KU x depa
MACHINE LEARNING
Training Course 2022

Support Vector Machines

DR. SETHAVIDH GERTPHOL

KASETSART UNIVERSITY

depa

Today's Outline

- Introduction
- Concepts
  - Mamimum Margin
  - Support Vector
  - Hard and Soft Margin
- Non-linear classification
  - Polynomial kernel
  - RBF kernel
- Scikit-learn implementation

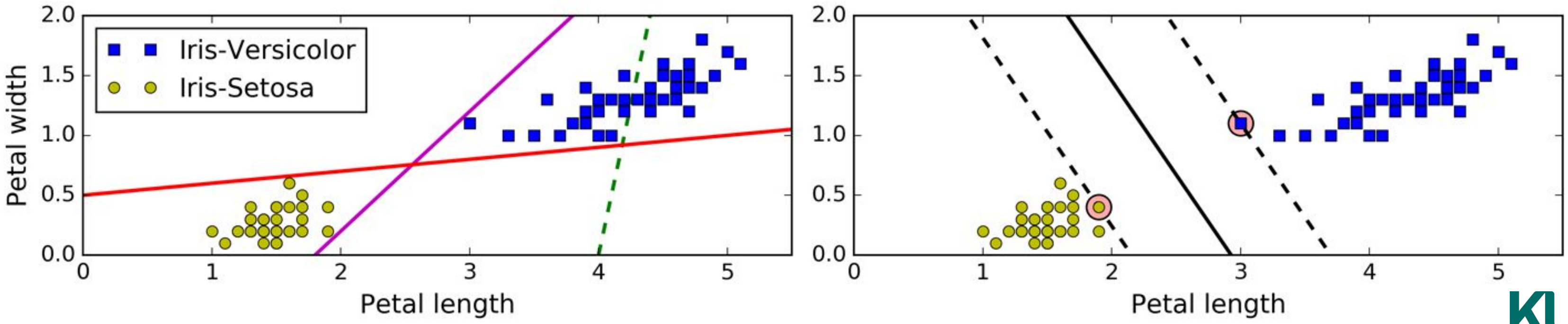
KU KASETSART UNIVERSITY depa

Support Vector Machines

- Very versatile technique, can do
  - Classification
  - Regression
  - Outliers detection
- Can find both linear and non-linear decision boundaries
- Support many kernel types, such as linear, polynomial, radial base function, etc.
- Give good performance, but must tune hyperparameters

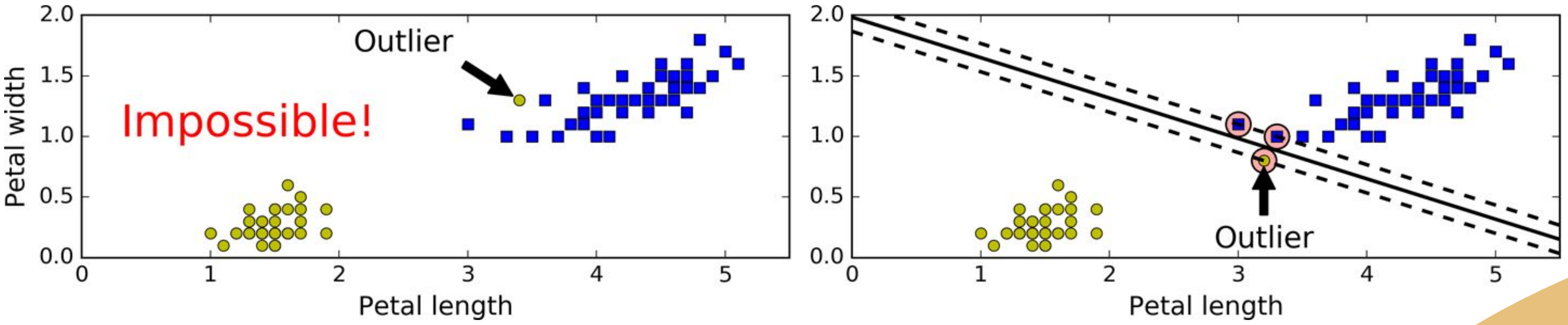
Concepts

- SVM finds a decision boundary with maximum margin (distance) to nearest data points of the classes
- The margin provides better separation between classes and gives good predictions for future data points
- Decision boundary depends on a few closest data points only; these are called support vector



