Geo Data Science

Support Vector Machines

Prof. Dr. Martin Kada

Chair Methods of Geoinformation Science (GIS)
Institute of Geodesy and Geoinformation Science

Technische Universität Berlin

Copyright Notice

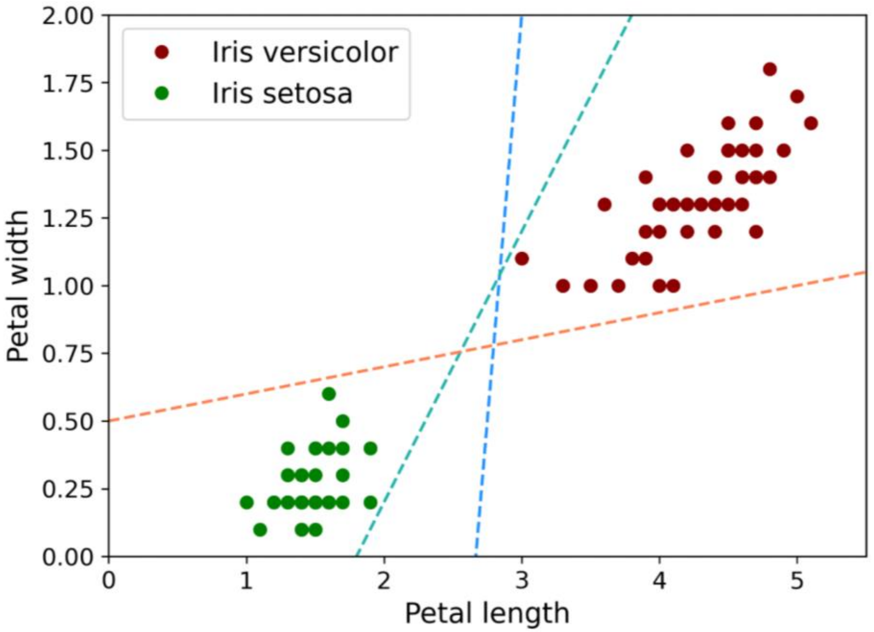
Content of this Lecture

- Logistic Regression
  - Binary classification
  - Linear model
  - Vulnerable to overfitting
    (especially with many features)
  - Based on linear regression,
    extension to multinomial regression

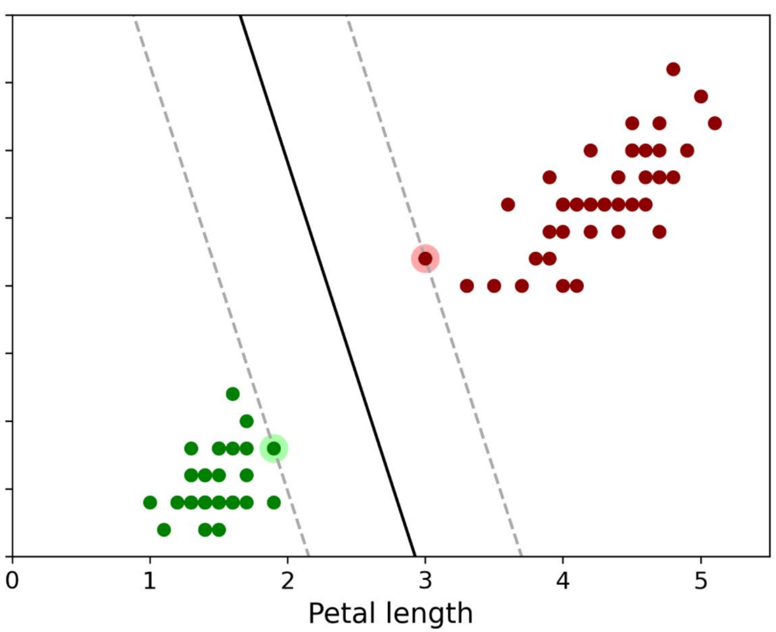
- Support Vector Machine (SVM)
  - Binary classification
  - Non-linear model using <u>kernels</u>
  - Less prone to overfitting due to <u>largest margin</u>
  - Can also be used for <u>regression</u>
    and <u>multi-class classification</u>
    (only very briefly covered in lecture)

Motivation

- Many possible ways to linearly separate two classes (e.g. of the Iris dataset)



(Very likely) poor linear classifier models as their decision boundaries come close to the training instances

