Support Vector Machines (SVMs)

ML2: AI Concepts and Algorithms (SS2025)

Faculty of Computer Science and Applied Mathematics
University of Applied Sciences Technikum Wien

Lecturer: Rosana Gomes
Authors: B. Knapp, S. Lackner, S. Rezagholi, R.O. Gomes

Classification

KNN classification
Logistic regression
Naive Bayes
Support vector machines
Classification trees
Ensembles & boosting
Random Forest
Neural networks

Regression

KNN regression
Regression trees
Linear regression
Multiple regression
Ridge and Lasso regression
Neural networks

Supervised learning

Data handling

EDA
Data cleaning
Feature selection
Class balancing
etc

Clustering

k-means
Hierachical clustering
DB-scan

Non-supervised learning

AI

Dimensionality reduction

PCA / SVD
tSNE
MDS

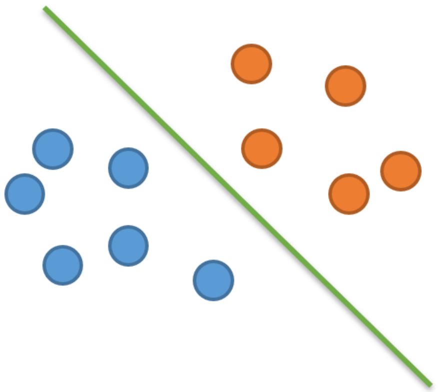
Reinforcement learning

Covered in a separate lecture.

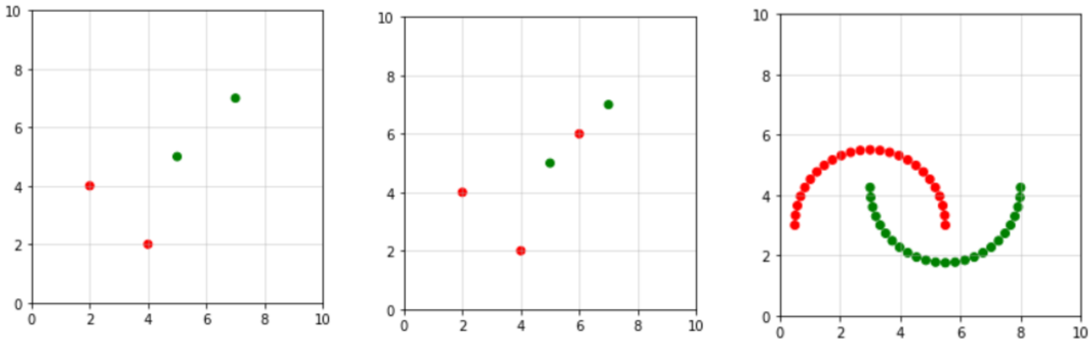
FH University of Applied Sciences TECHNIKUM WIEN

SVMs: Basic Idea

Find a hyperplane in n-dimensional space that separates datapoints from two classes.



Linear



Non-Linear

SVMs: Types

1. Maximum Margin Classifier (MMC)
   - For **perfectly linearly separable problems**.
2. Support Vector Classifier (SVC)
   - Accepts errors and is designed **for problems which are not linearly separable**.
3. Support Vector Machine (SVM)
   - Uses kernels to transform nonlinear problems into equivalent problems in higher-dimensional linear spaces.
   - Usable **for nonlinear separation problems**.

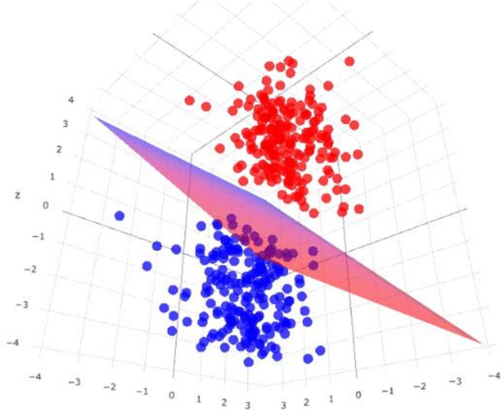
FH University of Applied Sciences TECHNIKUM WIEN

Hyperplanes

- **SVMs** use hyperplanes to separate data.
- A hyperplane is a **plane in higher dimensions**, formally a hyperplane in n-dimensional space:

$$\left\{ x \in \mathbb{R}^n : \alpha_1 x_1 + \ldots + \alpha_n x_n = c \right\}$$