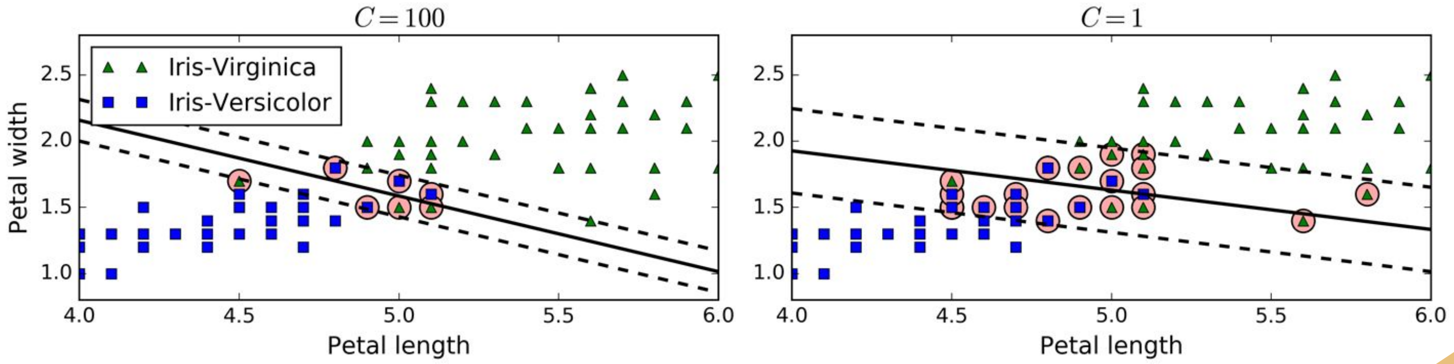
Hard Margin

- Does not allow any data point inside the margin
- Does not work when data points overlap (not linearly separatable)
- Sensitive to outliers



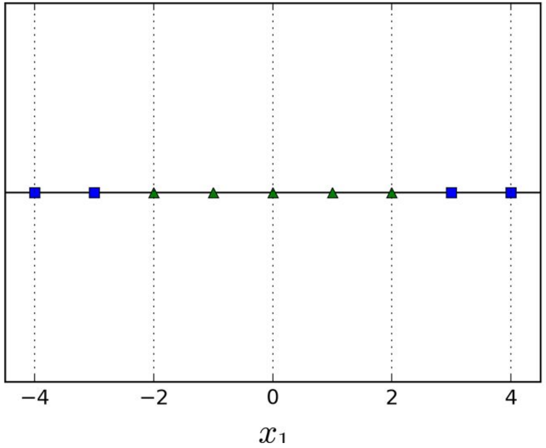
Soft Margin

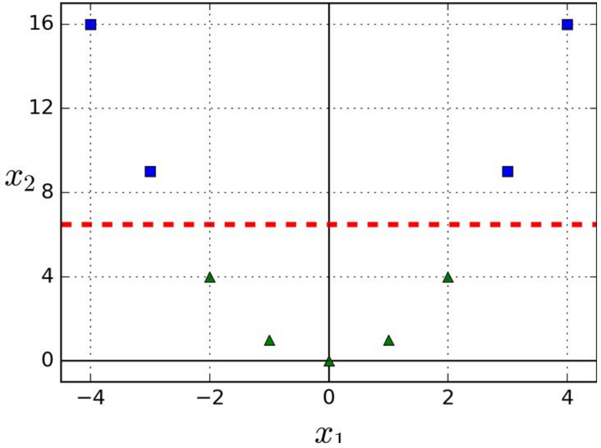
- Allows margin violations: data points inside the margin or even on the wrong side of the boundary
- How many violations do we allow? We can control it with a hyperparameter called C (regularization parameter)
  - Small C: less strict, wider margin, more violations
  - Large C: stricter, smaller margin, less violations
- Find optimal C by hyperparameter tuning techniques (GridSearch, etc)



Non-linear classification

- Sometimes there is no good linear decision boundary between classes
- One technique is to add polynomial features to the dataset to make the classes linearly separatable in higher dimensions
- We can add these features manually but…how many dimensions to add?
  - Low dimensions may not be enough to find good decision boundary
  - High dimensions add a lot of features and a lot of calculations (slow)