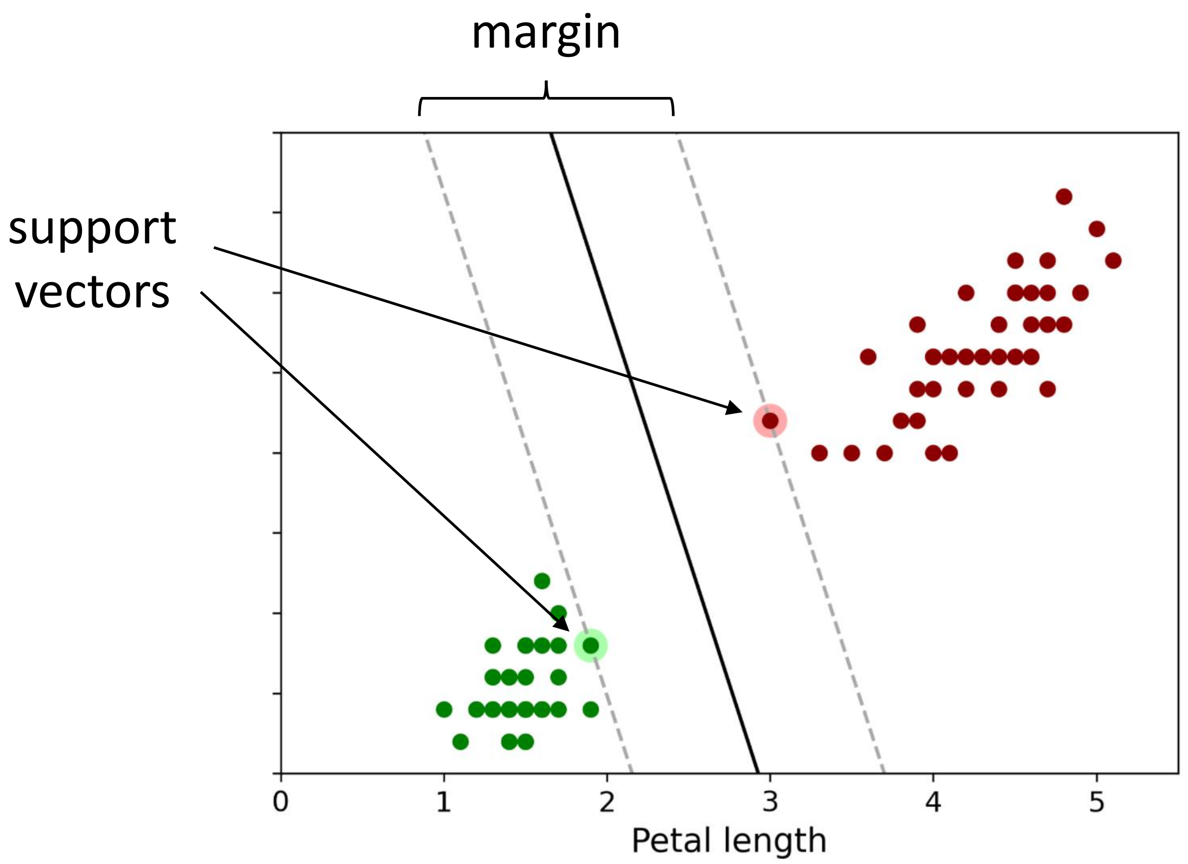


The decision boundary of a **support vector machine (SVM)** classifier is as far away from the closest training instance as possible