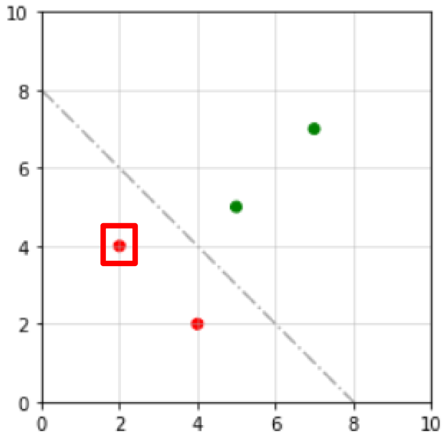
- A hyperplane is parametrized by $\alpha_1, \ldots, \alpha_n, c$.
- A hyperplane in 2 dimensions: A line.
- A hyperplane in 3 dimensions: A plane.



[https://www.kdnuggets.com/2019/09/friendly-introduction-support-vector-machines.html]

Hyperplanes for Classification

- To decide where a point is in relation to a hyperplane, plug the coordinate values of the point into the respective hyperplane's equation.
- This is the decision rule of SVMs.



This hyperplane (line) is characterized by the equation $x_1 + x_2 = 8$.

Points on the line: $\{x \in \mathbb{R}^2 : x_1 + x_2 - 8 = 0\}$

Points under the line: $\{x \in \mathbb{R}^2 : x_1 + x_2 - 8 < 0\}$

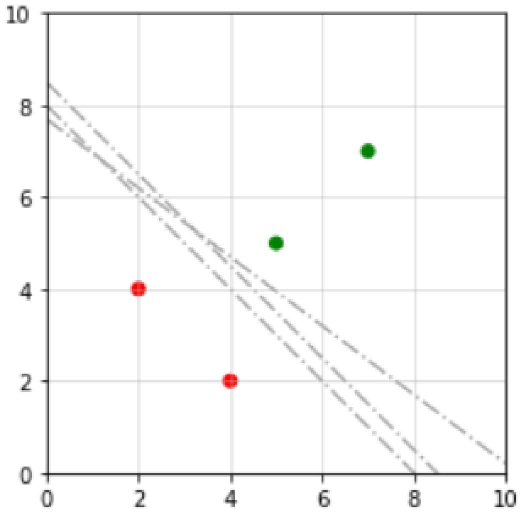
Points above the line: $\{x \in \mathbb{R}^2 : x_1 + x_2 - 8 > 0\}$

- Verify that the marked point is under the line.

FH University of Applied Sciences TECHNIKUM WIEN

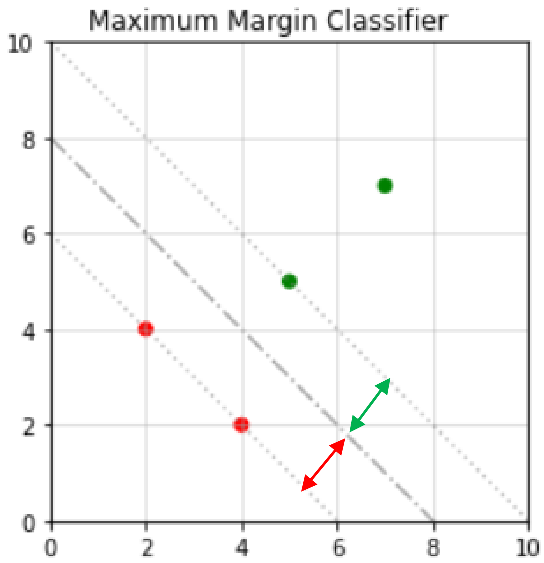
Which hyperplane separates best?

- For a linearly separable problem there are infinitely many separating hyperplanes.



These three hyperplanes separate the points.

Which hyperplane separates best?

