LECTURER: TAI LE QUY

MACHINE LEARNING

SUPERVISED LEARNING

SUMMARY

OVERALL WRAP UP

STUDY GOALS

- Overlook the entire **course structure**
- Explain the **main concepts** and **key findings**
- Discuss some big data **processing solutions**

EXPLAIN SIMPLY

1. What are the six phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM) design cycle?

2. Explain in one sentence what the kernel trick is.

3. What is meant by the term "**ensemble model**"?

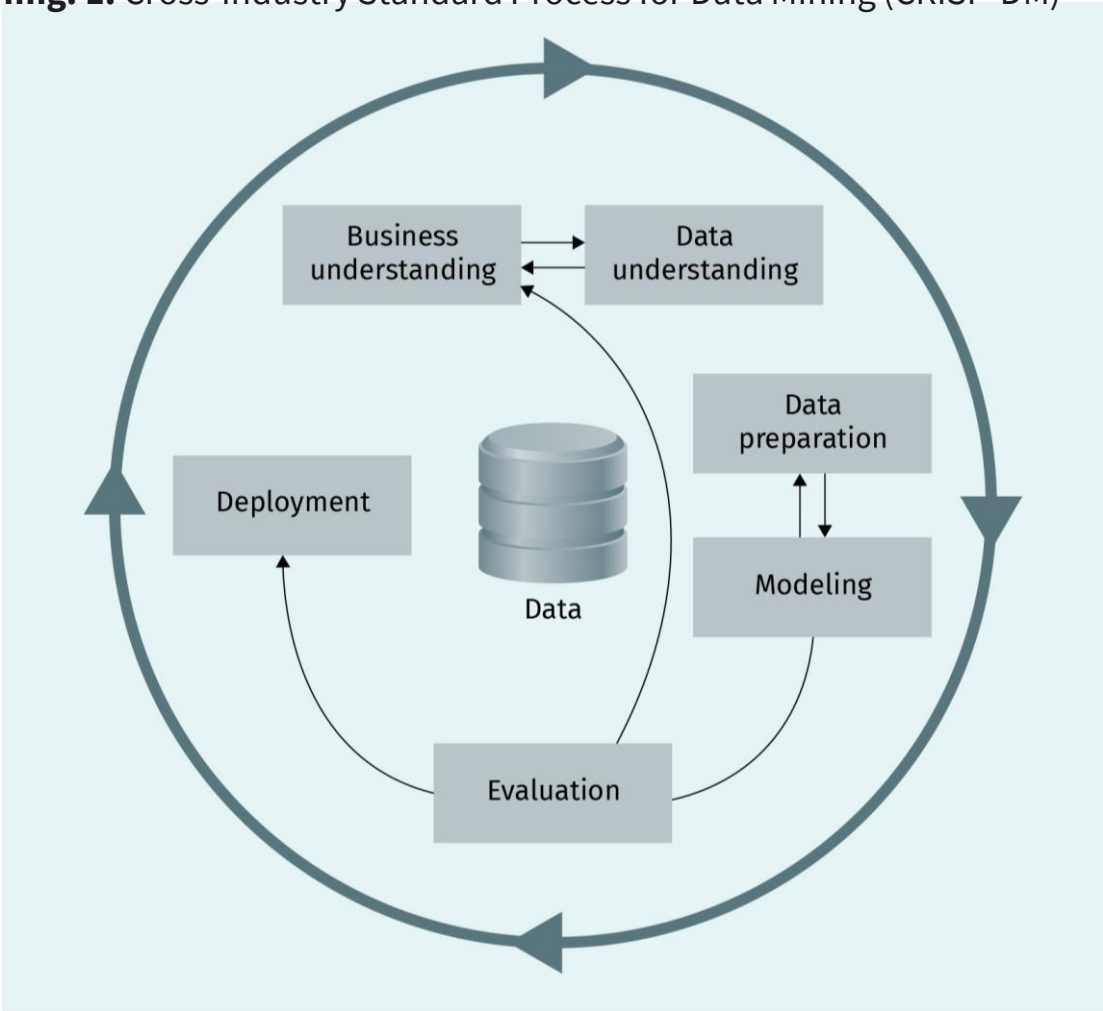
Tab. 1: Supervised and unsupervised machine learning