Motivation

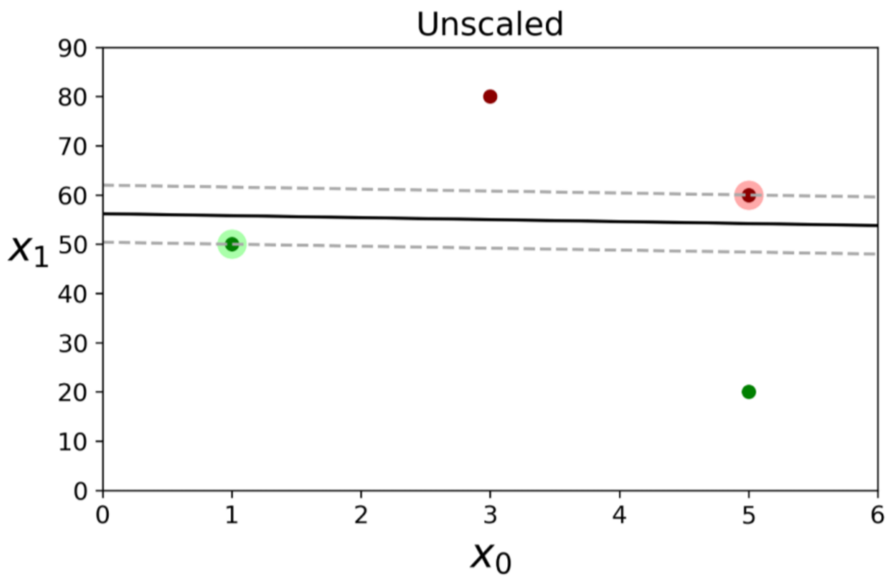
Technische Universität Berlin

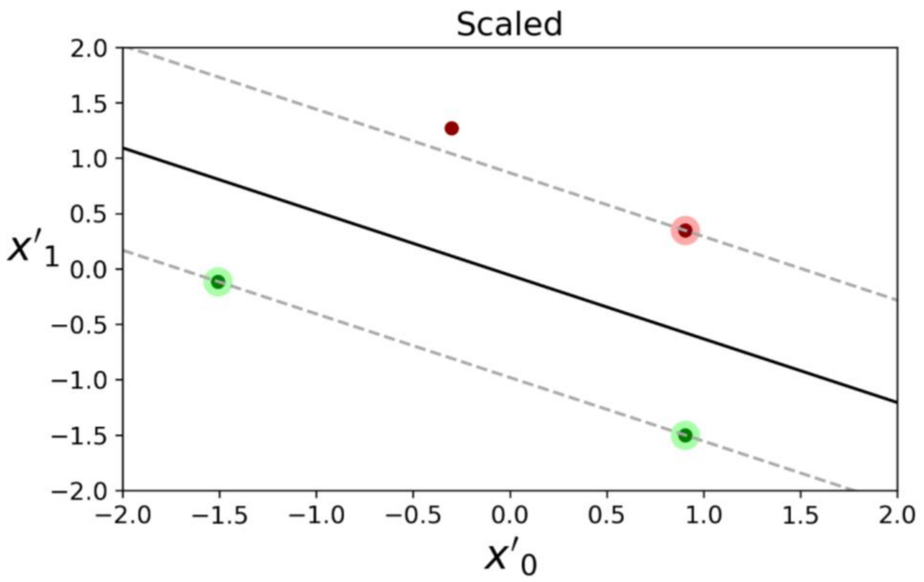
- Large Margin Classifier
  - The widest possible area around the class boundaries remains free of objects
  - Decision boundary determined by the support vectors
  - Further points outside the margin will not affect the decision boundary



Sensitivity to Feature Scales

- Support vector machines are sensitive to feature scales



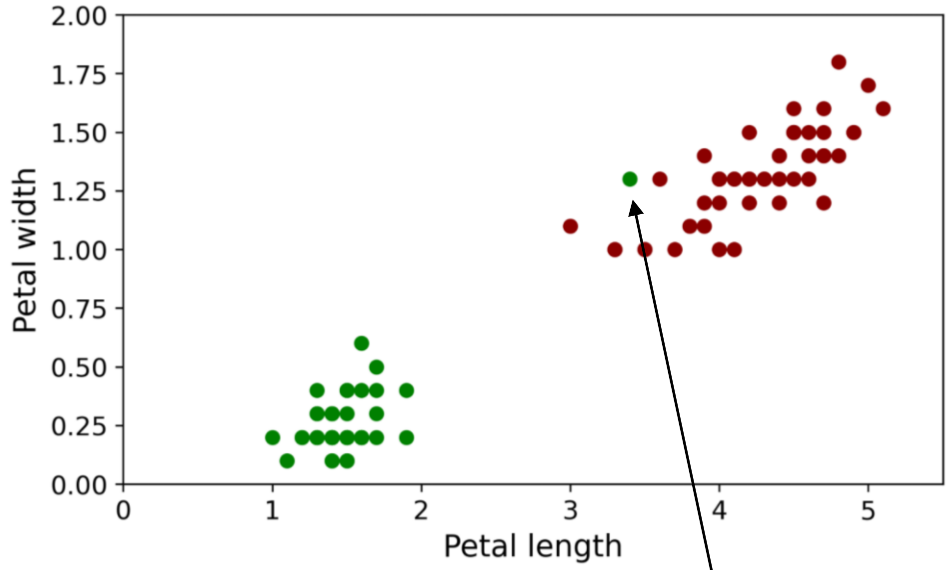


Fitted line is mainly influenced by the large differences in x1, as these have the highest impact on the calculated distances

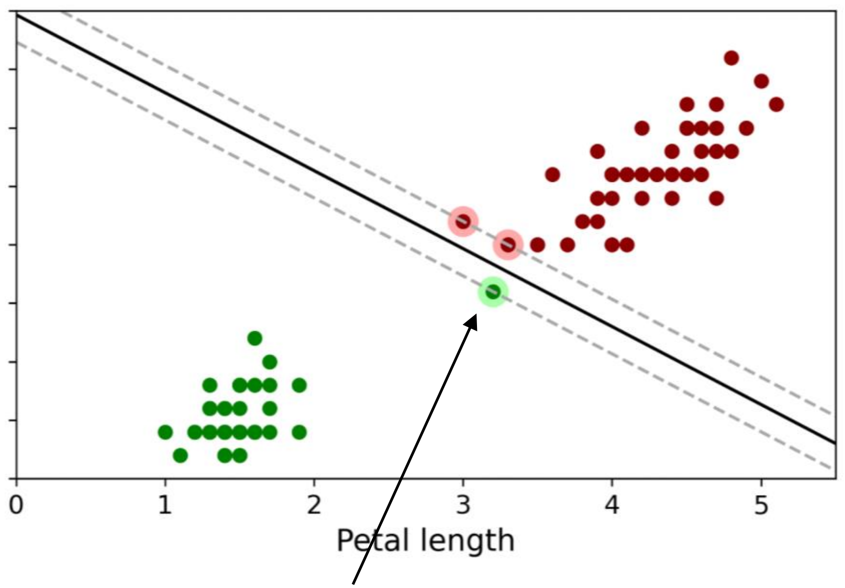
The fitted line is equally influenced by both features, since the differences have equal influence on the calculated distances

Hard vs. Soft Margin Classification

- Hard margin classification
  - All training instances must be outside the margin and on the correct side