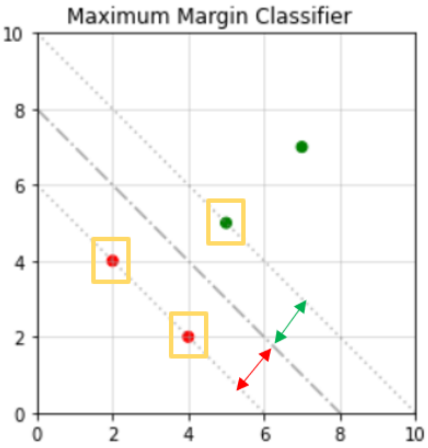


margin to green class = margin to red class

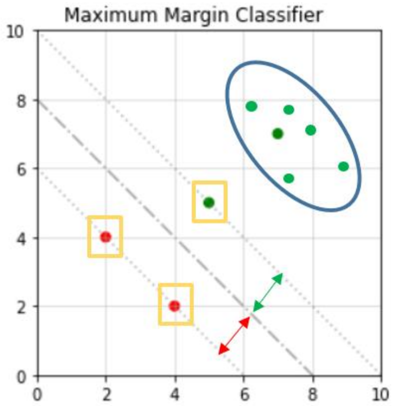
- The **maximum margin classifier** (MMC) chooses the hyperplane that maximizes the margin.
- The margin for a class is defined as the **distance** from the hyperplane to the **closest point(s)** to the hyperplane.

FH University of Applied Sciences TECHNIKUM WIEN

Support Vectors



- Only some of the points are important in defining the maximum margin hyperplane.
- These are called **support vectors.**
- Maximum margin classifiers are **very sensitive to outliers**.



Adding new points makes no difference if

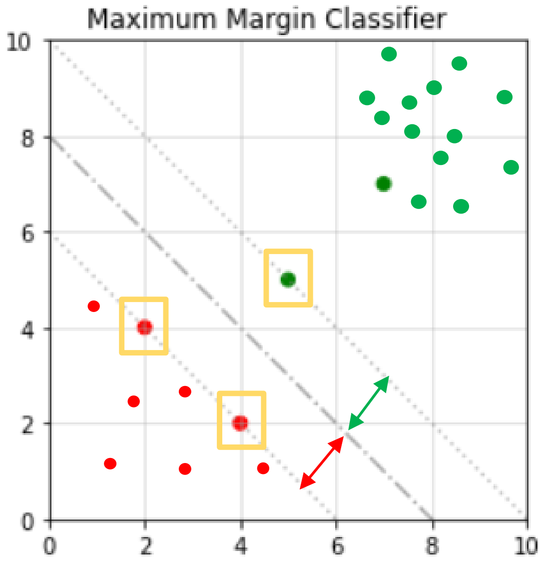
(1) the points do not produce an error, or

(2) the points do not become a support vectors.

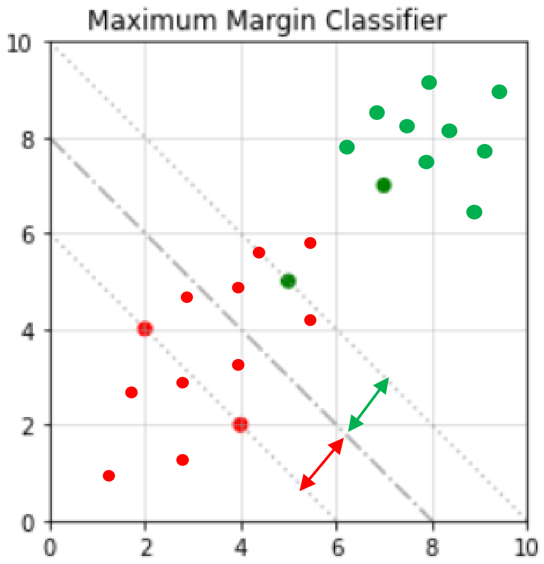
Support Vectors

One outlier can have large influence.

Training data

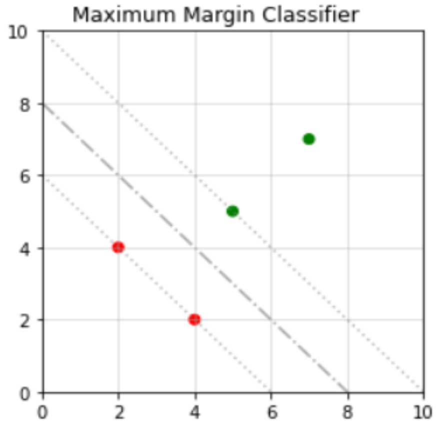


Test data



FH University of Applied Sciences TECHNIKUM WIEN

Formulation of Maximum Margin Classifiers (MMCs)



How to calculate the maximum margin hyperplane?