	Feature 1	Feature 2	Label
Sampe 1	$x_{1,1}$	$x_{2,1}$	y_1
Sample 2	$x_{1,2}$	$x_{2,2}$	y_2
Sample 3	$x_{1,3}$	$x_{2,3}$	y_3

Source of table 1: Christian Müller-Kett, 2021

THE MACHINE LEARNING DESIGN CYCLE

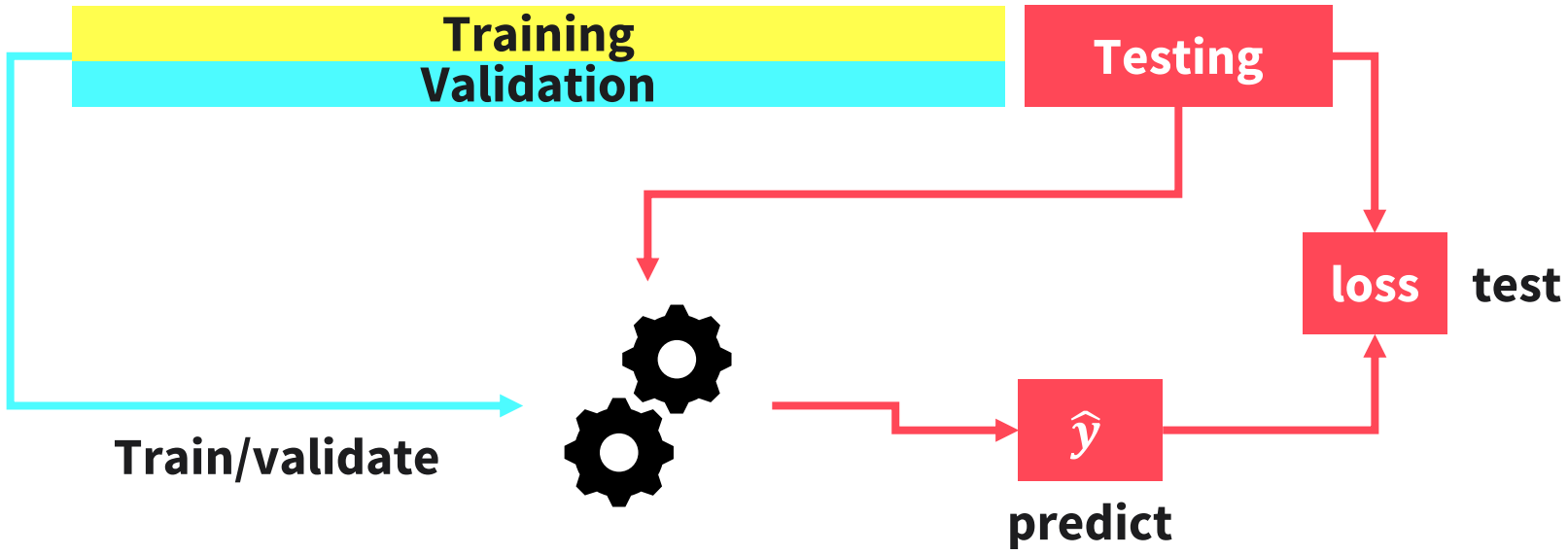
Img. 1: Cross-Industry Standard Process for Data Mining (CRISP-DM)



Source of image 1: Jensen, 2012 based on Wirth & Hipp, 2000.

