Fachhochschule
Dortmund
University of Applied Sciences and Arts

# Data Science

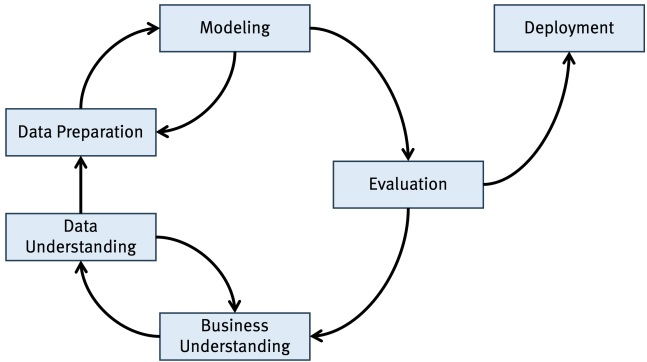
## 13: Machine Learning in a nutshell

Klaus Kaiser
WiSe 2024 / 2035

# Data Science

## Recap






**CRISP-DM** process model gives an overview on the steps needed for developing a machine learning model.

# Data Science

## Recap



Already seen: Depending on the source of data, the quality can differ extremely

- There could be different formats (e.g. German or international date format)
- There could be different sources which needs to be joined
- There could be missing values
- ...

The process of preparing one or multiple data sets for machine learning is called **data preparation**.

# Data Science

## Recap



A decision tree is an ordered directed tree, by following the tree to leafs decisions are done.

- In every node a test on an attribute of the data is done (e.g. height of a person is larger than 1.7 m)
- Each leaf describes an outcome, i.e. a class (e.g. van, sports car etc.)

# Data Science

## Recap



The **entropy** $H(M)$ of a set $M$ can be computed by

$$H(M) = \sum_{x \in M} -p_x \cdot \log_2(p_x)$$

where $p_x$ is the relative frequency of an element $x$ in the set $M$. Here, we define $0 \cdot \log_2(0) = 0$.

# Data Science

## Recap



For a split $S = \{s | s \subset M\}$, we can compute the **gain** $H_G$ **of order** by

$$H_G = H_M - \oplus(S),$$

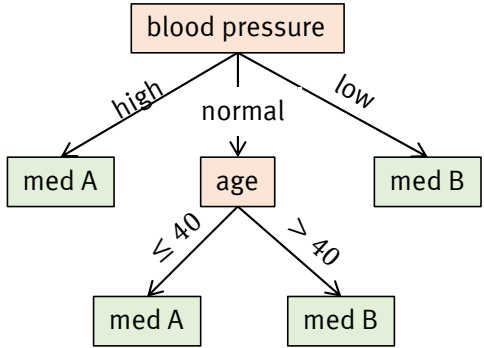
where $\oplus(S)$ denotes the weighted sum of the entropy of the splitted subsets.

$$\oplus(S) = \sum_{s \in S} \frac{|s|}{|M|} H(s).$$

# Data Science

## Recap





# Data Science

## Recap

we focus on students

Decision trees are a simple method for decision or classification problems.

### Pros

- Easy to be implemented (if-then-rules)
- Fast in usage (only follow the path through tree)
- Explainable and easy to understand (no black-box)
- Can be extended to regression problems

### Cons

- Not useful for numerical attributes (discretization)
  - Need to define fixed boundaries (e.g. $age < 40$)
- Could lead to small-scale distinctions (overfitting)

# Data Science

## Today



### Evaluation of models

# Machine Learning in a nutshell

# Machine Learning in a nutshell

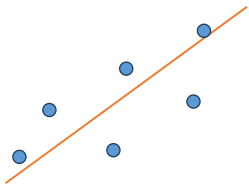
## Evaluating models

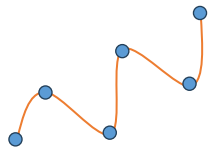
# Data Science
## Evaluating models



A machine learning model, like decision trees or linear regression, are fitted to a given dataset.

- Training process finds optimal fit for the given data
- **Problem:** The perfect fit for the training data might not be the perfect general model!

