Chapter 5: Support Vector Machines

Dr. Xudong Liu
Assistant Professor
School of Computing
University of North Florida

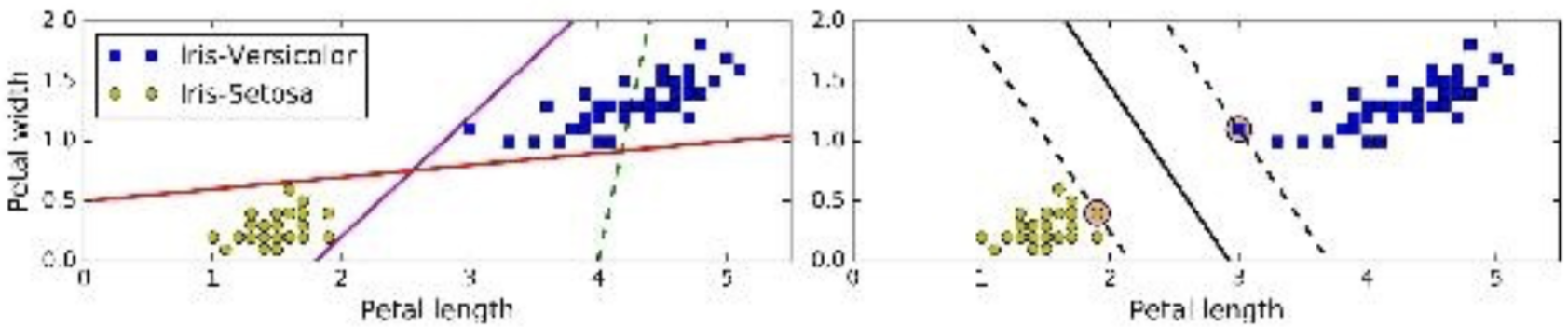
Monday, 9/23/2019

Support Vector Machines

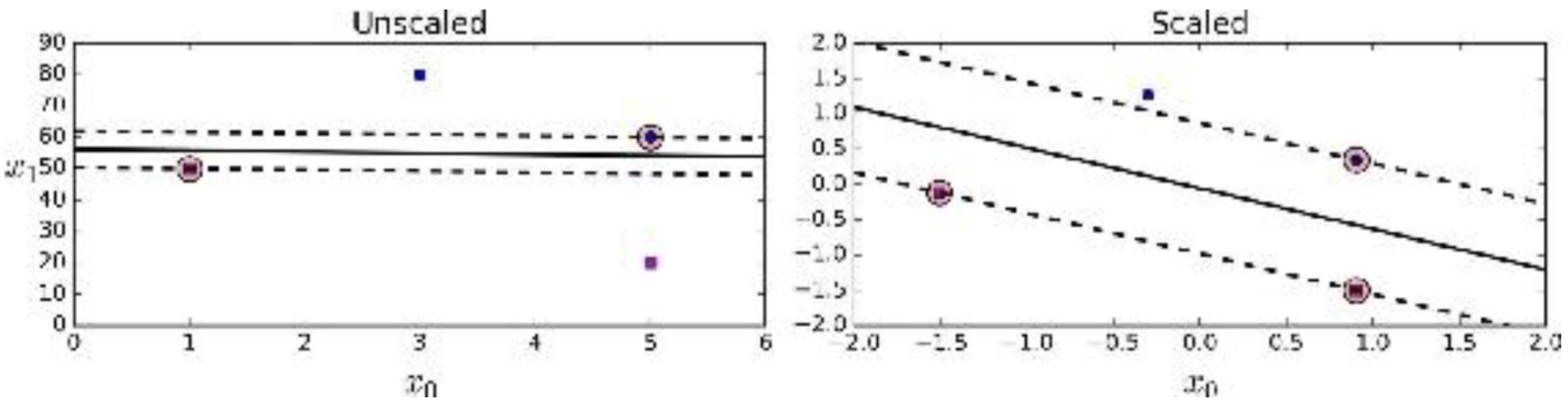
- Linear SVM classification: hard margin, soft margin
- Nonlinear SVM classification: polynomial features, similarity features
- Under the hood: decision function, objective functions

Linear Support Vector Machines

- A *linear SVM* classifier fits the "widest possible street" between the classes.
  - The solid line in the right image represents the decision boundary of this SVM.
  - It not only separates the two classes, but also stays as far away from the closest training example as possible.
- The decision boundary is determined or supported by the examples located on the edge of the street. These examples are called "support vectors."