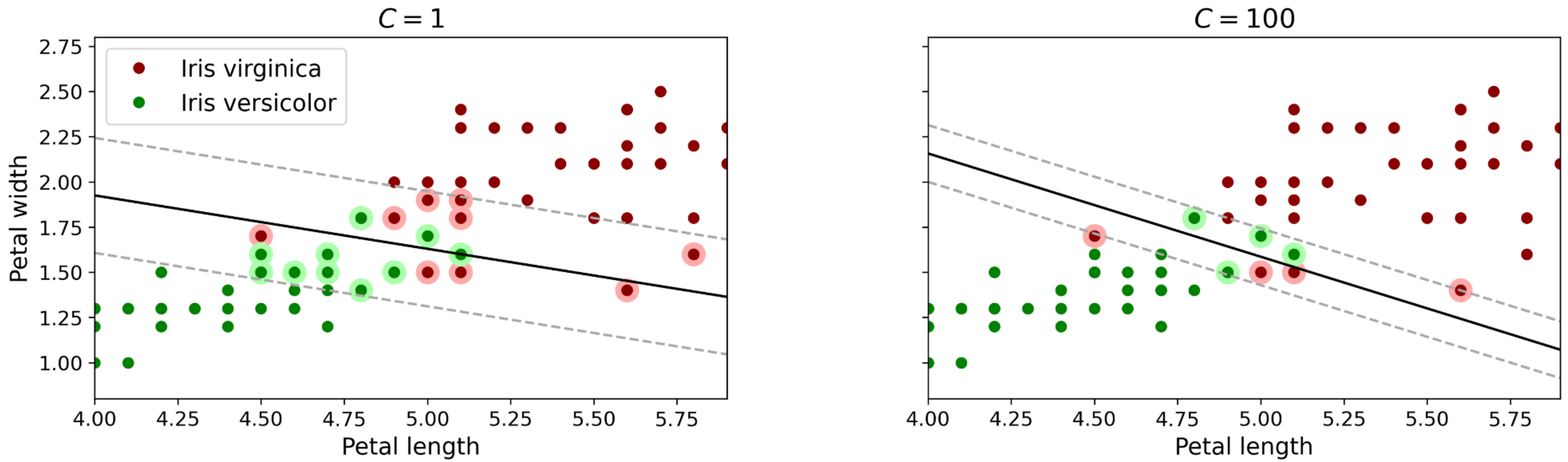
Impossible if data is not linearly separable (here because of the top green point)



Sensitive to outliers, which results in bad decision boundaries

Hard vs. Soft Margin Classification

- Soft margin classification
  - Flexible model with the objective to find a good balance between a large margin and a limited number of margin violations
  - Hyperparameter C determines a penalty that is added (to the cost value) for each misclassified training sample



Linear SVM Classifier Cost Function

number of samples

hyperparameter C

hinge loss

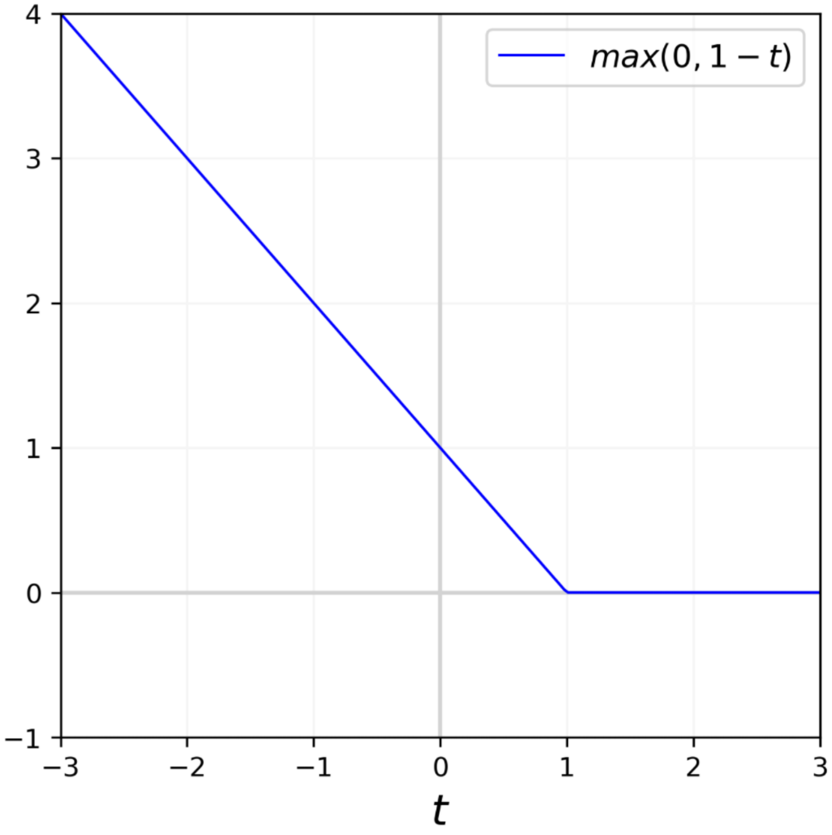
$$J(\theta) = \underbrace{\frac{1}{2}\sum_{i=1}^{n} \theta_i^2}_{\text{regularization to have a small weight vector}} + C \sum_{i=1}^{m} \overbrace{\max(0, 1 - y^{(i)}\theta^T \mathrm{x})}^{\text{hinge loss}}$$