- Encode classes -1 or 1: $y_i \in \{-1, 1\},\ y \in \{-1, 1\}^n$
- The hyperplane equation imposes that

$$\alpha_1 x_1 + \ldots + \alpha_n x_n - c = \begin{cases} > 0 \text{ if } y_i = 1 \\ < 0 \text{ if } y_i = -1 \end{cases}$$

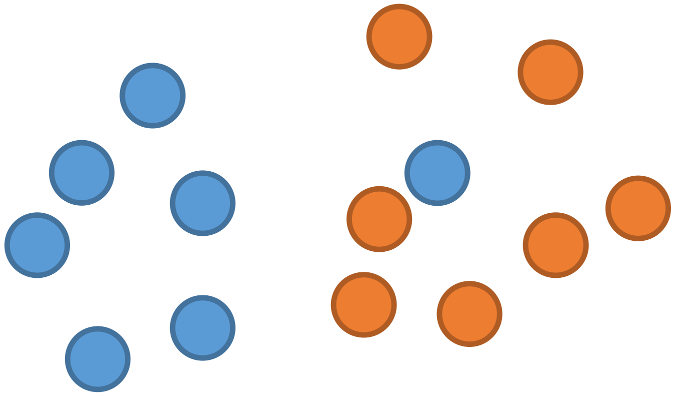
The *x-vector* defines how far a point is from the hyperplane (what side of the boundary condition). The *alpha-vector* controls the direction and orientation of the hyperplane, and parameter *c* shifts the hyperplane without changing its orientation.

We want to maximize the margin while keeping the expression true:

$$y_i\left(b + \alpha_1 x_{i,1} + \ldots + \alpha_p x_{i,p}\right) > 0 \text{ for all observations } i \in \{1, \ldots, n\}$$

FH University of Applied Sciences TECHNIKUM WIEN

Support Vector Classifiers (SVCs)



Motivation

Most real data is not linearly separable.

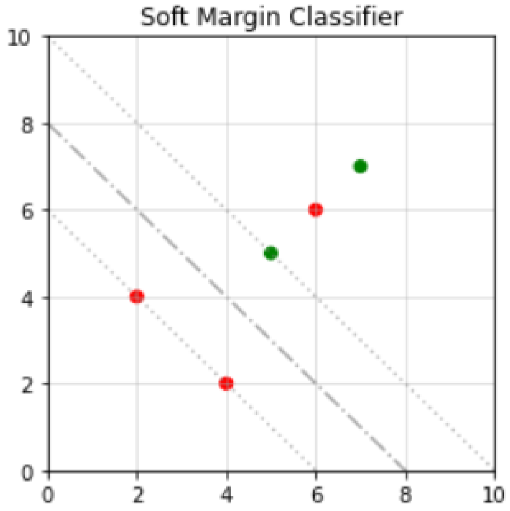
Maximum margin classifiers are of limited use.

FH University of Applied Sciences TECHNIKUM WIEN

Support Vector Classifiers: Basic Idea

Soft Margins

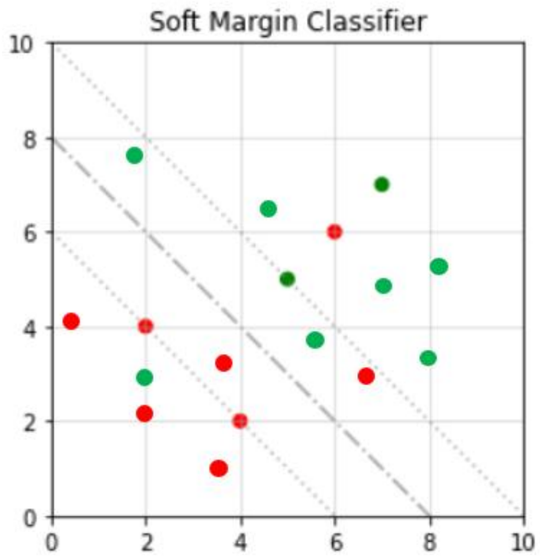
- The “hard margin” of maximum margin classifiers is too strict a criterion. Some error must be tolerated.
- The support vector classifier (SVC) is a variation of the maximum margin classifier (MMC) such that error can be accepted. SVCs are also called soft margin classifiers.