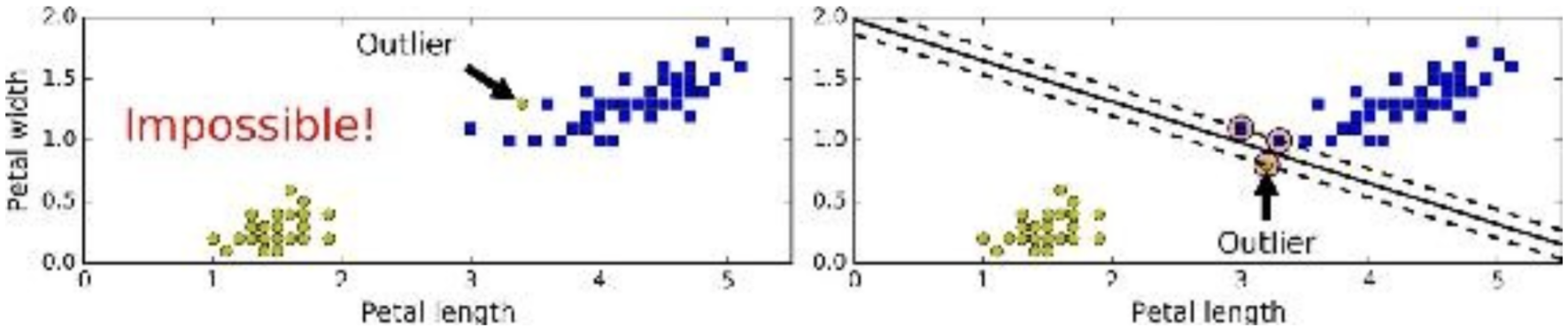


Sensitivity to Feature Scales



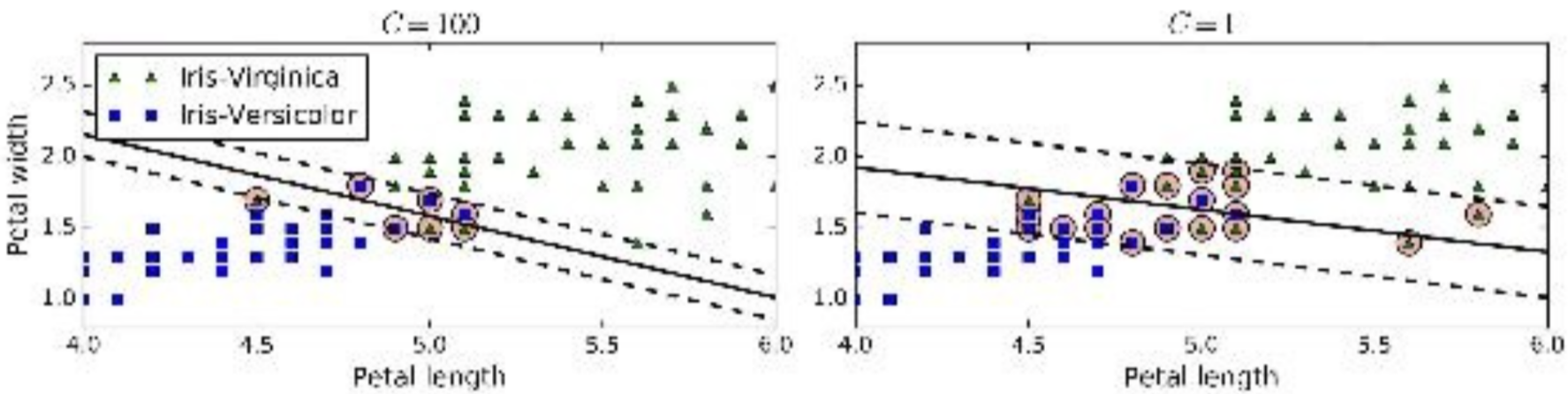
Hard Margin Classification

- Hard margin classification: if we impose that all examples must be off the street and on the correct side.
- Problems: only works for linear separable, and is sensitive to outliers.