CROSS-VALIDATION

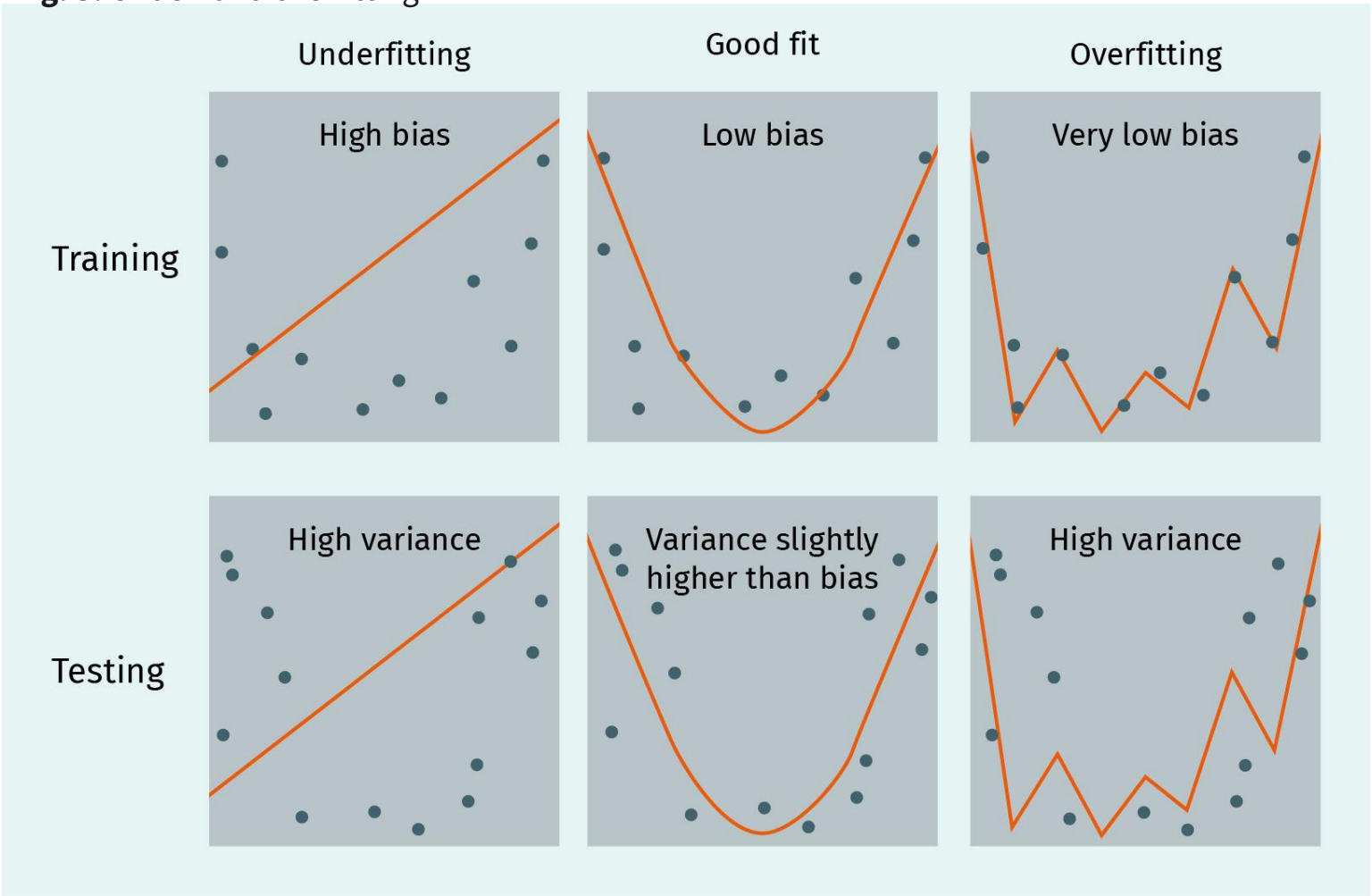
Img. 2: Cross-validation



Source of image 2: Christian Müller-Kett, 2019

UNDER- AND OVERFITTING

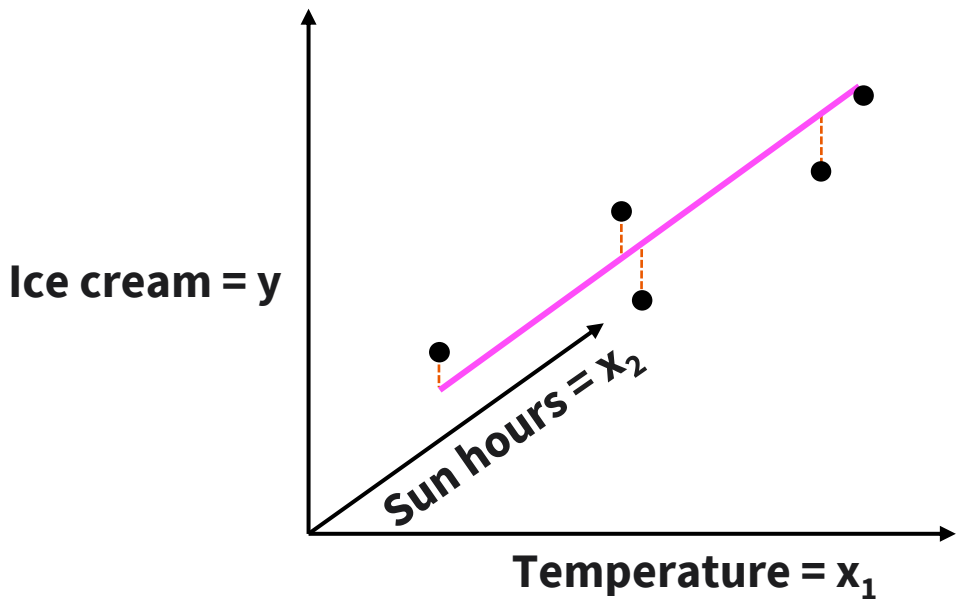
Img. 3: Under- and overfitting



Source of image 3: Christian Müller-Kett, 2022