FH University of Applied Sciences TECHNIKUM WIEN

SVCs: Soft Margins

- In a support vector classifier both the margin and the hyperplane side can be violated. This allows
  - the construction of a hyperplane even if **no perfect separation** is possible,
  - to deal with **outliers.**
- The degree of allowed violations can often be regulated with a hyperparameter.
- All **points inside the margin are support vectors**.



Formulation of Support Vector Classifiers (SVCs)

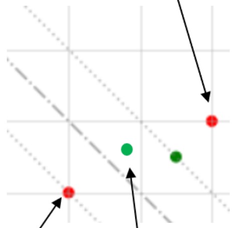
- Maximize margin obeying the following expressions:

$$y_i\left(b + \alpha_1 x_{i,1} + \ldots + \alpha_p x_{i,p}\right) > -\epsilon_i \text{ for all observations } i \in \{1, \ldots, n\}.$$

$\epsilon_1, \ldots, \epsilon_n$ are called slack variables.

One often enforces $\sum_{i=1}^{n} \epsilon_i \leq C$ where C is a hyperparameter of the algorithm.

C balances margin width vs classification accuracy.



$$\mathcal{L}_{\text{hinge}}(f(x_i), y_i) = \max(0, 1 - y_i f(x_i))$$

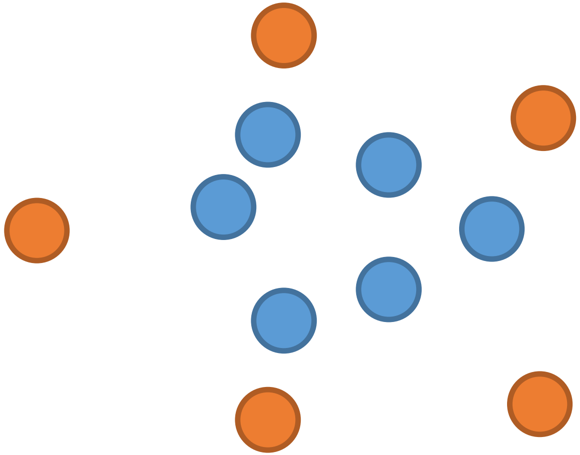
Where:

- $f(x_i) = \mathbf{w}^\top \mathbf{x}_i + b$ is the SVM decision function
- $y_i f(x_i)$ is positive when the classification is correct

FH University of Applied Sciences TECHNIKUM WIEN

Support Vector Machines: Motivation

Even a support vector classifier can not deal with this problem.



- MMCs and SVCs are meant for linear separation.
- Since SVCs are algorithms with very clear geometric meaning it would be nice to extend them to nonlinearly separable data.
- **Support Vector Machines (SVMs) Idea:** Embed the data into a higher-dimensional space in which it becomes linearly separable, do the separation there, and then map back to the original feature space.

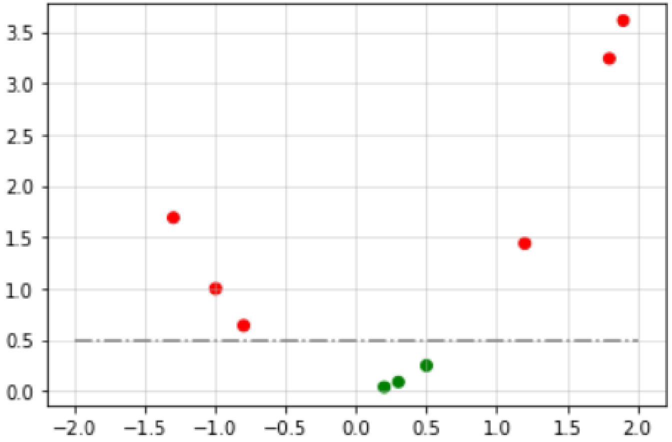
FH University of Applied Sciences TECHNIKUM WIEN

Example: Separability in Higher-Dimensional Space

Linearly nonseparable data in 1-dimensional space.