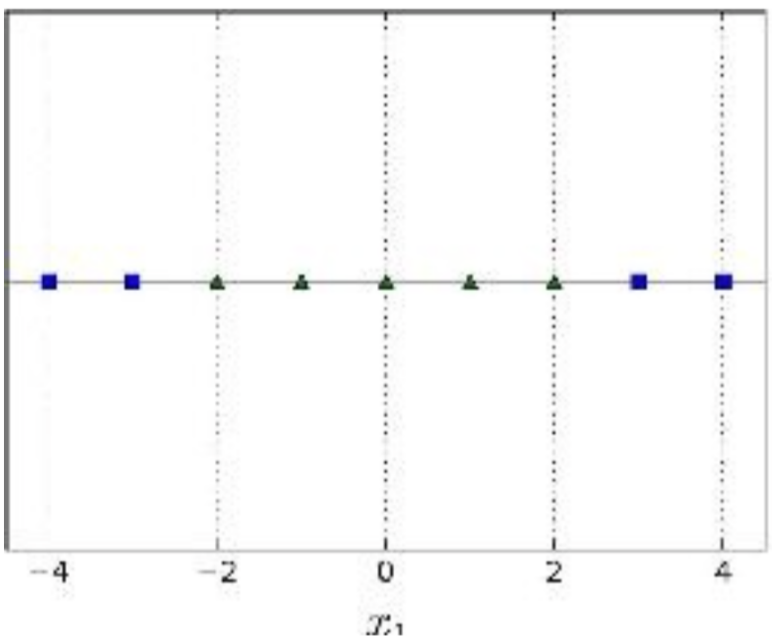
Soft Margin Classification

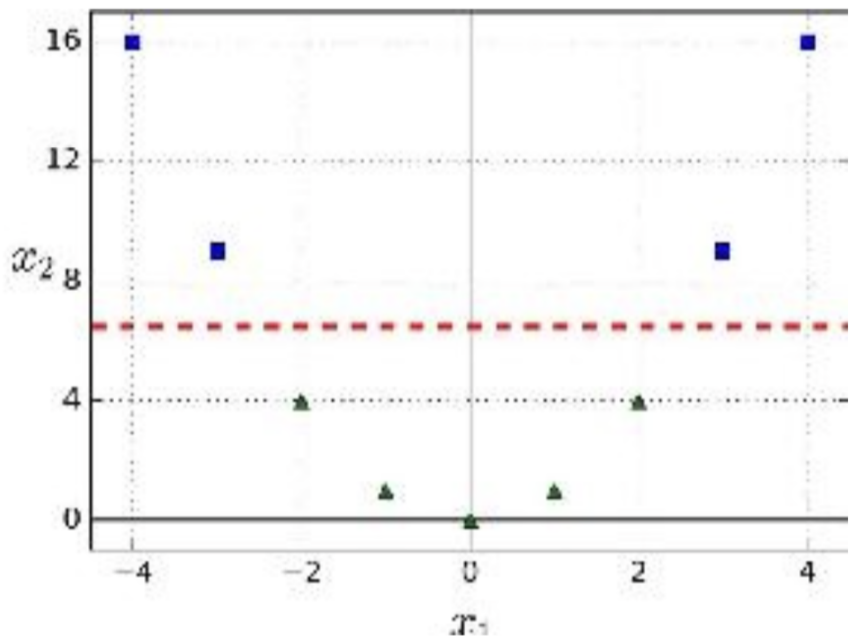
- To be more flexible, soft margin classification tries to balance between *margin size* (keep the street as wide as possible) and *margin violations* (keep the number of examples in the street or even on the wrong side as small as possible).
- Scikit-Learn: *LinearSVC* provides a hyperparameter called C. The smaller it is, the wider the street with more violations.
- Reducing C can help with overfitting.