LINEAR REGRESSION

Img. 4: Linear regression



$$y = \omega_0 + \sum_{i=1}^{n} \omega_i * x_i + \varepsilon$$

$$Residual = R_i = y_i - \hat{y}_i$$

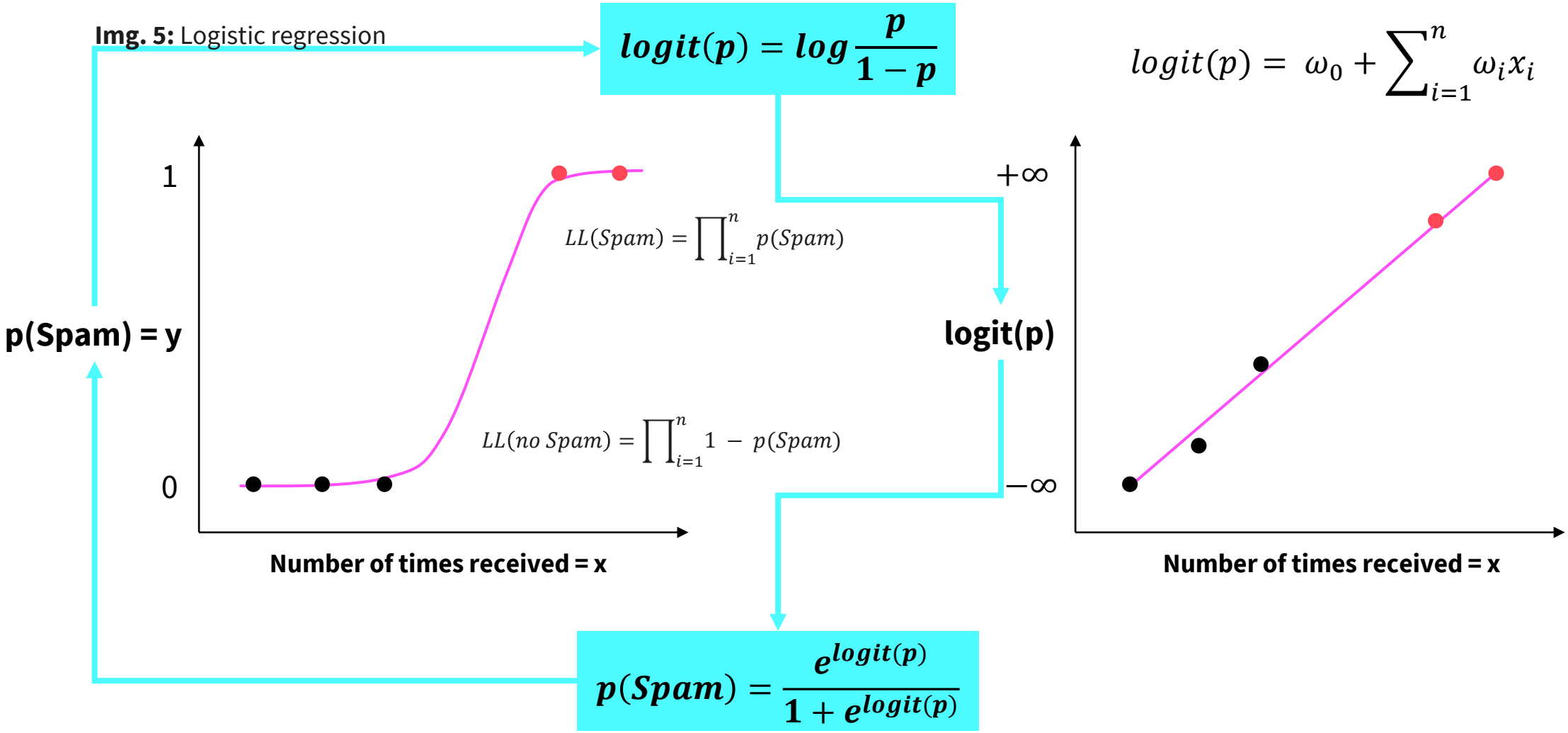
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

Source of image 4: Christian Müller-Kett, 2021.