regularization to have a small weight vector

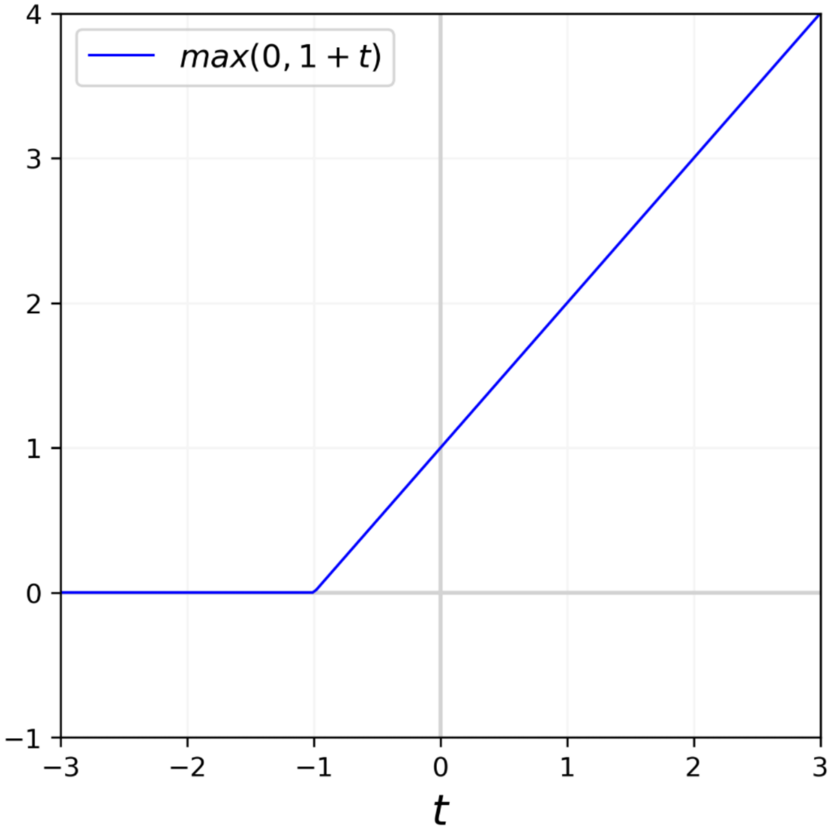
positive class ($y^{(i)} = +1$)
negative class ($y^{(i)} = -1$)

Hinge Loss

$y^{(i)} = +1$



$y^{(i)} = -1$



Linear SVM Classifier Cost Function

Technische Universität Berlin

number of samples

hyperparameter C

$$J(\theta) = \underbrace{\frac{1}{2}\sum_{i=1}^{n}\theta_i^2}_{\text{regularization to have a small weight vector}} + C \sum_{i=1}^{m} y^{(i)} \overbrace{\max(0, 1-\theta^T \mathrm{x})}^{\text{hinge loss for positive samples}} + (1-y^{(i)})\overbrace{\max(0, 1+\theta^T \mathrm{x})}^{\text{hinge loss for negative samples}}$$

positive class
$(y^{(i)}= 1)$

negative class
$(y^{(i)}= 0)$

Often seen alternative version
of the SVM cost function

Linear SVM Classifier Cost Function

Categories of points in cost function:

- Point is outside of margin ($y^{(i)}\theta^T\mathrm{x} > 1$) → no contribution to cost
- Point is on margin ($y^{(i)}\theta^T\mathrm{x} = 1$) → no contribution to cost (as in hard margin)
- Point violates margin constraint ($y^{(i)}\theta^T\mathrm{x} < 1$) → contributes to cost (linearly proportional to the distance of the point to the margin of this class)

If the support vector machine model is overfitting, then the model can be generalized by reducing the value for C
(→ increases the regularization part of the cost function)

Nonlinear SVM Classification

- Some datasets are not linearly separable
  - Introduce polynomial combinations of input features
  - High computational costs by the exploding number of features

- SVM with Polynomial Kernel using the kernel trick
  - Same results as adding polynomial features
  - But without explicitly adding any features
  - Hidden from the user by the implementation

Gaussian Radial Basis Function Kernel

- Similarity feature:
  - Add another feature that measures how similar each instance (of x) resembles some landmark ℓ
  - Gaussian Radial Basis Function (RBF)