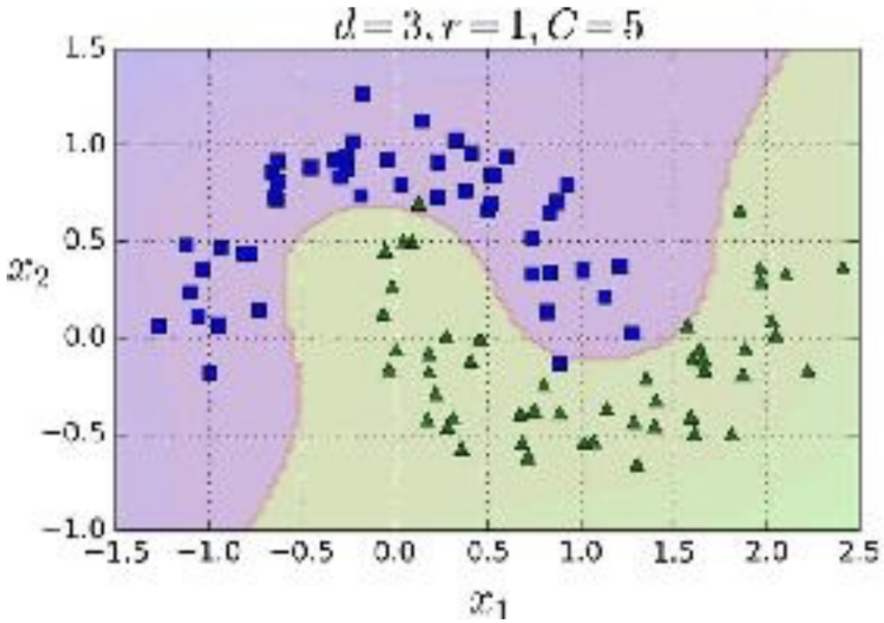
Nonlinear Support Vector Machines

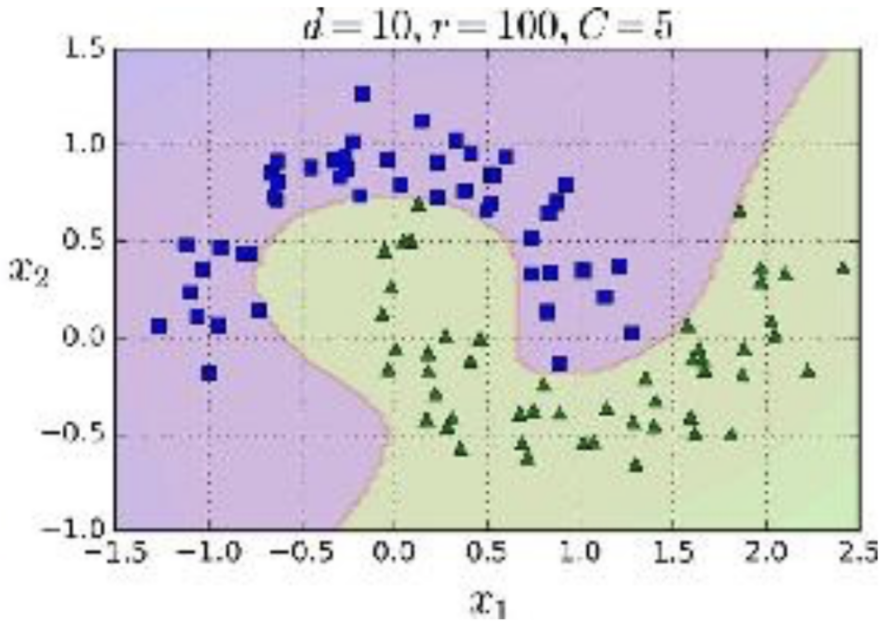
- Many datasets are not even close to being linearly separable.
- By *adding features* we can add extra features to the dataset to make it linear separable.
  - One feature x_1 alone is not separable.
  - Adding feature $x_2 = x_1^2$ makes the dataset separable.