LOGISTIC REGRESSION

Img. 5: Logistic regression



Source of image 5: Christian Müller-Kett, 2021.

MODEL PERFORMANCE METRICS

- $\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
- $\boldsymbol{Sensitivity} = \boldsymbol{Recall} = \frac{TP}{TP+FN}$
- $\boldsymbol{Specificity} = \frac{TN}{TN+FP}$
- $\boldsymbol{Precision} = \frac{TP}{TP+FP}$
- **ROC** curve: Sensitivy vs. 1-Specificity
- Area Under the Curve (**AUC**)

Tab. 2: Confusion matrix

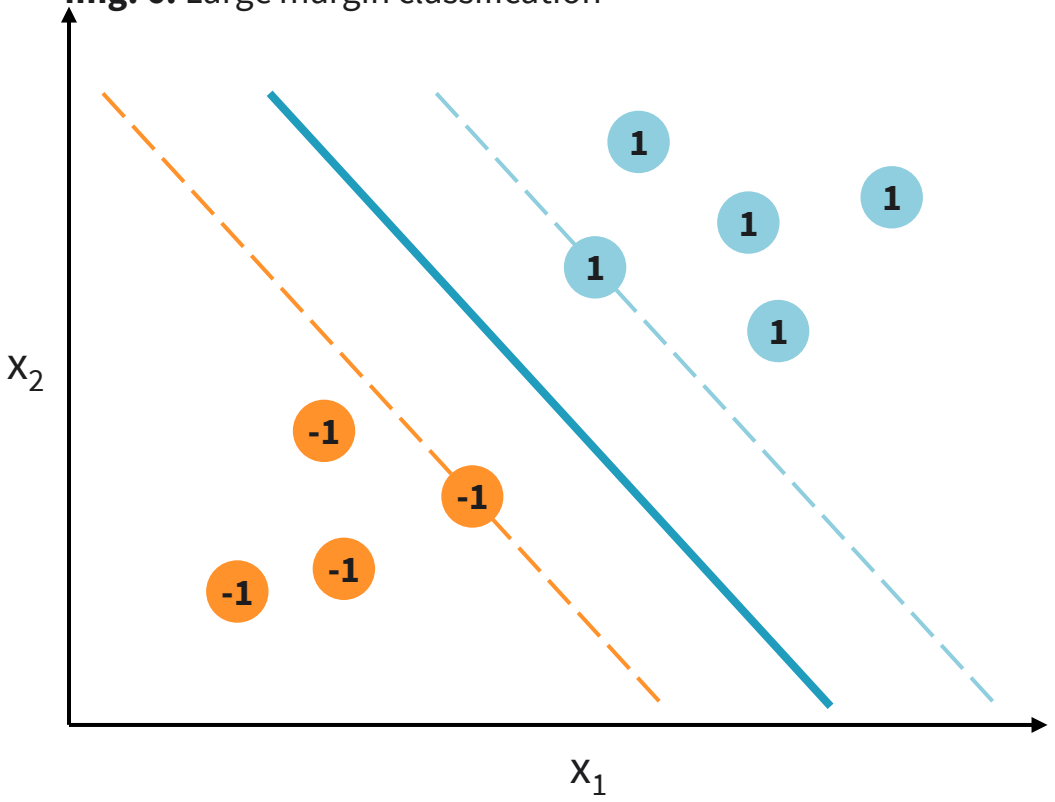
	Predicted +	**Predicted -**
Actual +	True **+** (TP)	False **-** (FN)
Actual -	False **+** (FP)	True **-** (TN)

Source of table 2: Christian Müller-Kett, 2023

1. **Margin** (dotted lines) are described by $\boldsymbol{y_i(\vec{w} \cdot \vec{x} - b) = 1}$
2. **Maximize** $\frac{2}{\|w\|}$

