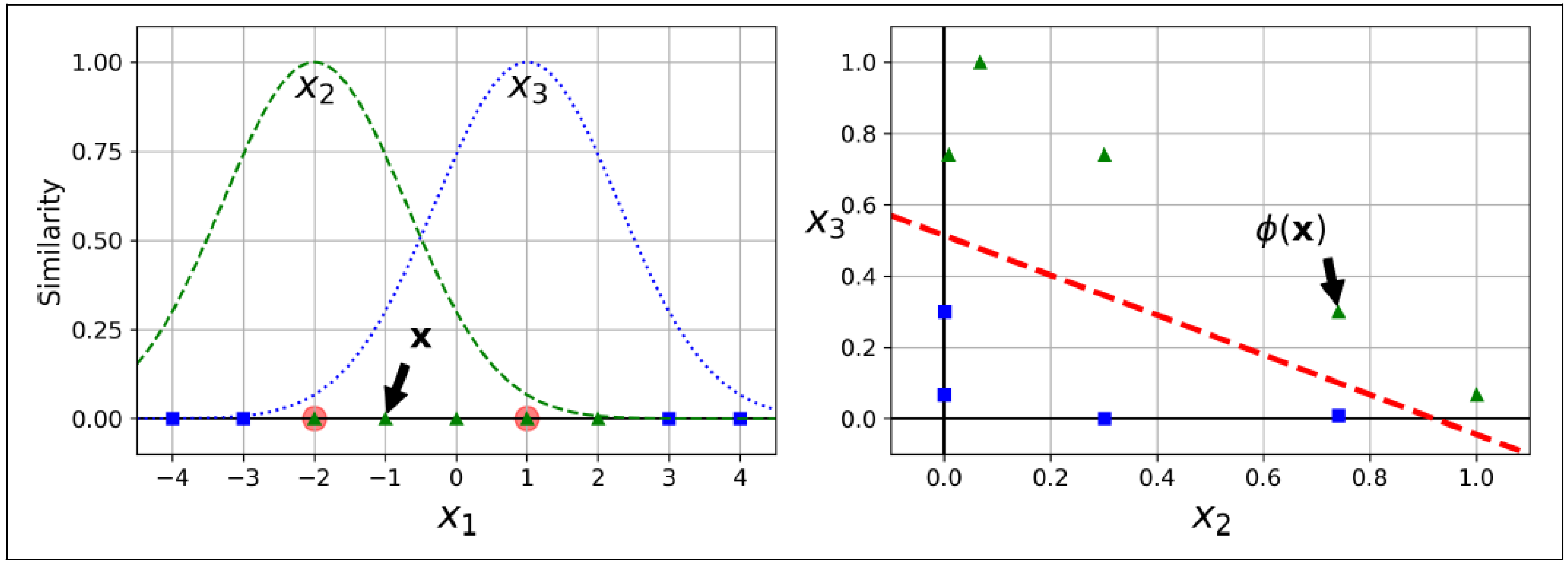

Figure 5-8. Similarity features using the Gaussian RBF

Gaussian RBF Kernel

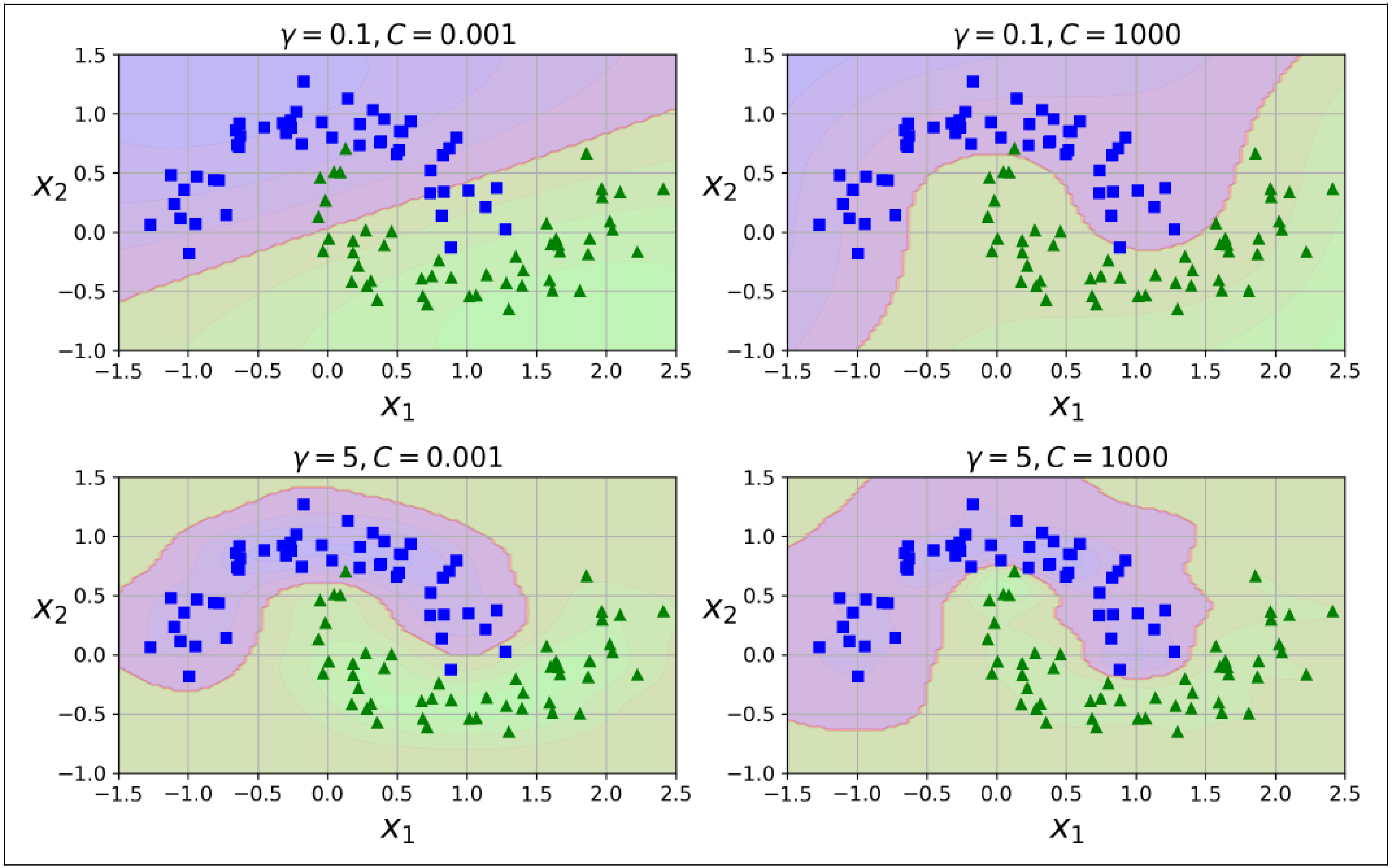

Figure 5-9. SVM classifiers using an RBF kernel

Computational Complexity

- Training time is between $O(m^2 \times n)$ and $O(m^3 \times n)$
  - m: number of instances
  - n: number of features
- Algorithm is good for complex but small/medium training sets
- Very slow for large datasets
- Scales well with the number of features

SVM Regression

- Supports linear and non-linear regression
- Objective: fit as many instances as possible on the street
- Street width controlled by hyperpameter ε