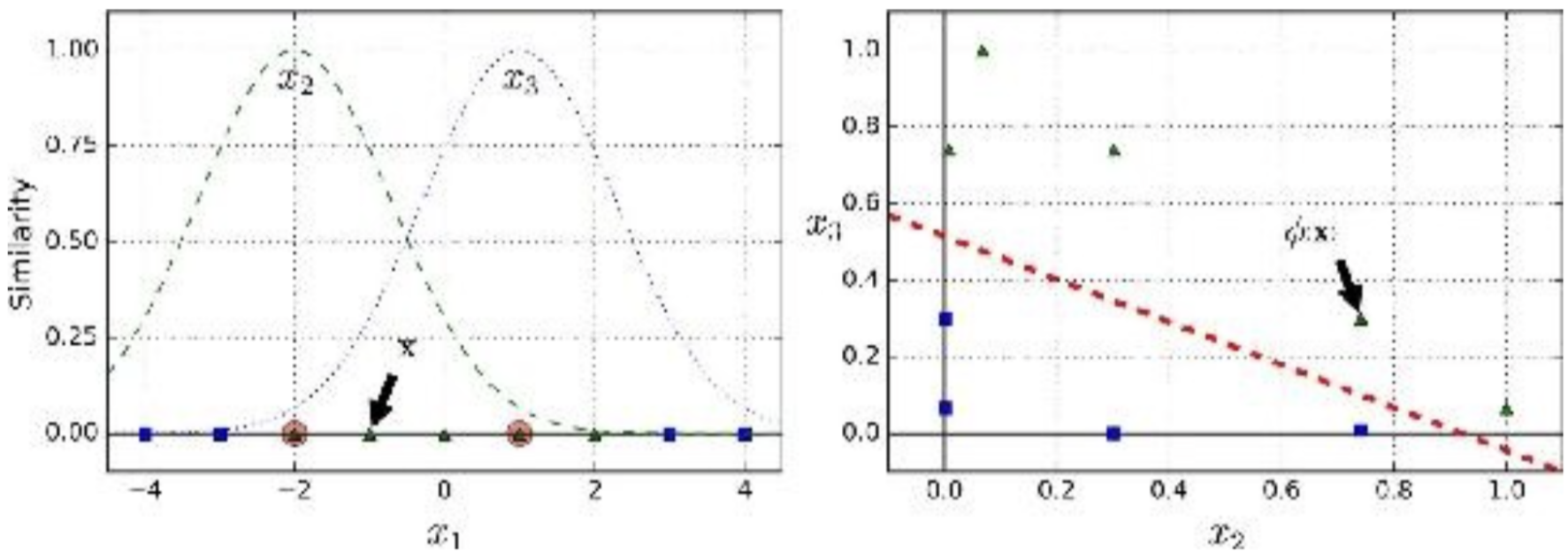
Polynomial Features





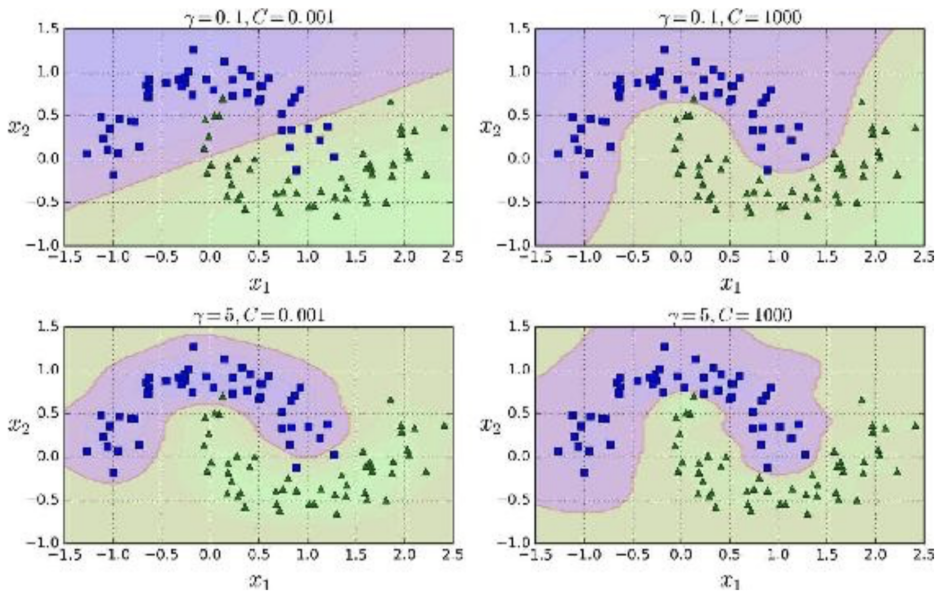
Similarity Features

- Similarity features measure how much each training example resembles a particular landmark.
- Gaussian Radial Basis Function: $\phi_\gamma(\mathbf{x}, \mathbf{l}) = \exp(-\gamma||\mathbf{x} - \mathbf{l}||^2)$.



Gaussian Radial Basis Function

- Hyperparameter γ: the bigger, the decision boundary becomes more irregular, more wiggling around the examples.
- If overfitting, try reduce γ, and C as well.



Computational Complexity

- When the dataset is very large and has many features, try Linear SVM.
- It dataset is not very large, try SVC's Guassian RBF which tends to work better than Linear SVM.

Class	Time complexity
`LinearSVC`	$O(m \times n)$
`SGDClassifier`	$O(m \times n)$
`SVC`	$O(m^2 \times n)$ to $O(m^3 \times n)$