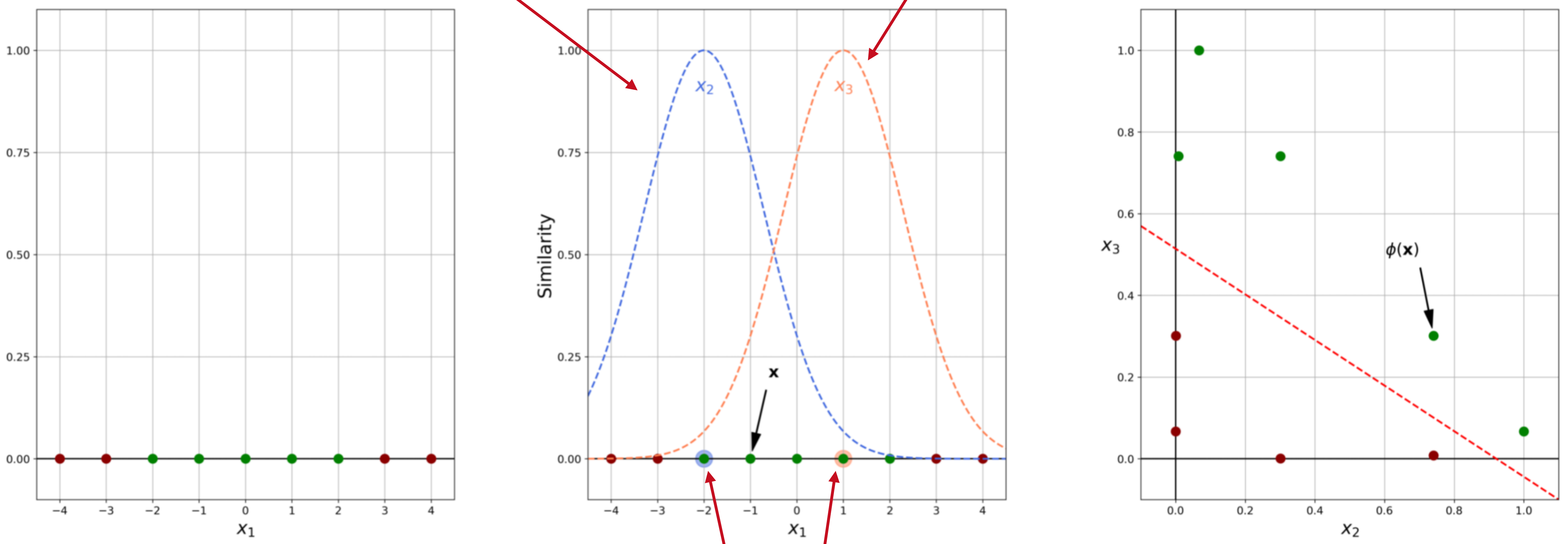
$$\phi_\gamma(\mathrm{x}, \ell) = \exp(-\gamma \|\mathrm{x} - \ell\|^2)$$

    is a bell-shaped function varying from 0 (very far away from the landmark) to 1 (at the landmark)
  - Hyperparameter γ determines the width of the bell function
    - Larger γ values → narrower curve (and vice versa)

Gaussian Radial Basis Function Kernel

landmark feature curve
for landmark x_2

landmark feature curve
for landmark x_3



linearly non-separable by
input feature x_1

landmarks

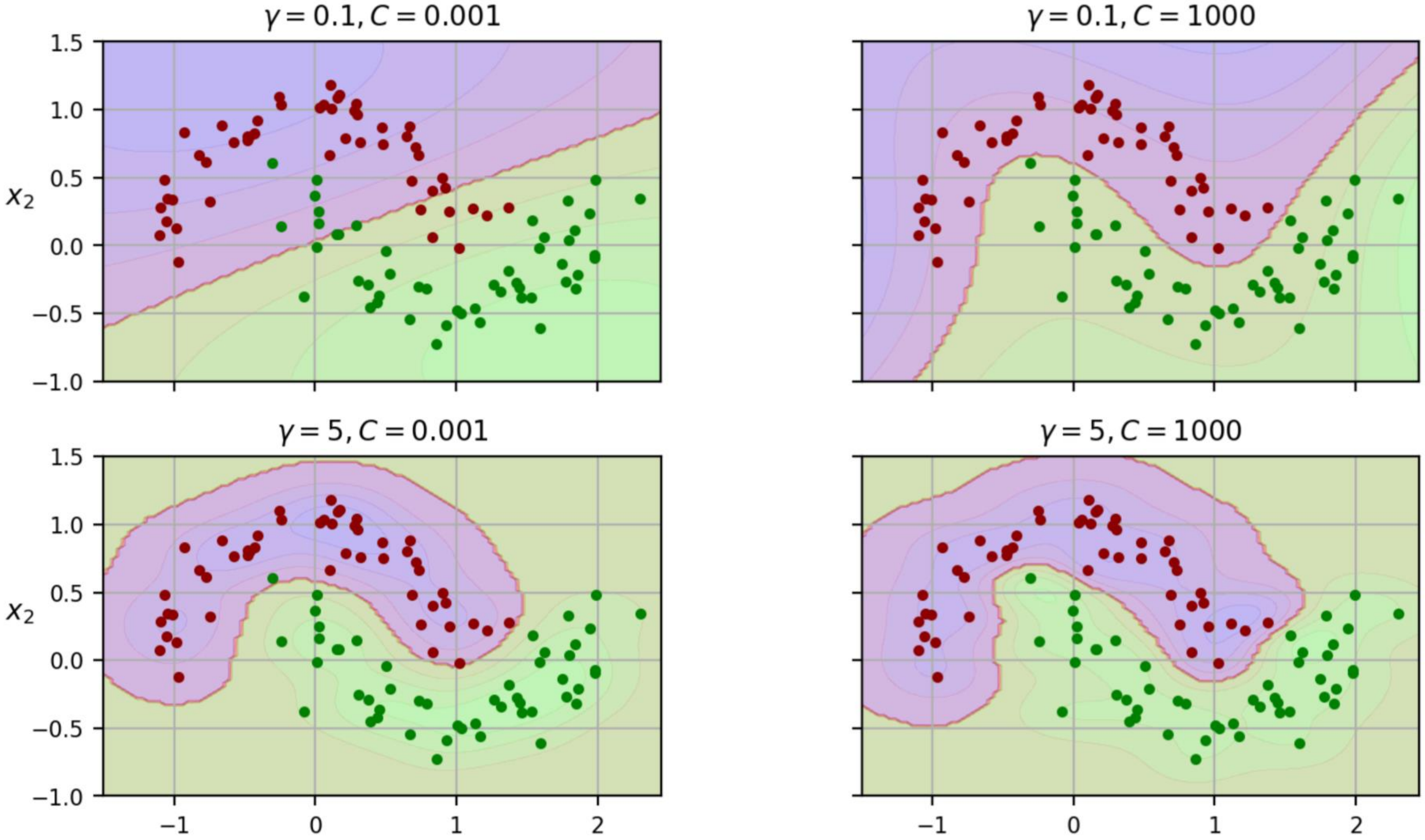
linearly separable by
landmark features x_2 and x_3