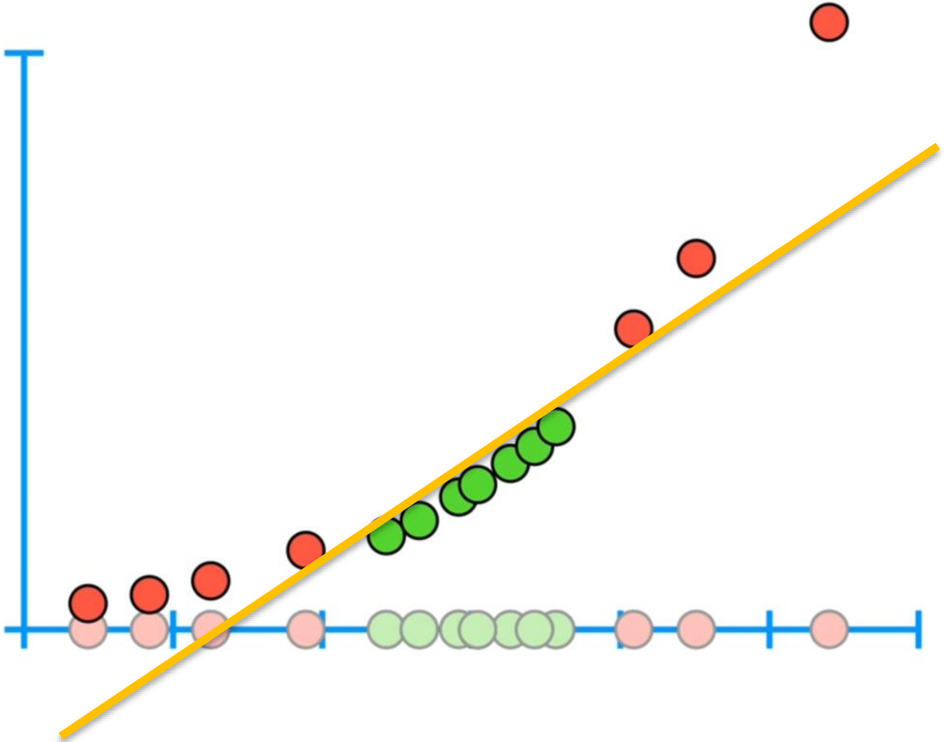
$$x \mapsto (x, x^2)$$

Introducing a second coordinate (the original value squared), makes the data linearly separable in 2-dimensional space.



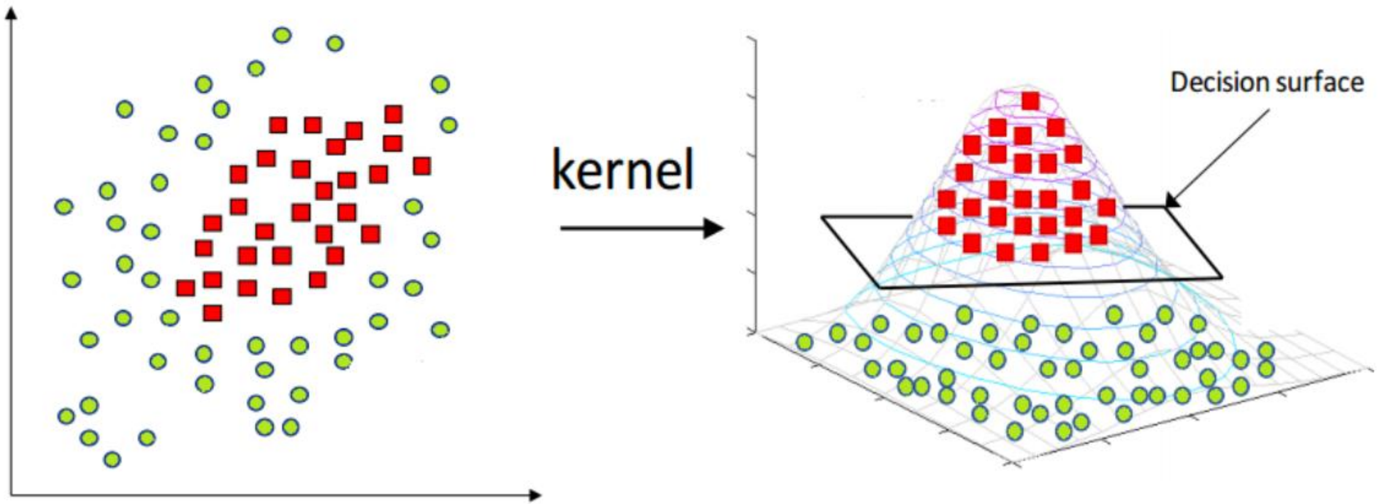
FH University of Applied Sciences TECHNIKUM WIEN

Example: Separability in Higher-Dimensional Space



[https://www.youtube.com/watch?v=efR1C6CvhmE]

Example: Separability in Higher-Dimensional Space



[https://www.kdnuggets.com/2019/09/friendly-introduction-support-vector-machines.html]

FH University of Applied Sciences TECHNIKUM WIEN

Separability in Higher-Dimensional Space: Challenges

How to find a good higher-dimensional embedding of the data?