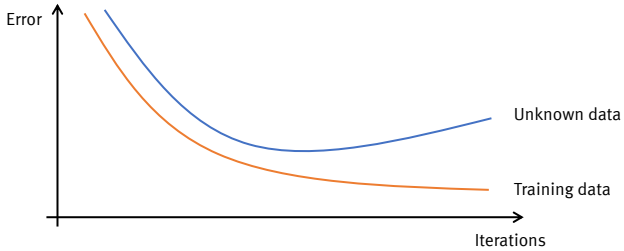




# Data Science

## Evaluating models






The problem of **overfitting** occurs, if a model fits too good to the training-data, and therefore performs worse on other data.

The problem of **underfitting** occurs, if a model can not adequately capture the structure of the data.

# Data Science
## Evaluating models



- Compute error (e.g. residuum) on independent dataset

In machine learning it is common to split the given data set into a training set and a test set. This is called **train-test splitting**.

- The train-test splitting can be obtained in different ways, depending on the application and data:
  - Random split: Randomly separate the data in train/test
  - Temporal split: For temporal data, split could be obtained by putting data before a date into train and after a date into test
- Often $80/20$ or $70/30$ split, i.e. use 80% of the data for training and 20% for testing

# Data Science
## Evaluating models



In order to compare machine learning models, it is useful to add a additional dataset, i.e. one performs a **train-validation-test split**.

- **Train set:** Used to train the different models
- **Validation set:** Used to compare the different models
- **Test set:** Used to verify that the best model performs well

How to measure the quality of an algorithm?

# Data Science
## Evaluating models

we focus on students

The **mean squared error (MSE)** of a predictor $\hat{\theta}$ for an unknown value $\theta$ is given by $E\left(\left(\hat{\theta}-\theta\right)^2\right)$. Thus, for predicted values $\hat{y}_i$ with exact values $y_i$ the MSE is given by

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

There are several alternative

- Mean average error (MAE): $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Mean absolute percentage error: $\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{|y_i|}$

**Task:** What is the difference of these errors?

# Data Science

## Evaluating models



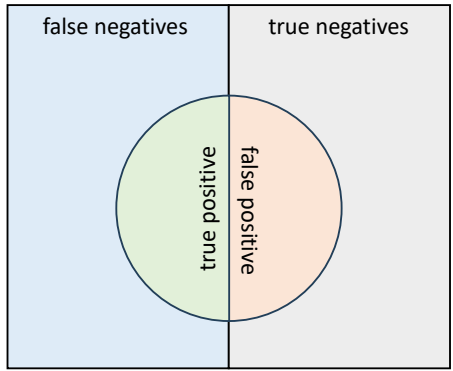
The **confusion matrix** is a table, collecting the quality of predictions on a dataset for binary classification problem.

| | | **actual values** | |
|---|---|---|---|
| | | Positive | negative |
| **predicted values** | Positive | True positive (TP) | False positive (FP) |
| | Negative | False negative (FN) | True negative (TN) |

# Data Science

## Evaluating models





# Data Science

## Evaluating models



**Task:** Assume that the following table gives the results of a machine learning model on a test-set. What is the confusion matrix?

| $i$ | $y_i$ | $\hat{y}_i$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 0 | 1 |
| 7 | 0 | 1 |
| 8 | 0 | 0 |
| 9 | 1 | 0 |
| 10 | 1 | 1 |

# Data Science

## Evaluating models



**Task:** Assume that the following table gives the results of a machine learning model on a test-set. What is the confusion matrix?

| $i$ | $y_i$ | $\hat{y}_i$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 0 | 1 |
| 7 | 0 | 1 |
| 8 | 0 | 0 |
| 9 | 1 | 0 |
| 10 | 1 | 1 |

| | | **actual values** | |
|---|---|---|---|
| | | 1 | 0 |
| **predicted values** | 1 | 4 | 2 |
| | 0 | 3 | 1 |