Gaussian Radial Basis Function Kernel

- Which landmarks to choose?

- Select a landmark at every instance of the dataset
  - Number of features of a training dataset with m instances is increased by m
  - Kernel trick leads to similar results, but with less computational burden
  - The many dimensions increase the chance that the transformed dataset is linearly separable

Hyperparameters γ and C with RBF



Hyperparameter γ

Hyperparameters γ acts like a regularization:

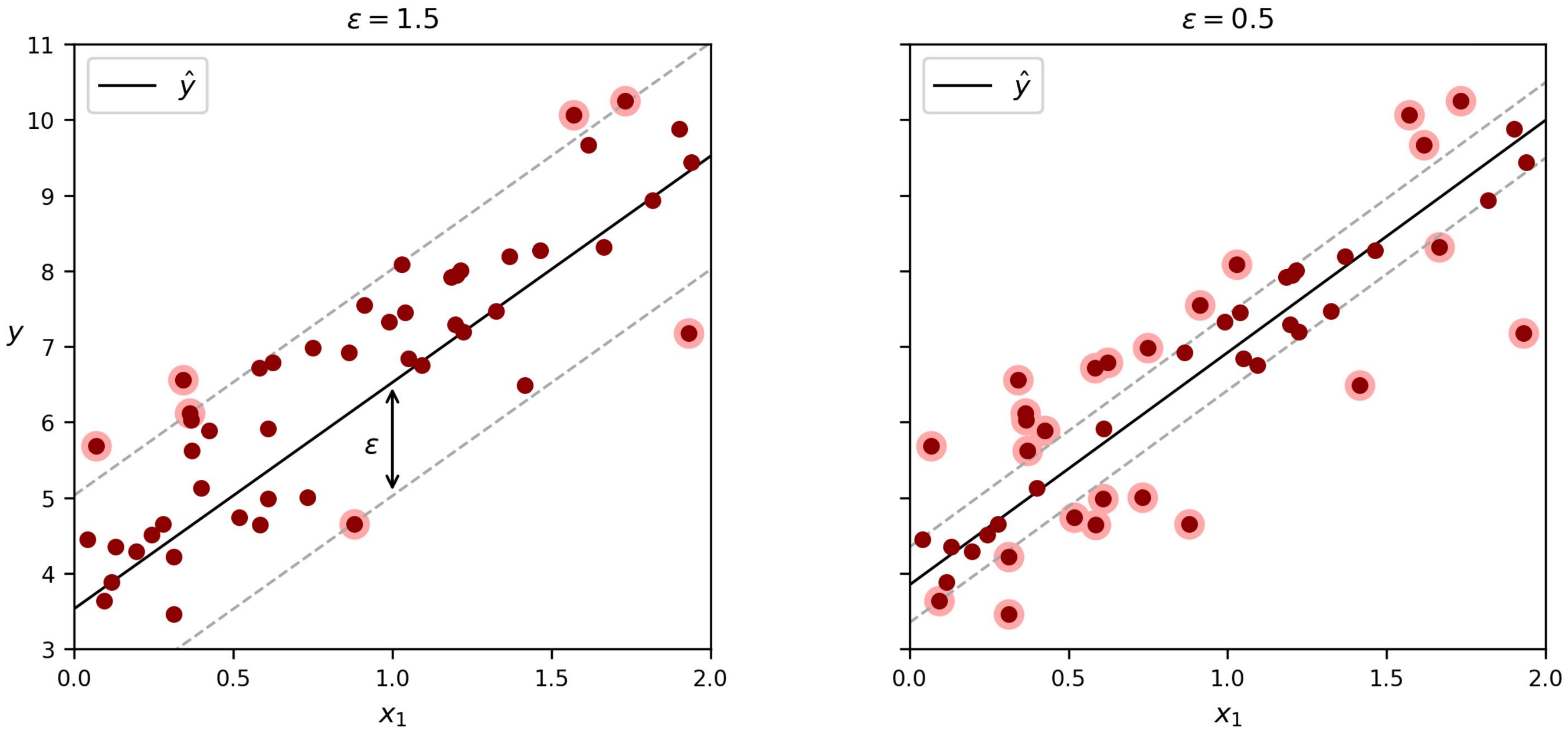
- If the model is overfitting, reduce the value of γ
- If the model is underfitting, increase the value of γ