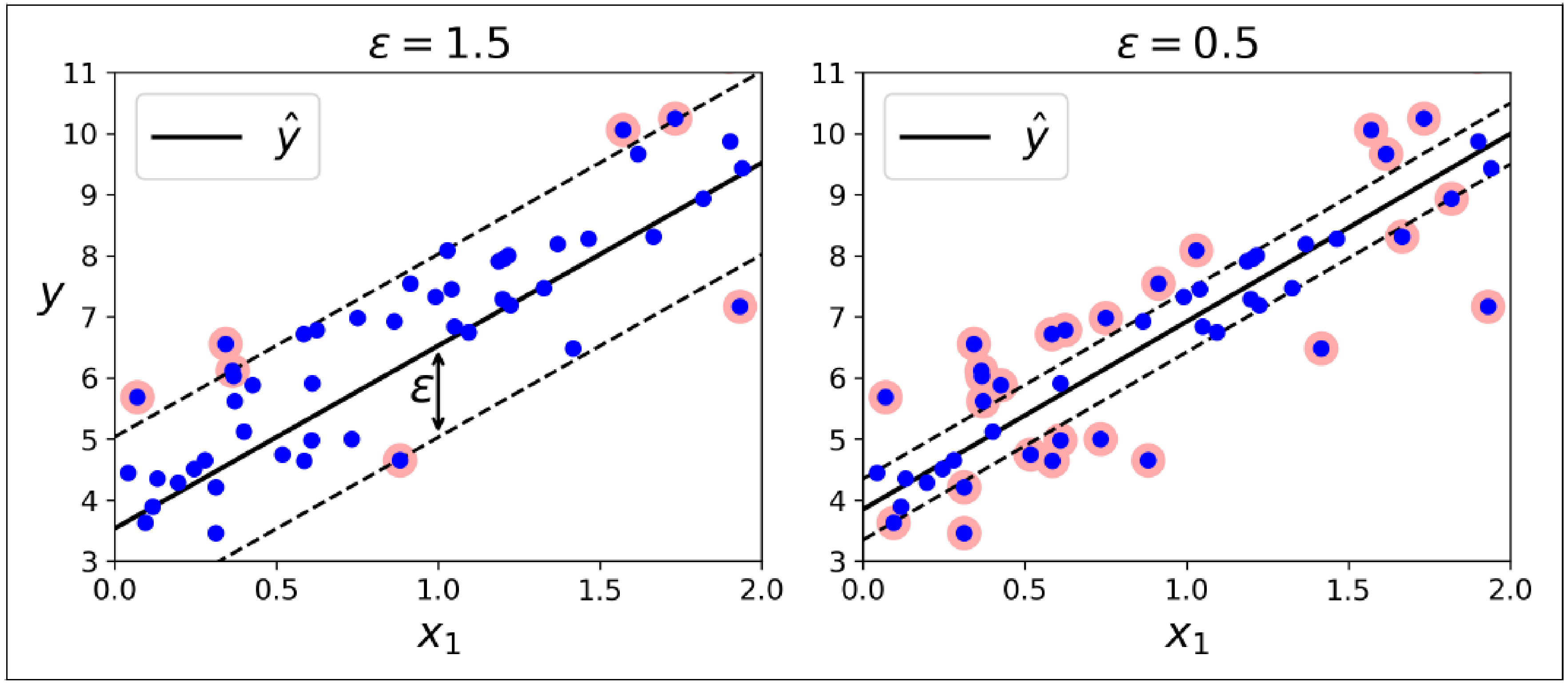

Figure 5-10. SVM Regression

- For non-linear regression, use kernelized SVM

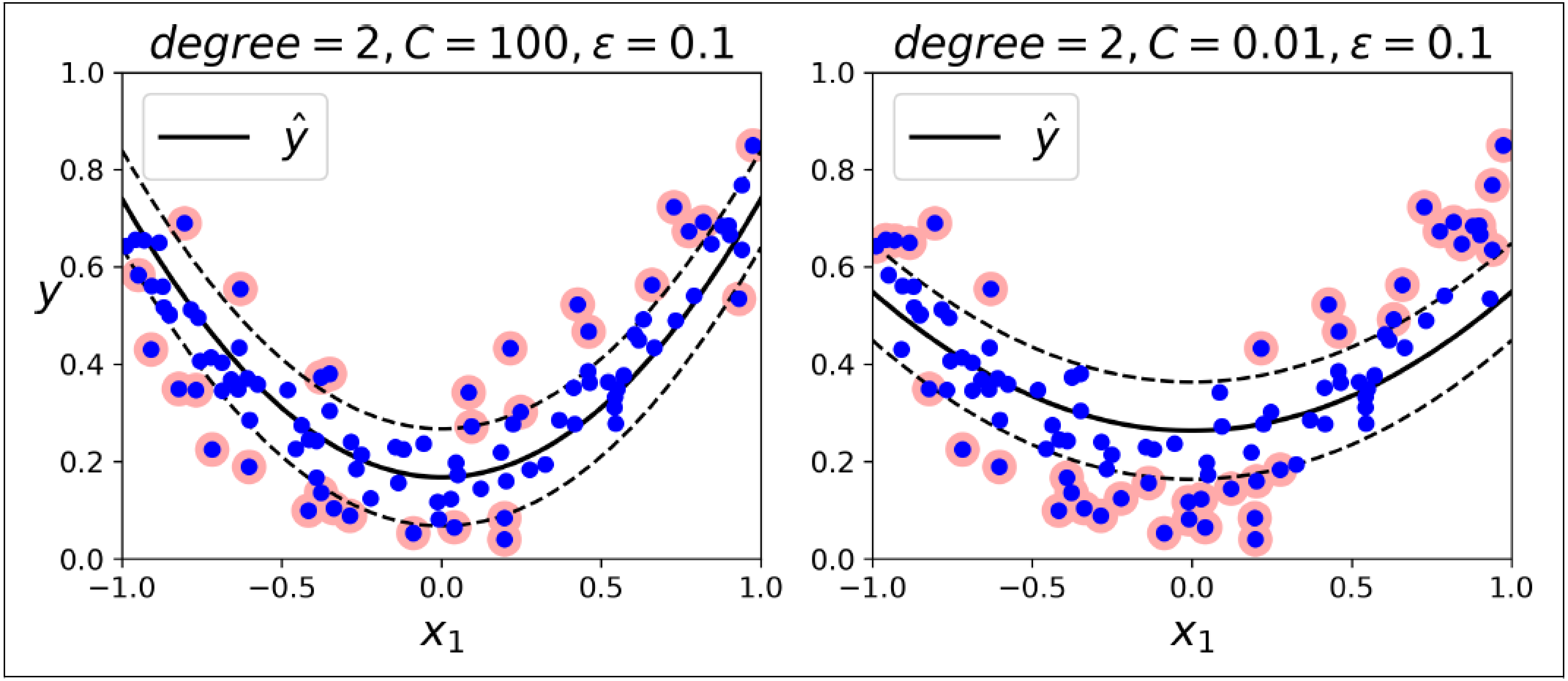

Figure 5-11. SVM regression using a 2nd-degree polynomial kernel