In principle, there is always an embedding in an exponentially higher-dimensional space that allows separability:

Make non-linearly separable data linearly separable in some higher space

The dimensionality of the embedding needs to be limited for the outcome to be computationally feasible.

FH University of Applied Sciences TECHNIKUM WIEN

SVMs: The Kernel Trick

- In the context of SVMs **kernel functions** are **generalizations of the inner product** (dot product, scalar product).

$$K(x, x') = \langle \phi(x), \phi(x') \rangle$$

Kernel is a function that computes the **dot product between two data points in a higher-dimensional feature space**, without explicitly performing the transformation to that space (kernel trick).

- The inner product between orthogonal vectors is zero.
- The inner product can be thought of as a **measure of similarity** between the two vectors, the same is true for kernel functions.
- The function $\phi(x)$ does not need to be known.
- Kernel functions can be used to “hide” the higher-dimensional embedding.

FH University of Applied Sciences TECHNIKUM WIEN

- There are **specialized kernel functions in different application areas** (text, images, sound, …).

Kernel trick: Polynomial case

- The following function is a 3-dimensional embedding of our 2-dimensional data:

$$\phi([x_1, x_2]) = \left[x_1^2, \sqrt{2}x_1x_2, x_2^2\right]$$

- Degree d polynomial kernel with real parameter r is

$$k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$$

$$k(x, y) = (x \cdot y + r)^d$$

Hyperparameters r and d are found via cross-validation.

The polynomial kernel computes relationships between pairs of observations (x,y)

One can calculate that

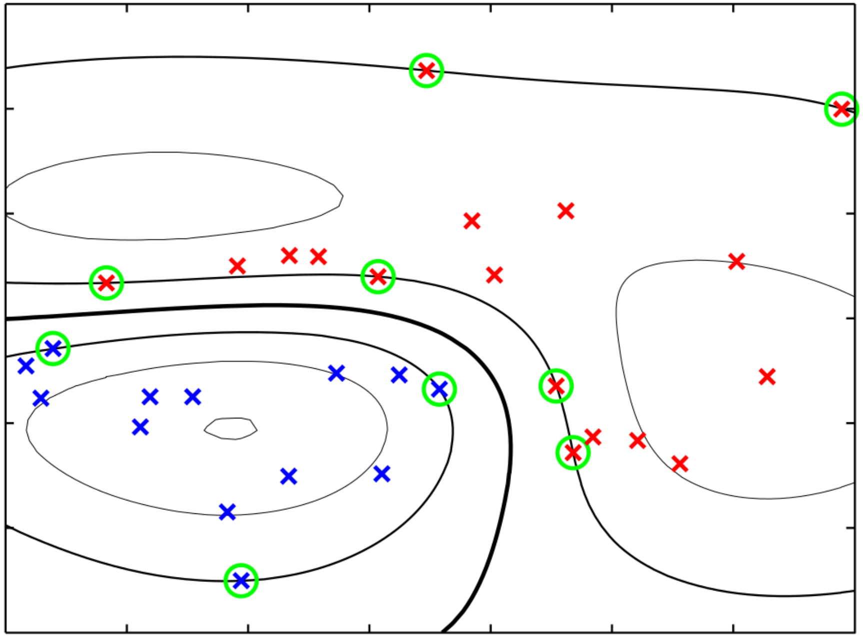
$$\phi(a) \cdot \phi(b) = (a \cdot b)^2 = k(a, b)$$

FH University of Applied Sciences TECHNIKUM WIEN

Kernel Trick: Save computation time by not calculating the actual transformation.

Kernel Trick & Hard Margin: Example

Example of synthetic data from two classes in two dimensions showing contours of constant $y(\mathbf{x})$ obtained from a support vector machine having a Gaussian kernel function. Also shown are the decision boundary, the margin boundaries, and the support vectors.