SVM Regression

Technische Universität Berlin

- SVM supports linear and nonlinear regression
  - Change the objective:
    - Classification: fit a decision boundary between two classes with the largest possible margin and as few margin violations as possible
    - Regression: fit a line with a defined margin ε, so that as many training instances are located within the margin as possible, while limiting the number of instances outside the margin

SVM Regression



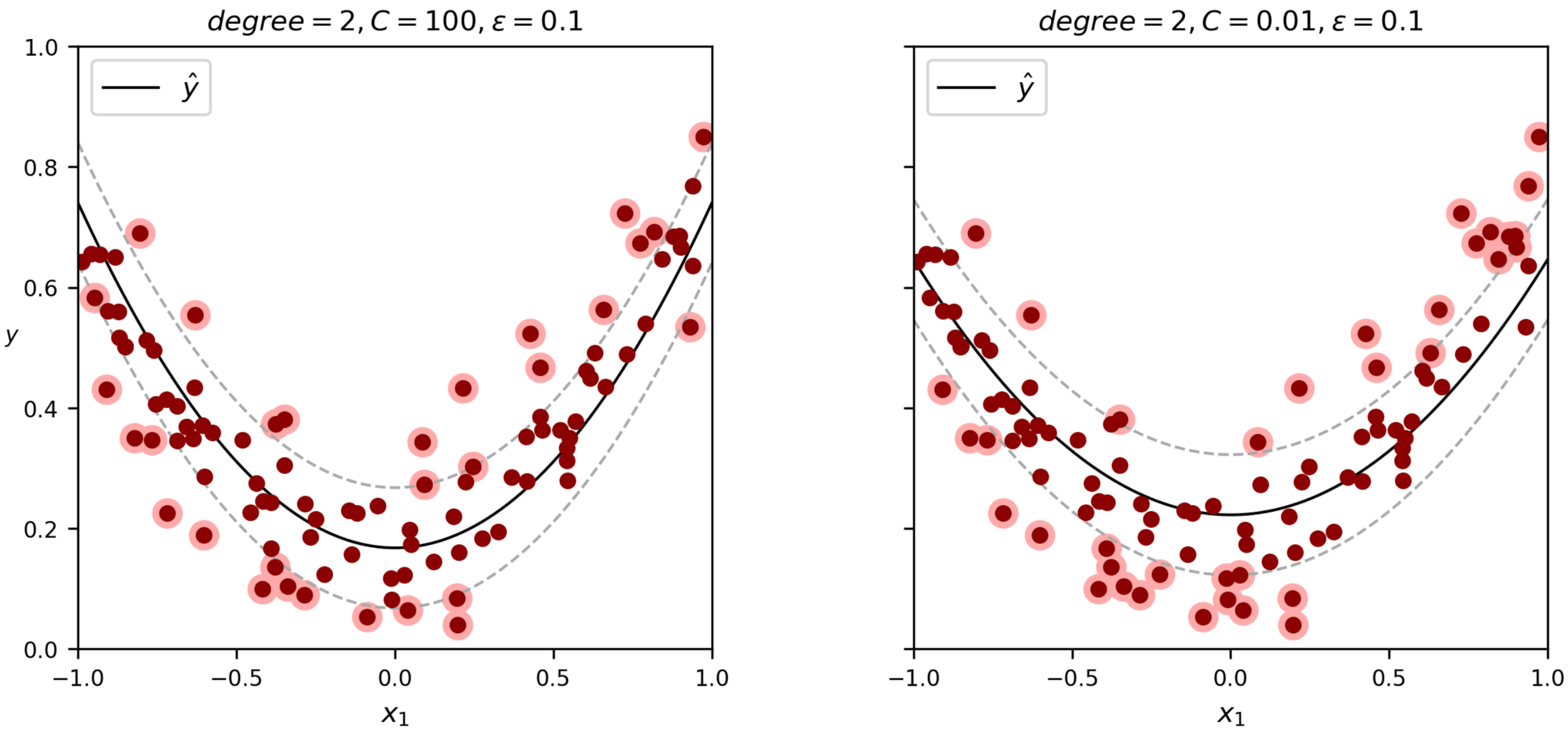
SVM Regression

- Non-linear regression:

    - Use (polynomial, RBF, or other) kernel

    - Hyperparameter C used for regularization
        - Large C value → little regularization
        - Small C value → more regularization

SVM Regression with Polynomial Kernel



Multiclass Classification with SVMs

- Multiclass classification SVM uses binary classification SVM:

  - One vs. one approach:
    - Each classifier separates points of two different classes
    - As many binary classifiers as there are pairs of classes

  - One vs. rest approach:
    - Each classifier separates points of one class with all other points
    - As many binary classifiers as there are classes

  - Combination of binary classifiers leads to multiclass classifier

Thank you for your attention!