– Margin maximization depends on $\boldsymbol{x_i \cdot x_j}$

Img. 6: Large margin classification

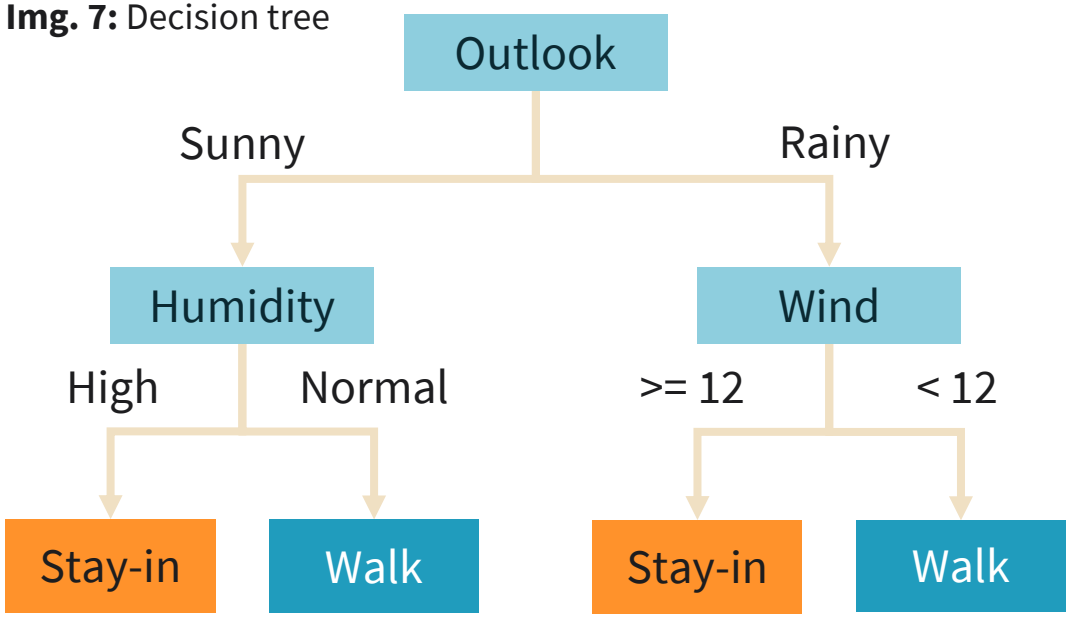


Source of text: Boehmke & Greenwell, 2019, Hastie et al., 2017
Source of image 6: Christian Müller-Kett, 2021

Tab. 3: Training data

Outlook	Humidity	Wind	Label
Rainy	High	20 km/h	Stay
Sunny	Normal	4 km/h	Walk
Sunny	Low	18 km/h	Walk
…	…	…	…

Img. 7: Decision tree



Source of image 7: Christian Müller-Kett, 2021
Source of table 3: Christian Müller-Kett, 2021

REVIEW STUDY GOALS

- Overlook the entire **course structure**
- Explain the **main concepts** and **key findings**
- Discuss some big data **processing solutions**

- **Combine weak estimator** to form **one strong estimator**
- **Bagging**
  - Several decision trees trained by **varying data subsets**
  - The final prediction as the **majority** or **average** of individual predictions
- **Boosting**
  - Several decision trees trained **sequentially**
  - Each tree is trained with a dataset **exaggerating** the **misclassified samples** from the previous tree

Source of text: Boehmke & Greenwell, 2019, Breiman et al., 1984, Mitchell, 1997

SESSION 1

TRANSFER TASK

TRANSFER TASKS

A start-up that sells **sustainable products in smaller stores** has been very successful in recent years. As a result, more stores are to be opened worldwide.

As a Data Scientist, you and your team are tasked with training a **machine learning model predicting product demand** one week ahead. Eventually, this model is supposed to be connected to the **company's ordering system**, giving well-informed advice about how many products should be ordered per store.

Create a rough **project plan** and briefly describe the **work items for each of the project's phases**.

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

LIST OF SOURCES

Breiman, L., Friedman, J., Olshen, R. A., & Stone, J. S. (1984). *Classification and regression trees*. Chapman & Hall. https://doi.org/10.1201/9781315139470

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with R*. Chapman & Hall.

Hastie, T., Tibshirani, R., Friedman, J. H. (2017). *The elements of statistical learning. Data mining, inference, and prediction*. Second edition. New York, NY: Springer.

Jensen, K. (2012). CRISP-DM process diagram (CRISP-DM_process_diagram.png) [Illustration]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29—39). Springer.

How did you like the course?