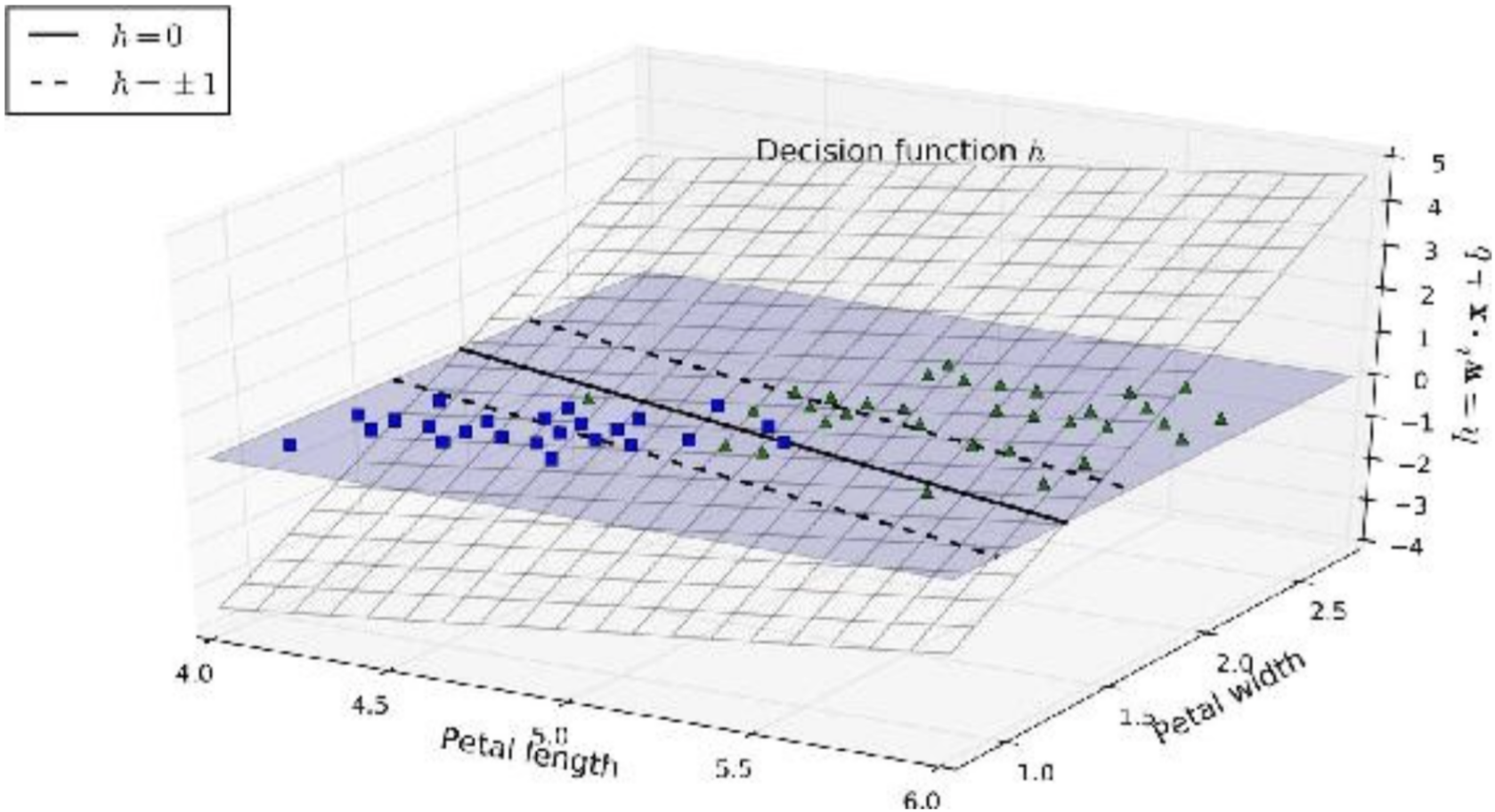
Linear Support Vector Machines

- Linear SVM classifier predicts the class of a new instance $\mathbf{x}$ by computing the *decision function*:

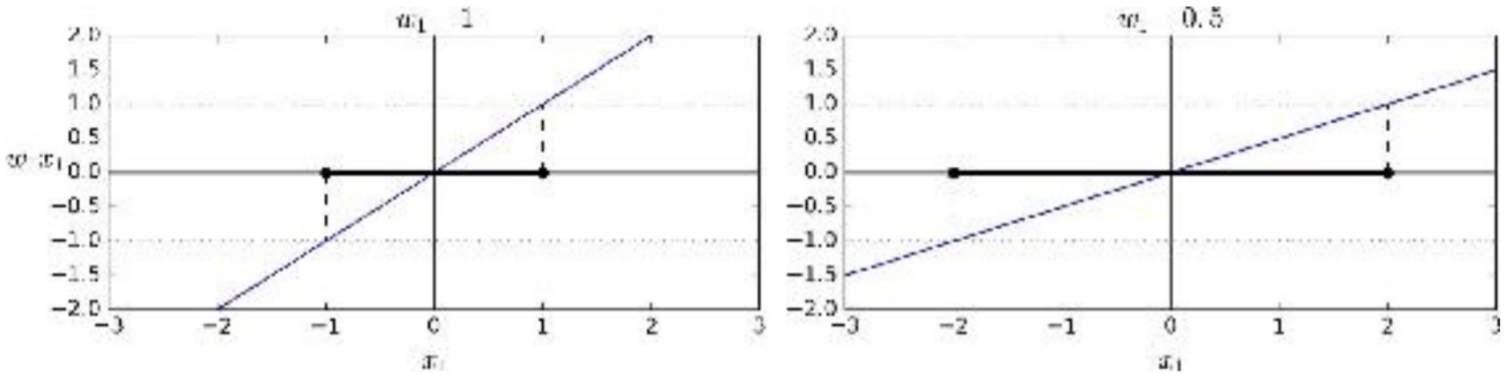
$$\mathbf{w}^T \cdot \mathbf{x} + b = w_1 x_1 + \ldots + w_n x_n + b.$$

- Then, if the result is negative, the prediction is the negative class; otherwise, the positive class.

Linear Support Vector Machines



Objective Function in Hard Margin Linear SVM Learning



- Because the smaller the $||\mathbf{w}||$, the wider the street, we want to minimize it.
- At the same time, we want to avoid margin violations, then we want the decision function to be greater than 1 for positive examples and smaller than -1 for negative examples.
- Thus, the following objective function:

$$\underset{\mathbf{w},\mathbf{b}}{minimize}\ \frac{1}{2}\mathbf{w}^T \cdot \mathbf{w}$$

$$subject\ to\ t^{(i)}(\mathbf{w}^T \cdot \mathbf{x}^{(i)} + b) \geq 1 \text{ for } i = 1, \ldots, m$$