Kernels

- A function that can calculate the dot product of transformed vectors based on the original vector only.
  - No need to actually transforming the vectors
  - Save calculation time

Transforming function

$$\phi(\mathbf{x}) = \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1{}^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2{}^2 \end{pmatrix}$$

$$\begin{aligned}\phi(\mathbf{a})^T\phi(\mathbf{b}) &= \begin{pmatrix} a_1{}^2 \\ \sqrt{2}\, a_1 a_2 \\ a_2{}^2 \end{pmatrix}^T \begin{pmatrix} b_1{}^2 \\ \sqrt{2}\, b_1 b_2 \\ b_2{}^2 \end{pmatrix} = a_1{}^2 b_1{}^2 + 2a_1 b_1 a_2 b_2 + a_2{}^2 b_2{}^2 \\ &= (a_1 b_1 + a_2 b_2)^2 = \left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^T \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right)^2 = (\mathbf{a}^T \mathbf{b})^2\end{aligned}$$

Polynomial kernels

- Calculate relationship between two vectors in higher dimensions
- 3 more parameters
  - Gamma: weight of the dot product term
  - r: constant term
  - d: degree of polynomial to use

$$\text{Linear:} \quad K(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}$$

$$\text{Polynomial:} \quad K(\mathbf{a}, \mathbf{b}) = (\gamma\mathbf{a}^T\mathbf{b} + r)^d$$

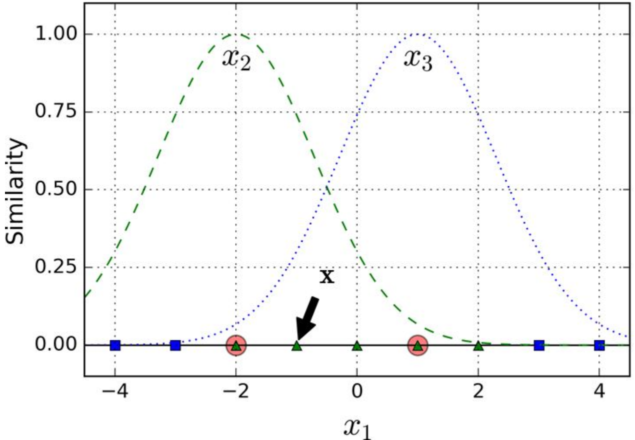
$$\text{Gaussian RBF:} \quad K(\mathbf{a}, \mathbf{b}) = \exp(-\gamma \| \mathbf{a} - \mathbf{b} \|^2)$$

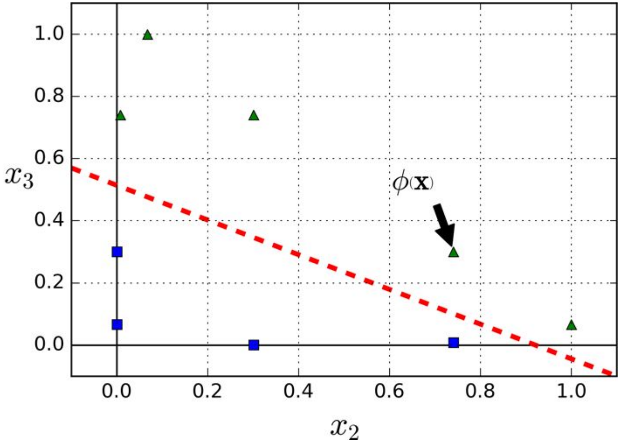
$$\text{Sigmoid:} \quad K(\mathbf{a}, \mathbf{b}) = \tanh(\gamma\mathbf{a}^T\mathbf{b} + r)$$

Gausian RBF (Radial Basis Function)

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

- Find similarity between a data point and a landmark
- Landmark can be every other data points
- Gamma controls the width of bell curve
  - High gamma makes the bell curve narrower, less effect from far away data points