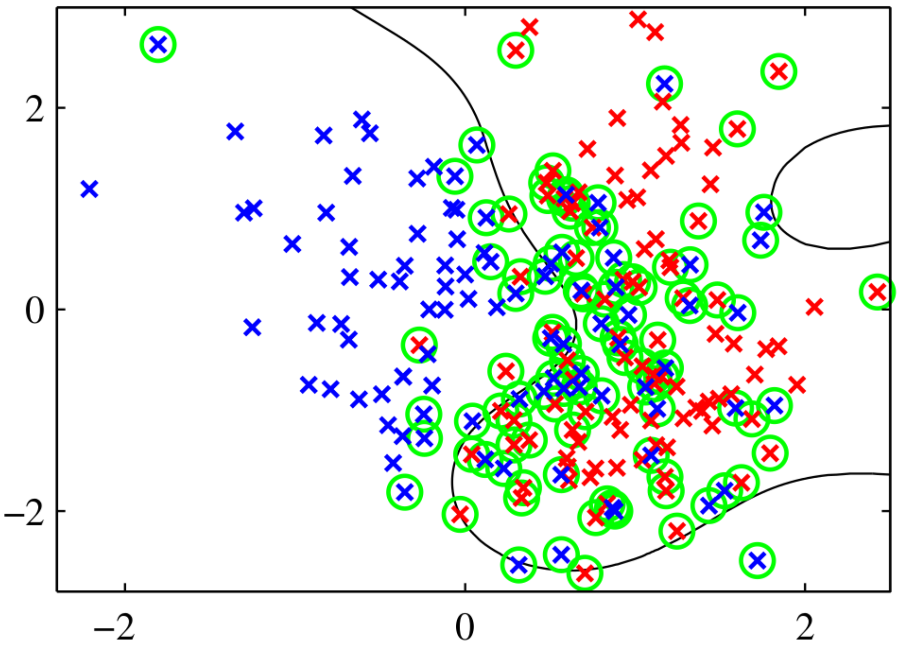


[Bishop (2006): Pattern Recognition and Machine Learning, Figure 7.2]

FH University of Applied Sciences TECHNIKUM WIEN

Kernel Trick & Soft Margin: Example

Illustration of the ν-SVM applied to a nonseparable data set in two dimensions. The support vectors are indicated by circles.



[Bishop (2006): Pattern Recognition and Machine Learning, Figure 7.4]

FH University of Applied Sciences TECHNIKUM WIEN

Kernel Trick: Radial Basis Kernel

- Popular kernel for SVMs, corresponds to an embedding into an infinite-dimensional function space.
- Can be interpreted as a Gaussian similarity function being assigned to each datapoint.
- Will not be discussed in this course (see [this video])

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right)$$

FH University of Applied Sciences TECHNIKUM WIEN

Takeaway

- SVMs are good “of-the-shelf“ algorithms. Nowadays boosting is often preferred, but it can lead to larger and slower models.
- SVMs can be used for classification (binary and multiclass), regression, and outlier detection.(The classical application is to binary classification.
- SVMs can sometimes solve nonlinear problems (kernel trick).
- SVMs can be sensitive to outliers.
- Fitting SVMs is computationally intensive (for large high-dimensional datasets).
- SVMs are mathematically more complicated than some other algorithms (especially the kernel-trick).

FH University of Applied Sciences TECHNIKUM WIEN

Next Up: Gaussian Mixture + RECAP class

Assignment: SVMs

a) Explain SVMs and the kernel trick in 1 page/slide each.

Use self-made images or even hand drawings (of which you take a photo).

Use self-written explanations.

Do not copy from the lecture slides or the internet (neither text nor images).

b) Implement an MMC by yourself and test it on a 2D dataset of your choice.

If you cannot find a better solution you can just approximate the hyperplane that maximizes the margin by brute force.

FH University of Applied Sciences TECHNIKUM WIEN

References

- Géron A. (2017): Hands-On Machine Learning with Scikit-Learn & Tensorflow. – O‘Reilly.
- James G., Witten D., Hastie T., Tibshirani R. (2017): An introduction to Statistical Learning. – Springer.
- Kuhn M., Johnson K. (2016): Applied Predictive Modeling. – Springer.
- Berk R. (2016): Statistical Learning from a Regression Perspective. – Springer.

FH University of Applied Sciences TECHNIKUM WIEN