# Data Science
## Evaluating models



For the TP, FP, TN, FN values of confusion matrix we define

- the **accuracy** as

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- the **precision** as

$$\frac{TP}{TP + FP}$$

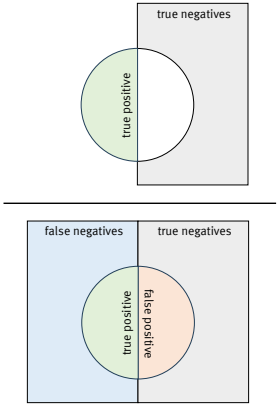
- the **recall** as

$$\frac{TP}{TP + FN}$$

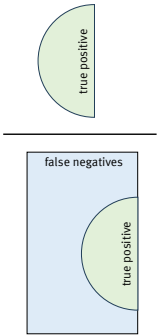
# Data Science

## Evaluating models

we focus on students








| | | |
|---|---|---|
| Proportion of all predicted correct. | Proportion of the true predicted correct. | Proportion of the true class found. |

# Data Science
## Evaluating models



**Task:** Compute accuracy, precision and recall for the given confusion matrix

| | | actual values | |
|---|---|---|---|
| | | 1 | 0 |
| **predicted values** | 1 | 4 (TP) | 2 (FP) |
| | 0 | 3 (FN) | 1 (TN) |

- **accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$
- **precision:** $\frac{TP}{TP+FP}$
- **recall:** $\frac{TP}{TP+FN}$

# Data Science
## Evaluating models



**Task:** Compute accuracy, precision and recall for the given confusion matrix

| | | actual values | |
|---|---|---|---|
| | | 1 | 0 |
| **predicted values** | 1 | 4 (TP) | 2 (FP) |
| | 0 | 3 (FN) | 1 (TN) |

- **accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$
- **precision:** $\frac{TP}{TP+FP}$
- **recall:** $\frac{TP}{TP+FN}$

- **accuracy:** $\frac{TP+TN}{TP+TN+FP+FN} = \frac{4+1}{4+1+3+2} = \frac{5}{10} = 0.5$
- **precision:** $\frac{TP}{TP+FP} = \frac{4}{4+2} = \frac{4}{6} = 0.66$
- **recall:** $\frac{TP}{TP+FN} = \frac{4}{4+3} = \frac{4}{7} = 0.57$

# Data Science

## Evaluating models



The predictive performance of a classifier can be measured with help of the **F1** score, which is given by

$$F_1 = 2\frac{\mathit{precision} \cdot \mathit{recall}}{\mathit{precision} + \mathit{recall}}$$

- Harmonic mean of precision and recall
- If recall and precision equals to $1$, the $F_1$ score equals to $1$.
- If recall or precision equals to $0$, the $F_1$ score equals to $0$.

# Data Science
## Evaluating models



**Task:** Compute the $F_1$ score of the given precision and recall values!

- **precision:** $\frac{TP}{TP+FP} = \frac{4}{4+2} = \frac{4}{6} = 0.66$
- **recall:** $\frac{TP}{TP+FN} = \frac{4}{4+3} = \frac{4}{7} = 0.57$

**Task:** Compute the $F_1$ score of the given precision and recall values!

- **precision:** $\frac{TP}{TP+FP} = \frac{4}{4+2} = \frac{4}{6} = 0.66$
- **recall:** $\frac{TP}{TP+FN} = \frac{4}{4+3} = \frac{4}{7} = 0.57$

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} = 2\frac{0.66 \cdot 0.57}{0.66 + 0.57} = 0.31$$

# Data Science
## Evaluating models



Computing the metrics of a machine learning model gives a glimpse on the quality. Unfortunately, this not always gives an information if the results are useful or not.

- Is $90\%$ accuracy good?
  - Assume $100$ data points, $90$ are of class *A* and $10$ of class *B*.
  - Predicting always class *A* would lead to an accuracy of $90\%$
- Defining a naive reference solution could help in interpreting the results

# Data Science

## Evaluating models



Consider the following setting: A machine learning model should predict the price of a car, based on the properties brand, style and engine.

**Task:** What are possible naive reference solutions?

# Summary

# Data Science

## **Summary:** Summary



- You know how to evaluate machine learning models and compare them