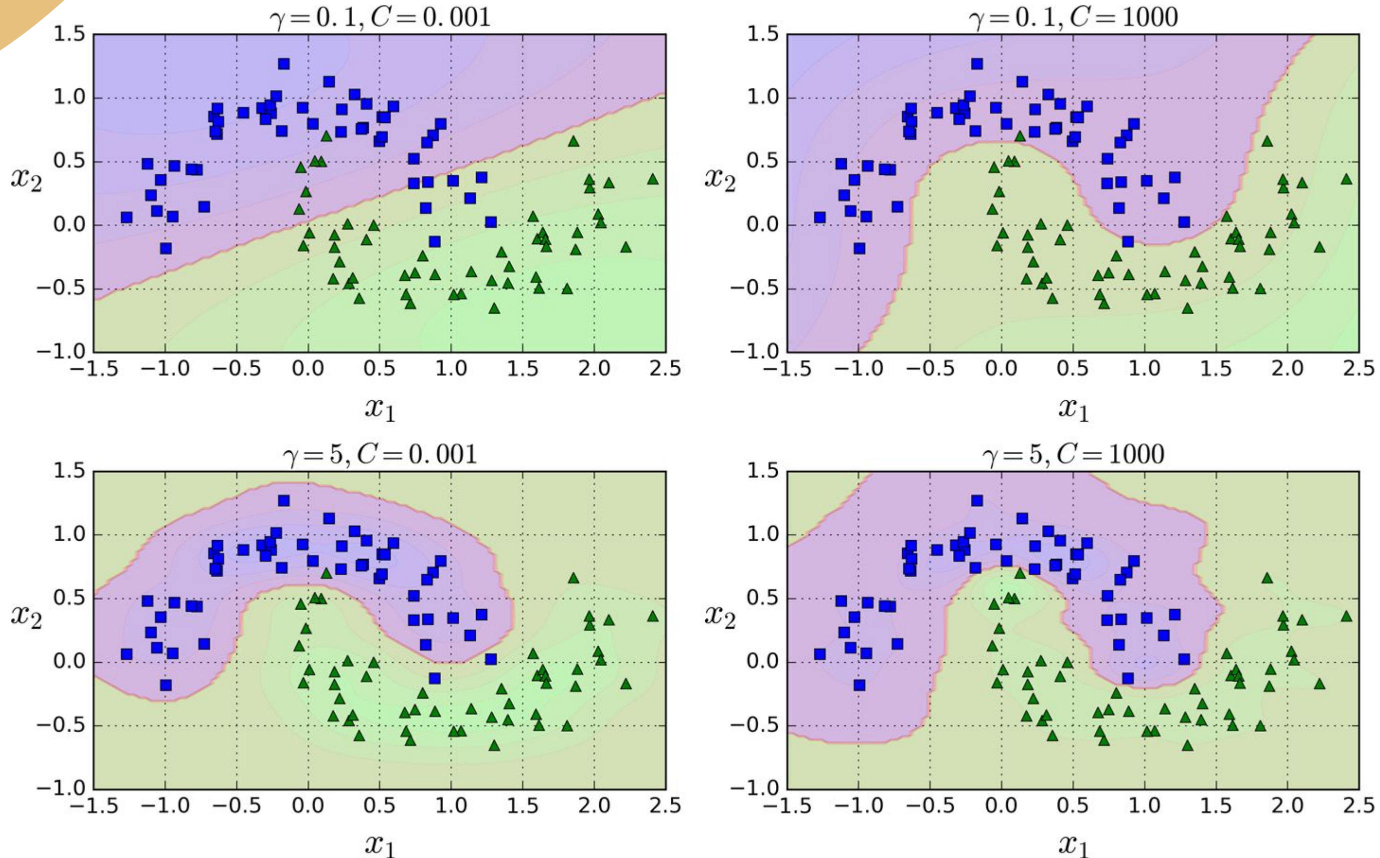




RBF kernel

- Each landmark will become one dimension, so total features can equal total data points in the training set!
- Can use kernel tricks to calculate up to infinite dimensions
- 2 parameters to tune: Gamma and C
  - C: margin
  - Gamma: range of influence of landmarks
  - High gamma: small range of influence, boundary irregular
  - Low gamma: large range of influence, boundary smooth

Effects of C and gamma



Scikit-learn implementation

- LinearSVC
  - only linear kernel, no kernel trick, but fast
- SVC
  - can choose different type of kernels, but slower, good for complex (nonlinear) small and medium datasets
  - Good with large number of features due to kernel tricks

References

- Aurélien Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow", O'Reilly Media, Inc., March 2017.